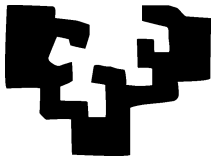
eman ta zabal zazu



Universidad del País Vasco
Euskal Herriko Unibertsitatea

Técnicas multivariantes de apoyo a la gestión del conocimiento. Extensiones y Aplicaciones

Juan Ignacio Modroño Herrán

Directoras:

Karmele Fernández Aguirre
María Isabel Landaluce Calvo

Bilbao, 3 de julio de 2012

A SusanA
n
d
o
n
Iker.

Agradecimientos

Quiero agradecer, en primer lugar, a mis directoras de tesis. La paciencia, el apoyo y el llevarme a donde yo no podía ir sólo que tanto Karmele como Marisa me han concedido para preparar esta tesis, excede la capacidad de este párrafo para agradecerlo con palabras. Solamente espero que este trabajo les llene de satisfacción tanto como a mí.

Me gustaría agradecer especialmente a las profesoras Mónica Bécue, Ana Martín y Pilar Zorrilla por su ayuda, comentarios y colaboración en diversos elementos de esta tesis. Sin ellas, ésta no sería lo que es.

No puedo olvidarme de mis compañeros y ex-compañeros del Departamento de Economía Aplicada III, de la Facultad de Ciencias Económicas de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU). Me gustaría agradecer a todos en su conjunto. Todos en algún momento y de la manera que les ha sido factible me han ayudado y alentado en el desarrollo de esta tesis. No obstante, me siento obligado a mencionar, de una manera especial, a Vicente, Inmaculada, Araceli, con quienes he trabajado también estos años. A Mariví, Marta, Petr, Iñaki. Susan, Mari Paz, Aurora. Ignacio, Jorge, Pilar, Javier. Fernando, Eva, Amaya, Txusa, Ana, Isa, Cecilia. La lista es corta, probablemente injusta, el orden es irrelevante y premeditadamente omite la conjunción "y".

Agradezco también el apoyo financiero del Grupo de investigación en Estadística y Computación creado en 2001 por Fernando Tusell, cuyo nombre actual es *Estadística y Optimización* IT-347-10 y está ahora liderado por Araceli Garín. También tengo que agradecer a la UPV/EHU por el contrato *Plan de negocio y estudio de mercado de una tienda de regalos institucionales* y al EUROLAB Central Archive for Empirical Social Research (actual Leibniz Institute for the Social Sciences) de la Universidad de Colonia, Alemania.

Finalmente, mis últimos y no menores agradecimientos son para mi familia. Para mis padres, Emilio y Basi, que siempre estuvieron ahí para apoyarme, lo necesitase o no. También para Susana, Iker y Andoni, a los que les doy un

grandísimo beso y les pido perdón por cuando no les dedico todo el tiempo que necesitan y se merecen. También a Porfi que, al igual que mis padres, hacen de aita cuando yo no estoy, y casi siempre mejor que yo.

Índice general

Índice de Figuras

Índice de Tablas

CAPÍTULO 1

Introducción

1.1. Contextualización

En los últimos años ha existido un gran crecimiento en nuestra capacidad de generar datos, debido principalmente a dos motivos. Por un lado, el gran poder de procesamiento y, por otro lado, al bajo coste de almacenamiento. La riqueza de estas voluminosas masas de datos reside en que en el seno de las mismas existe una gran cantidad de conocimiento *implícito*, de gran interés para la toma de decisiones o para mejorar nuestra comprensión de la realidad que nos rodea.

El descubrimiento de este conocimiento *implícito* y su transformación en conocimiento *explícito* conlleva todo un complejo y completo proceso que recibe el nombre de Gestión del Conocimiento. De forma general, los datos son la materia prima bruta. En el momento que un usuario atribuye a los datos, en un contexto concreto, algún significado pasan a convertirse en información. Cuando los especialistas encuentran o elaboran un modelo que permita la interpretación de esa información aportando una valor añadido, entonces se puede hablar de conocimiento.

En la Gestión del conocimiento existe, por tanto, una jerarquía de la información que comienza con los **datos**, que consisten en un conjunto discreto de elementos objetivos acerca de distintos eventos, que pueden derivar de hechos, información, estadísticas o similares, tanto históricos como derivados del cálculo o de la experimentación. Los datos en sí mismos carecen de sentido, ya que sólo describen lo que sucede de manera parcial y no proporcionan juicios ni interpretaciones, por lo que no ayudan a la toma de decisiones.

En el siguiente nivel se sitúa la **información**, que consiste en datos dotados de relevancia o propósito. La información se construye a través de datos que han sido contextualizados para que puedan ser analizados estadísticamente. Para ello es necesario efectuar un proceso de corrección de los mismos con el fin de eliminar datos erróneos y finalmente, se debe efectuar una condensación para que la información contenida sea resumida, es decir, sean datos concisos.

Es en el tercer nivel cuando se puede hablar del **conocimiento**, que consiste en un conjunto de hechos, verdades o principios obtenidos como resultado del estudio, del análisis, de la investigación de la información.

El paso del dato al conocimiento no es trivial, y requiere una serie de técnicas que ayuden al descubrimiento de patrones de comportamiento, relaciones, reglas, asociaciones o incluso excepciones útiles para el estudio de la realidad en cuestión. Todas estas técnicas constituyen una disciplina que se ha denominado Minería de Datos.

Minería de Datos es un término genérico que engloba técnicas, herramientas y resultados de investigación, usados para extraer información útil de grandes bases de datos. Existen muchas y muy diversas tecnologías de apoyo a la Minería de Datos, que se han utilizado desde hace mucho tiempo y la integración de estas tecnologías con la administración de datos ha contribuido mucho a mejorar el proceso.

Las más importantes de estas tecnologías son los métodos estadísticos y el aprendizaje automático. El aprendizaje automático consiste en la obtención de reglas de aprendizaje y modelos de los datos, para lo que a menudo también se necesita la ayuda de la Estadística. Por esta razón los métodos estadísticos constituyen uno de los componentes más importantes de la Minería de Datos. Además, existen otras tecnologías entre las que se incluyen visualización, procesamiento paralelo y apoyo a la toma de decisiones. Las técnicas de visualización ayudan a presentar los datos para facilitar la Minería de Datos. Las técnicas de procesamiento paralelo ayudan a mejorar el rendimiento de la Minería de Datos. Los sistemas de apoyo a la toma de decisiones ayudan a discriminar los resultados y proporcionan los resultados esenciales.

La Minería de Datos utiliza una gran variedad de técnicas, pero todas ellas siguen el siguiente esquema general:

1. Selección y procesamiento de los datos: generalmente los datos disponibles no se encuentran en la forma más adecuada para el tratamiento de los mismos, por lo que es necesario realizar una operación de filtrado de valores incorrectos o efectuar un muestreo.

2. Selección de características a estudiar: con el propósito de simplificar el procesamiento de los datos y de agilizarlo. Las técnicas de visualización

de los datos son adecuadas para localizar patrones o para obtener una referencia de la calidad del conjunto de los datos.

3. Uso de algoritmo de extracción de conocimiento: con el objetivo de elaborar modelos de conocimiento a partir de los patrones de comportamiento y de asociación entre las distintas variables del estudio.

4. Análisis predictivo: ciertas técnicas permiten, a partir de datos históricos y/o de suposiciones sobre determinadas condiciones, predecir el comportamiento de eventos determinados.

5. Interpretación y evaluación de los resultados: se verifica si los resultados obtenidos aportan alguna novedad en las relaciones ocultas entre los datos que permitan guiar la toma de decisiones.

La Minería de Datos tiene distintas variaciones, siendo una de ellas el Text Mining o Minería de Textos. En este caso, las técnicas de Minería de Datos se aplican a descubrir patrones ocultos en textos, esto es, constituye el proceso de extraer conocimiento a partir de la información contenida en bases de datos textuales.

El proceso de Minería de Datos tiene aplicaciones en muy diversos dominios y con diferentes objetivos como en, por ejemplo, empresas, gobiernos, universidades u hospitales. Esto es, su potencial tiene utilidad para todas aquellas organizaciones interesadas en explotar sus bases de datos. La naturaleza de estos datos puede ser métrica, nominal o incluso, como ya se ha mencionado, textual.

Resumiendo, diremos que la Gestión del Conocimiento se refiere al conjunto de procesos desarrollados en una organización para crear, organizar, almacenar y transferir el conocimiento. La Minería de Datos es la disciplina que tiene por objetivo la extracción del conocimiento implícito en la bases de datos. Por tanto, tiene un papel fundamental en el proceso de convertir en explícito al conocimiento implícito y en las distintas etapas del proceso de Gestión del Conocimiento. La Estadística es, a su vez, la disciplina que proporciona las técnicas y herramientas apropiadas para la adecuada extracción del conocimiento permitiendo encontrar relaciones y patrones dentro de los datos que contribuyen a la creación de modelos, esto es, representaciones abstractas de la realidad, que debidamente interpretados dan un significado a las relaciones y patrones encontrados.

Es en este contexto *Gestión del conocimiento* → *Minería de datos* → *Estadística* en el que se sitúa la presente Tesis Doctoral. En ella, a través de algunas aportaciones teóricas y empíricas, queremos reflejar la utilidad y la versatilidad de algunos métodos estadísticos en todas y cada una de las etapas que conlleva la Gestión del Conocimiento y, por ende, la Minería de Datos.

1.2. Estructura y organización

La tesis queda estructurada en dos partes. En la primera se presentan, de manera homogénea, las principales técnicas multivariantes que permiten realizar un análisis exploratorio de grandes masas de datos. Se han seleccionado metodologías apropiadas tanto para el análisis de datos métricos, como datos cualitativos e incluso textuales, pudiendo conformar los datos una sola tabla, o estar estructurados en distintas tablas atendiendo a diversos criterios. En esta parte se ha querido reflejar una de las grandes ventajas que permiten las técnicas exploratorias seleccionadas: *aprender de los datos*, por lo que constituyen una herramienta esencial en las primeras etapas de la Gestión del Conocimiento-Minería de Datos.

En el capítulo 2 también se han incluido algunas técnicas confirmatorias o predictivas de gran utilidad para la elaboración de modelos así como para las últimas etapas del proceso de extracción del conocimiento y que han sido utilizadas en las aplicaciones de los capítulos posteriores.

Una aplicación empírica con datos reales constituye el capítulo 3 de esta primera parte: Estudio de la viabilidad de una tienda de regalos corporativos en la UPV/EHU. A través de este estudio se pone claramente de manifiesto la riqueza interpretativa de las técnicas seleccionadas, permitiendo extraer un conocimiento exhaustivo de la realidad analizada, de gran utilidad para la organización mencionada.

La tesis también consta de una segunda parte, integrada por los capítulos 4 y 5, centrada en el análisis de datos estructurados en diferentes tablas. En ella quedan reflejadas las principales aportaciones de este trabajo de investigación, focalizadas en la exploración conjunta de varias tablas de datos partiendo de la filosofía propia del Análisis Factorial Múltiple. Concretamente, en el capítulo 4 se presenta una extensión adaptada al tratamiento simultáneo de varias tablas cualitativas referidas a distintos conjuntos de individuos. Por su parte, en el capítulo 5 se presenta una metodología que permite el análisis de lo que se denominan tablas mixtas, esto es, tablas múltiples compuestas por tablas de distinta naturaleza, métrica, cualitativa y textual. Ambos capítulos, una vez presentada la herramienta estadística, se completan con los resultados obtenidos en diferentes aplicaciones con datos reales, utilizando los datos del estudio de la viabilidad de la tienda corporativa como principal hilo conductor.

Las conclusiones alcanzadas en todos y cada uno de los capítulos y las futuras líneas de investigación que han quedado abiertas a raíz de este trabajo, y que aparecen en el capítulo 6, así como las referencias bibliográficas utilizadas para su elaboración, constituyen el broche de esta tesis.

CAPÍTULO 2

Técnicas Multivariantes: del dato al conocimiento

2.1. ¿Qué es el data mining?

La minería de datos o data mining surge de la disponibilidad creciente de datos a raíz del desarrollo y la práctica universalización de la informática a la que se ha llegado ya desde finales del siglo XX. Ésta ha permitido, por un lado, la casi instantánea obtención y almacenamiento de datos susceptibles de ser analizados y, por otro, un rápido procesamiento de los mismos de cara a la extracción y análisis de las relaciones que hay entre y dentro de las entidades o variables que esos mismos datos están midiendo. Esta tarea, la de procesamiento de la información, habría llevado en los inicios de la aplicación de los análisis multivariantes semanas de trabajo de un equipo de expertos trabajando a tiempo completo.

El data mining surge auspiciado por el desarrollo tecnológico y se desarrolla, de manera casi imperceptible pero constante, en el ámbito empresarial. En este ámbito los datos se recogen con mayor avidez en la búsqueda de una mayor competitividad y así se empiezan a formar bases de datos cada vez más grandes. Este proceso se generaliza extendiéndose a empresas cada vez más pequeñas. A la vez que las bases de datos crecen tanto en tamaño (número de datos) como en complejidad (número de variables que se miden, diferentes periodos de tiempo en que se toman) se hace patente la necesidad de usar herramientas de análisis que exploren esas bases de datos (o subconjuntos de interés de las mismas) en busca de la información que contienen. Es de particular interés encontrar

información oculta, no previsible, y que esta información pueda obtenerse de la manera más automatizada posible. Este proceso de búsqueda de información y, sobre todo, las herramientas necesarias para ello, provienen de la Estadística.

El data mining es una disciplina científica originada, por tanto, en el ámbito empresarial. Este sector empresarial es un entorno que, a escala global, está prácticamente atomizado, lo que ha provocado que existan no sólo muchos tipos de procesos y de técnicas apropiadas, sino múltiples definiciones de este área. Es, todavía, un campo en fase de unificación, aunque ya se ha avanzado mucho en este sentido.

Probablemente, la primera definición de data mining está en Fayyad et al. (1996). En ella se reduce el data mining a un simple y único paso de un proceso más amplio del denominado Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases -KDD- en inglés). Este paso consiste en la búsqueda de patrones de interés en una forma representativa concreta, como pueden ser reglas o árboles de clasificación, regresión o clustering. Las tareas de concreción del problema, selección del subconjunto de datos, limpieza de datos, elección del método de análisis o la tarea final de elaboración del informe sobre el trabajo serían parte del proceso KDD, como otras etapas del mismo y diferentes al data mining. Hoy día, sin embargo, los conceptos de *data mining* y *knowledge discovery* se consideran mayoritariamente sinónimos.

Una definición de data mining más reciente y proveniente del mundo científico es la de Hand et al. (2001):

> *Data mining es el análisis de (a menudo grandes) conjuntos de datos observados para encontrar relaciones no sospechadas y resumir los datos de forma que sean entendibles y útiles para el propietario de los datos.*

Esta definición hace hincapié en varios aspectos. Primero, que los datos son observados, no se obtienen a partir del diseño de un experimento para resolver un problema planteado previamente, a diferencia de como ocurre en varias áreas de la Estadística, particularmente las asociadas a las Ciencias Experimentales, como la Física. Esta es una particularidad de las bases de datos a partir de las cuales ha surgido esta disciplina, provenientes de Ciencias no experimentales o en las que la posibilidad de experimentar es parcial y reducida, pero en las que existe abundancia de datos disponibles. Esto ocurre en disciplinas como el Marketing, el comportamiento en Internet en el área de las Ciencias Sociales y en otras como la Meteorología, la Ingeniería, la Genética o la Astronomía. En este sentido, se denomina con frecuencia al *data mining* como análisis de datos *secundario*, por ejemplo, en Marketing.

En segundo lugar, la definición establece que un primer objetivo es encontrar relaciones no sospechadas. El que sean no sospechadas es, probablemente,

más un deseo que una condición *sine qua non* para que un proceso de análisis pueda catalogarse como de *data mining*. Si la base de datos es suficientemente grande y compleja es relativamente fácil encontrar relaciones no previsibles a priori, pero esto no tiene por que ser siempre así. En cualquier caso, encontrar (cuantificándolas) relaciones no sospechadas requiere de un proceso de selección de datos y de preparación para su uso mediante una o varias herramientas de análisis. El análisis es, en realidad, un proceso donde tras preparar los datos para buscar relaciones, se elige una o varias herramientas estadísticas que se consideran adecuadas y se evalúan los resultados obtenidos a través de ellas. Las herramientas, normalmente, pueden utilizarse de varias maneras y el proceso puede implicar el uso de varias de ellas e incluso desechar una herramienta para probar otra distinta.

Finalmente, se establece un segundo objetivo, que es el de resumir los datos de forma que sean entendibles y útiles para su propietario. Este objetivo, al igual que el anterior, tiene un sustrato de fondo que tiene mucho que ver con el entorno empresarial, orientado a resultados inmediatos, en el que el data mining ha nacido y crecido: la utilidad. Además, parece sugerir que el propietario de los datos no tiene que ser necesariamente un experto en análisis de datos, al poner énfasis en que el resumen ha de ser entendible. Nuevamente este objetivo está orientado a satisfacer las necesidades de gerentes empresariales, normalmente poco o nada expertos en el uso de herramientas estadísticas, pero que necesitan información basada en datos reales para la toma de decisiones.

Existen otras muchas definiciones y no es éste el lugar para enumerarlas ni para arriesgarse a poner una como la única válida. Por ejemplo, la Wikipedia (`http://es.wikipedia.org`), quizás no es un repositorio de conocimiento científicamente reconocido pero sí casi universalmente conocido y utilizado, sin embargo se atreve a dar una definición clara y precisa, cuando hace unos pocos años daba dos definiciones de las existentes. Simplemente dice que *consiste en la extracción no trivial de información que reside de manera implícita en los datos*. En general, las definiciones ponen el énfasis en la extracción de la información, en el resumen de la misma y en que sea útil para la toma de decisiones. Algunos usuarios cualificados opinan que, en realidad, data mining es lo que hace cualquier analista de datos, incluso sin saberlo.

Para finalizar, señalar que algunas definiciones recientes añaden a las anteriormente expuestas que la extracción de información y su proceso ha de hacerse de forma automatizada. Sobre esto hay que comentar que de forma automatizada probablemente pueden proporcionarse únicamente algunos estadísticos univariantes y multivariantes sencillos; el resto de herramientas de análisis requieren de una reflexión por parte de un especialista sobre su conveniencia y el alcance de la interpretación de los resultados que se obtienen con

ellas. Sin embargo, para avanzar en el área de la automatización, al menos por el lado de la preparación de los datos y de la especificación de los métodos a emplear en el proceso, un número importante de empresas, entre las que están IBM, SAS, SPSS, Visa y Oracle, han creado o apoyan el *Data Mining Group* (`http://www.dmg.org`) con el objetivo de integrar los diferentes sistemas de bases de datos existentes en un sistema de lenguaje común que permita su análisis mediante las herramientas de data mining conocidas y que, a su vez, pueden estar implementadas mediante software propietario distinto (como SAS o SPSS, por ejemplo) de difícil interoperatividad. A este lenguaje lo llaman PMML acorde con las especificaciones del metalenguaje basado en etiquetas XML[1] de uso frecuente en Internet y en bases de datos.

2.1.1. Herramientas de data mining

Existen multitud de herramientas de análisis de datos útiles en el proceso de data mining, entre las que destacan las de clasificación en cualquiera de sus variantes. Sin embargo, éstas no son las únicas o, al menos, aparecen combinadas con otras en muchas ocasiones.

Hand et al. (2001) hacen una clasificación de las herramientas agrupadas por tareas, que corresponden a objetivos diferentes para la persona que analiza los datos. Sin pretender ser única, es, posiblemente, de las más completas.

- *Análisis de Datos Exploratorios* Según Hand et al. (2001), es la simple exploración de los datos sin objetivos definidos, a lo que habría que añadir un establecimiento mínimo de supuestos en los procedimientos a emplear. Generalmente usan con profusión métodos gráficos de visualización con un grado elevado de interactividad con el usuario. Generalmente son útiles para conjuntos de datos de escasa dimensionalidad. Si la dimensionalidad es grande, generalmente se asocian con técnicas de proyección como las técnicas factoriales. Estos métodos incluyen desde un simple histograma o un diagrama X-Y hasta un Análisis de Componentes Principales, por ejemplo.

- *Modelización Descriptiva* El objetivo de estas herramientas es describir completamente los datos, o el proceso que los genera. Se incluyen:

    - Modelos para la distribución de probabilidad de los datos o estimación de su densidad.

[1] XML no es un lenguaje en sí; define una manera de cómo debe estructurarse un lenguaje. Está basado en etiquetas; en este sentido se parece a otros lenguajes anteriores de uso diverso como HTML o, incluso a LaTeX.

  - Algoritmos de particionado en grupos sobre un cierto espacio, como segmentación y clasificación. La segmentación es una partición realizada a priori sobre características de los, por ejemplo, individuos como edad, género o salario, mientras que la clasificación se obtiene a partir de un algoritmo que busca grupos heterogéneos de fuerte homogeneidad interna.
  - Modelos que describen la relación entre las variables (modelización de la dependencia).

- *Modelización Predictiva* Este tipo de tareas suponen el establecimiento de modelos que permiten predecir el valor de una variable (no en la acepción temporal del término) dados los valores que toman otras variables. La diferencia con la modelización descriptiva es que en ella todas las variables tienen la misma importancia y entran en el análisis de la misma manera. Hand et al. (2001) diferencian según el tipo de variable a predecir: si es categórica hablan de clasificación (como en el caso anterior) y si es cuantitativa de regresión. En este caso, la clasificación incluye métodos como los modelos logit o el análisis discriminante.

- *Descubrimiento de Patrones y Reglas* Se centran en reglas de asociación para encontrar, por ejemplo, individuos con cierto comportamiento, por ejemplo, el uso fraudulento de números de teléfono o patrones de morosidad bancaria. Se basa en el uso de valores lógicos y/o condicionales del tipo `and, or, if, then, ...` y en el cálculo de frecuencias como estimaciones de probabilidades condicionadas. En esta tarea a veces es muy importante la detección y estudio de observaciones anómalas o *outliers*, cuando éstos son datos realmente producidos (y no un simple error de introducción de datos en la base).

- *Recuperación por Contenido* En este tipo de tareas, el investigador tiene un patrón o conjunto de características de un individuo u objeto y quiere encontrar aquéllos otros individuos u objetos de características similares. Aquí entra en juego la noción de distancia y es de particular importancia su elección. Este tipo de tareas suele utilizarse en análisis textual (localización de documentos en la Web, atribución de un documento a un autor) y análisis de imágenes (reconocimiento de imágenes). Este tipo de tareas son relativamente poco automatizables, en el sentido de que el especialista ha de tomar muchas decisiones en el proceso de análisis hasta llegar a un resultado.

A pesar de que las tareas anteriores están bien diferenciadas, en la práctica estas diferencias pueden llegar a diluirse. Unas veces métodos de tareas diferen-

tes usan elementos comunes (como una medida de distancia) y otras veces esas mismas tareas resultan más complementarias que competidoras entre sí. Así, Lebart et al. (2006, págs. 5-7) tras reivindicar los métodos exploratorios multivariantes como herramientas útiles en data mining, distinguen entre dos tipos de aproximaciones a la estadística multidimensional: las de tipo descriptivo y exploratorio, por un lado, y las de tipo inferencial o confirmatorio, por otro. Las herramientas multivariantes de tipo descriptivo se corresponden en buena parte con los dos primeros tipos de tarea mencionadas por Hand et al. (2001) y con el análisis textual, mientras que las herramientas multivariantes de tipo inferencial hacen lo propio con las tareas de modelización predictiva. Lebart et al. (2006) afirman que ambos tipos de aproximaciones son complementarias dado que, a menudo, una fase del proceso de análisis de datos de tipo explicativo y predictivo es precedido por otra de tipo exploratorio o descriptivo. Por otro lado, las técnicas que de forma clásica se clasifican como de un tipo u otro a menudo están altamente relacionadas con otras de la aproximación alternativa. Por ejemplo, los ejes de un Análisis de Componentes Principales están muy próximos a los factores extraídos de un análisis factorial que proviene de un modelo especificado a priori, mientras que otras veces el análisis de regresión es utilizado de forma exploratoria dado que los coeficientes de regresión tienen una relación cercana con los coeficientes de correlación, simples o múltiples, característicos del análisis exploratorio.

2.2. Herramientas exploratorias de data mining para el análisis de grandes encuestas

En esta sección se hace referencia a diversas herramientas exploratorias procedentes del análisis multivariante para el análisis de datos de encuestas de, potencialmente al menos, gran tamaño. Con este tipo de técnicas la relación entre variables es totalmente simétrica, en el sentido de que ninguna variable se ve privilegiada ni tratada de forma diferente a otras de las que intervienen en el análisis. Dentro de la terminología empleada en entornos de minería de datos, estás técnicas se engloban dentro de las tecnicas de Estadística Descriptiva y Exploratoria (Lebart et al. 2006) o de Análisis de Datos Exploratorios (Hand et al. 2001).

Las técnicas a considerar permiten estudiar, interpretar y elaborar material sobre conjuntos de variables tanto cuantitativas como cualitativas. El tratamiento de variables cualitativas incluye casos particulares como las frecuencias que dan lugar a las tablas léxicas del Análisis Textual (Lebart et al. 1998).

Se consideran dos tipos de técnicas: las basadas en métodos factoriales

enfocados a la reducción de dimensionalidad y a la interpretación de las nuevas variables o ejes resultantes y los métodos de clasificación, cuyo objetivo es proporcionar grupos de individuos u observaciones mediante criterios de proximidad y algoritmos iterativos diseñados para obtenerlos.

Entre las técnicas factoriales, consideramos:

1. *El Análisis de Componentes Principales (ACP):* es una técnica largamente conocida para el análisis de datos de tablas de individuos sobre los que se han medido variables cuantitativas continuas. El objetivo es obtener e interpretar una tabla resumida de la anterior que presenta un menor número de variables, denominadas factores o componentes.

2. *El Análisis de Correspondencias (AC):* se aplica a tablas de frecuencias correspondientes a los valores posibles de dos variables categóricas, que pueden ser cualitativas (también denominadas nominales) o continuas discretas. Es extensible a cualquier tabla de frecuencias, de manera que su uso es también aplicable al Análisis Textual. De hecho, el Análisis Textual se sitúa en el origen del AC.

3. *El Análisis de Correspondencias Múltiples (ACM):* es útil para el análisis de tablas de cualquier número de variables categóricas y está relacionado tanto con el ACP como con el AC.

En todos estos métodos factoriales son de suma importancia y utilidad las representaciones gráficas de las proyecciones de las filas y las columnas de la tabla analizada sobre los ejes que se obtienen en cada caso. Dada la distinta naturaleza de las variables que se usan y de los métodos que se emplean, la interpretación de esas proyecciones son diferentes en cada caso.

Los métodos de clasificación, por su parte, buscan la obtención de grupos de individuos que son homogéneos internamente y, a la vez, diferentes entre ellos. Estos grupos se pueden obtener de y caracterizar a partir de unas variables que no son otras que los factores que se obtienen de los métodos factoriales antes expuestos, por lo que son, en ese sentido, complementarios.

2.3. El Análisis de Componentes Principales de variables continuas

El Análisis de Componentes Principales o ACP es una técnica estadística de corte descriptivo que permite una visualización aproximada de una tabla de datos medidos sobre variables cuantitativas y una descripción simultánea

de las asociaciones existentes entre esas variables y de las similitudes entre los datos correspondientes a individuos u observaciones.

La descripción que se realiza en ACP es efectuada a partir de unas pocas variables denominadas componentes principales. Estas se obtienen a partir de las variables cuantitativas que aparecen en la tabla de datos bidimensional que cruza individuos y variables. Generalmente, las variables son continuas, aunque en ocasiones se aplica sobre variables discretas de tipo ordinal (como, por ejemplo, variables que se obtienen a partir de respuestas a una pregunta, cuyo rango numérico se ajusta a una escala de Likert con un número suficiente de categorías). Este tipo de variables, que entran dentro de la clase de las variables categóricas, son susceptibles de ser analizadas alternativamente mediante análisis de Correspondencias Simples o Múltiples (ver Sección 2.5), partiendo de las distancias entre perfiles de frecuencias observadas de las categorías que abarcan los valores que pueden tomar las variables.

Los componentes principales permiten la observación de las similaridades entre los individuos de la tabla en base a unas pocas dimensiones, tantas como componentes se consideren, de forma que éstos reflejan la parte más importante de la información proporcionada por las variables originales de la tabla. Dichos componentes se obtienen como combinaciones lineales de esas mismas variables. Es frecuente, casi obligado, proyectar en uno o varios planos la dispersión bidimensional que mejor resume la dispersión multidimensional de las variables originales, al tiempo que se obtiene el grado de fidelidad de esta representación.

En los casos en los que el ACP es utilizado como una etapa previa de un método de tipo inferencial, confirmatorio o de dependencia generalmente el objetivo suele ser desechar variables redundantes o, al menos, ofrecer una alternativa si se quiere tratar con ellas sin eliminarlas, como en Regresión por Componentes Principales.

El ACP es un estudio de una tabla de dos dimensiones de corte dual. Lo es en el sentido de que, por un lado, las clases existentes de individuos quedan caracterizadas por las variables (según que alcancen valores bien elevados o bien menores en ellas) y, por otro, se obtienen grupos de variables que quedan vinculadas a través de individuos típicos que puntúan valores especialmente altos o bajos en ellas.

Finalmente, las variables que conforman una tabla pueden ser heterogéneas en cuanto a su naturaleza. Una manera de tratarlas es seleccionar un grupo como activas, que serán aquellas de las cuales se extraerán los componentes principales y considerar el resto como ilustrativas. Dichas variables pueden ser útiles para explicar las relaciones entre las variables originales que se obtienen en el análisis.

Preparación de los datos y distancia a utilizar

Distancia Si denotamos por $\boldsymbol{R}$ a la tabla de datos original de n individuos $\times$ p variables, r_{ij} será el valor que toma para el individuo i la variable j. La semejanza entre dos individuos, i e i' se mide a través de la distancia euclídea medida sobre todas las variables activas disponibles, de forma que la distancia al cuadrado entre ellos, en un espacio p-dimensional, se define como

$$d^2(i,i') = \sum_{j=1}^{p}(r_{ij} - r_{i'j})^2 \tag{2.1}$$

De la misma forma, dada la dualidad existente en la tabla, dos variables que tengan valores muy próximos para todos los individuos, con un alto grado de asociación, estarán representadas por dos puntos muy próximos de $\mathbb{R}^n$.

Normalización Es prácticamente universal el centrado previo de los datos, lo que facilita la interpretación de los resultados y la comparación con las medidas más habituales de Estadística Descriptiva. Sin embargo, la diferente dispersión de las variables activas influye en la distancia entre los individuos, de forma que en muchas ocasiones las variables también se tipifican, dando lugar a un ACP normalizado o normado. Esto es particularmente importante cuando las variables presentan diferentes unidades de medida de elección siempre arbitraria y que, a menudo, es inevitable si miden magnitudes de diferente naturaleza. A la matriz $n \times p$ de datos centrados o tipificados la denotamos igualmente $\boldsymbol{X}$, a sus filas $\boldsymbol{x(i)}$ (puntos-individuo de $\mathbb{R}^p$) y a sus columnas $\boldsymbol{x_j}$ (puntos-variable de $\mathbb{R}^n$).

Ponderación de los individuos Es posible que sea interesante en algunos casos que los individuos tengan cada uno un peso diferente p_i. Sin embargo, en general esto no es así, en cuyo caso se utiliza un peso idéntico $p_i = 1/n$ para todos ellos.

Ajuste de la nube de puntos-individuo

El objetivo del ACP es buscar unas variables nuevas, denominadas componentes, que recojan el máximo posible de la variabilidad contenida en la tabla de datos con el menor número posible de ellas. Para la primera componente, se trata de encontrar una recta de forma que se maximice la suma de las distancias al cuadrado entre las proyecciones de los puntos sobre esa recta, de forma que la recta recoja la mayor parte de la variabilidad de esa nube. Como se prueba, por ejemplo, en Lebart, Morineau & Piron (2000) esto equivale a

maximizar las distancias entre las proyecciones de esos mismos puntos sobre la recta y el centro de gravedad de la nube, que no es otra cosa que el vector de medias, y de ahí la operación de centrado de la tabla de datos. Para las componentes sucesivas se busca maximizar la misma suma de distancias sujeta a que la recta que se obtiene en segundo lugar sea ortogonal a la anterior y así sucesivamente hasta tener tantas componentes como variables linealmente independientes.

El objetivo de maximizar la variabilidad de la nube de puntos puede escribirse así:

$$\text{máx}\, I = \sum_{i=1}^{n} p_i d^2(i, G) \tag{2.2}$$

donde G es el centro de gravedad, p_i es el peso dado a los individuos y a la magnitud I se le denomina inercia. Si los pesos son todos iguales a $1/n$, inercia equivale a varianza.

Se denota ψ_i a la proyección del punto-individuo i sobre una recta. Si la recta está en la dirección del vector unitario $\boldsymbol{u}$, se calcula como

$$\psi_i = \boldsymbol{x(i)'u} = \sum_{j=1}^{p} x_{ij} u_j \quad \text{sujeto a} \quad \boldsymbol{u'u} = 1 \tag{2.3}$$

Matricialmente, considerando los n puntos:

$$\boldsymbol{\psi = Xu} \tag{2.4}$$

De forma que la inercia (o varianza) de todas las proyecciones sobre una recta es

$$\lambda = \sum_{i=1}^{n} p_i \psi_i^2 \tag{2.5}$$

la cual se denomina inercia proyectada sobre una recta, eje o componente. Matricialmente,

$$\lambda = \boldsymbol{\psi' N \psi = u' X' N X u} \tag{2.6}$$

donde $\boldsymbol{N}$ es una matriz diagonal que contiene los pesos p_i en la diagonal principal. La solución al problema de máximo de la ecuación (2.6) con la restricción $\boldsymbol{u'u = 1}$ viene dada por

$$\boldsymbol{X' N X u} = \lambda \boldsymbol{u} \tag{2.7}$$

de lo que se deduce que λ y $\boldsymbol{u}$ son valores propios y vectores propios asociados de $\boldsymbol{X'NX}$ que no es más que la matriz de covarianzas de la matriz de datos original (para ACP no normado) o de correlaciones (para ACP normado). Existen p valores y vectores propios, de forma que el eje de máxima inercia

coincide con la recta que contiene al primer vector propio $\boldsymbol{u_1}$. Este eje de máxima inercia es ortogonal al segundo y así sucesivamente.

Dada la ecuación (2.4), el vector que contiene las coordenadas de los n individuos sobre un eje α es

$$\boldsymbol{\psi_\alpha} = \sum_{j=1}^{p} u_{\alpha,j} \boldsymbol{x_j} \tag{2.8}$$

de forma que la varianza es $var(\boldsymbol{\psi_\alpha}) = \lambda_\alpha$

Nube de puntos variable

Distancias entre puntos-variable En un ACP normado, la tipificación supone que las variables se sitúan en una hiperesfera de radio unidad, ya que la distancia de una variable cualquiera j al origen es:

$$d^2(j, O) = \sum_{i=1}^{n} = p_i x_{ij}^2 = 1 \tag{2.9}$$

Puede demostrarse (Lebart, Morineau & Piron (2000)) que la distancia cuadrática entre dos variables se escribe en función de su coeficiente de correlación

$$d^2(j, j') = 2(1 - r_{jj}) \tag{2.10}$$

Por lo que esa distancia cuadrática ha de pertenecer al intervalo $[0, 4]$ y es mayor cuanto mayor sea la correlación y viceversa. Además, en un espacio $\mathbb{R}^n$, ese coeficiente de correlación es igual al coseno del ángulo que forman los dos vectores variable $\boldsymbol{x_j}$ y $\boldsymbol{x_{j'}}$.

En el caso de puntos variable proyectados sobre el plano formado por dos ejes factoriales, la proximidad entre dos puntos variables puede interpretarse en términos de correlaciones siempre que estén cerca del círculo unidad, lo que significa que los dos puntos han de estar bien representados en el plano.

Ejes factoriales y relaciones de transición De manera similar a como se realiza el análisis de la nube de puntos individuo, la búsqueda de la dirección de máximo alargamiento de la nube de puntos variable contenida en la matriz $\boldsymbol{X'}$ consiste en la obtención de un vector de proyecciones $\boldsymbol{\phi = X'v}$ de forma que se maximice $\boldsymbol{v'XX'v}$ sujeto a $\boldsymbol{v'v = 1}$, lo que da como resultados los mismos valores propios no nulos del análisis de la nube de puntos individuo de la ecuación (2.7). Esto resulta en las relaciones de transición

$$\boldsymbol{v_\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \boldsymbol{N^{\frac{1}{2}} X u_\alpha} \tag{2.11}$$

$$\boldsymbol{u_\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \boldsymbol{X' N^{\frac{1}{2}} v_\alpha} \tag{2.12}$$

de forma que el vector de proyecciones de los puntos variables sobre un eje α puede calcularse como

$$\boldsymbol{\phi_\alpha} = \boldsymbol{X'N^{\frac{1}{2}}v_\alpha} = \sqrt{\lambda_\alpha}\boldsymbol{u_\alpha} \tag{2.13}$$

Calidad de la representación y ayudas a la interpretación

Tanto la calidad de representación como las ayudas a la interpretación son indicadores numéricos que responden a definiciones comunes para los tres métodos factoriales básicos, ACP, AC y ACM. Las diferencias se establecen en función de los elementos particulares del análisis factorial de que se trate.

Calidad o fidelidad de la representación La calidad de representación de una nube de puntos en un espacio multidimensional al ser proyectada sobre un plano cuyos ejes corresponden a un ACP como el descrito anteriormente, se mide a través del porcentaje de la inercia total proyectada sobre dicho plano. Dado que la inercia de las proyecciones de los puntos de la nube sobre un eje α es el valor propio correspondiente a ese eje λ_α (ver ecuación 2.5). La calidad de representación sobre un subespacio de dimensión q es, pues,

$$\tau_q = \frac{\sum_{\alpha\in\{q\}}\lambda_\alpha}{\sum_{\alpha=1}^{p}\lambda_\alpha} \quad q < p \tag{2.14}$$

Esta medida es útil para seleccionar el número de componentes que se retienen para *resumir* la tabla de datos original. El número de componentes que se retienen es aquél que proporciona un porcentaje de inercia proyectado considerado suficiente. Cuando el ACP es normado, en el que la matriz $X'NX$ es la matriz de correlación de elementos diagonales iguales a 1, se suele escoger aquéllos componentes con un valor propio o inercia proyectada mayores que 1, lo cual corresponde a un valor siempre elevado de dicha tasa.

Ayudas a la interpretación Son medidas que ayudan a seleccionar los individuos y las variables más característicos de los ejes o componentes y, por lo tanto, más significativos para efectuar una interpretación de los mismos, dado el significado de las variables.

Cosenos cuadrados de los individuos El coseno cuadrado es una medida de la calidad de representación de un punto en un subespacio factorial dado (un eje, generalmente). Para un eje α, es el coseno del ángulo que forma la distancia al cuadrado de un individuo al centro de gravedad y

la proyección al cuadrado del individuo sobre ese eje (o distancia de la proyección sobre el eje α al mismo centro):

$$Cos^2_\alpha(i) = \frac{d^2_\alpha(i,G)}{d^2(i,G)} = \frac{\psi^2_{\alpha i}}{d^2(i,G)} \tag{2.15}$$

de forma que, por construcción, $\sum_\alpha Cos^2_\alpha(i) = 1$.

Contribuciones de los individuos La contribución de un individuo i a la formación de un eje α es la parte de la varianza del mismo debida a ese individuo:

$$CTR_\alpha(i) = \frac{p_i \psi^2_{\alpha i}}{\lambda_\alpha} \tag{2.16}$$

de forma que $\sum_i CTR_\alpha(i) = 1$.

Contribuciones de las variables La misma magnitud puede computarse en el espacio dual de las variables, dando lugar a la contribución de un punto variable:

$$CTR_\alpha(j) = \frac{\phi^2_{\alpha j}}{\lambda_\alpha} \tag{2.17}$$

donde igualmente $\sum_j CTR_\alpha(j) = 1$. En un ACP normado, la contribución de una variable en términos absolutos al cuadrado $\phi^2_{\alpha j}$ es el coseno cuadrado del ángulo que forma la variable j con el eje α, ya que las variables están a distancia 1 del origen. Así, la coordenada $\phi_{\alpha j}$ es el coseno y la correlación entre la variable y el factor, lo que ayuda también a la interpretación.

Proyección de elementos suplementarios

En ocasiones, se dispone de información adicional a la tabla objeto de análisis la cual contiene los elementos (individuos y variables) denominados activos a partir de los cuales se extraen los ejes principales. Dicha información puede venir en forma de individuos adicionales o de variables adicionales, que pueden ser continuas o categóricas. A este tipo de información adicional se le denomina elementos suplementarios, y la proyección sobre los ejes que se obtienen del análisis mediante los elementos activos ayudan a clarificar la interpretación de esos ejes.

Individuos suplementarios Para proyectar individuos adicionales se tipifican los valores obtenidos para ellos de las variables activas, x^+_{ij} (si el ACP es

normado), pero con las medias y las desviaciones típicas que se obtienen de la matriz de variables activas

$$x_{ij}^{+} = \frac{r_{ij}^{+} - \overline{r}_j}{s_j} \tag{2.18}$$

Y se obtienen sus coordenadas de forma análoga a los individuos activos en la ecuación (2.4) como $\boldsymbol{\psi}_{\boldsymbol{\alpha}}^{+} = \boldsymbol{X}_{+}\boldsymbol{u}_{\boldsymbol{\alpha}}$, es decir, utilizando los ejes unitarios obtenidos a partir de los elementos activos.

Variables continuas suplementarias De cara a poder posicionar estas variables en la esfera de radio unidad (análisis normado) es preciso tipificar estas variables suplementarias acorde a sus medias y varianzas:

$$x_{ij}^{+} = \frac{r_{ij}^{+} - \overline{r}_j^{+}}{s_j^{+}} \tag{2.19}$$

para luego proyectarlas sobre los ejes principales mediante $\boldsymbol{\phi}_{\boldsymbol{\alpha}}^{+} = \boldsymbol{X}_{+}'\boldsymbol{v}_{\boldsymbol{\alpha}}$.

Variables nominales suplementarias Cuando una variable es categórica, la proyección de sus categorías viene dada por la proyección en suplementario de tantos individuos promedio como número de categorías sean consideradas para esas variables categóricas.

2.4. Análisis de Correspondencias

El Análisis de Correspondencias (AC) es una técnica multidimensional de reducción de dimensionalidad en la tabla de frecuencias, o tabla de contingencia, que cruza las modalidades contenidas en dos variables cualitativas y desarrollada por Benzécri (1982). Otros autores han derivado total o parcialmente técnicas similares de forma independiente y con otros nombres, como Guttman (1941) y Hayashi (1956) con *Quantification Methods* o Nishisato (1980) con *Dual Scaling*.

El AC es un método de análisis exploratorio de datos cuyo objetivo es analizar las semejanzas entre las filas, por un lado, y las columnas, por otro, de una manera totalmente simétrica. Consiste en una reducción de dimensionalidad, como en el Análisis de Componentes Principales (ACP), pero en este caso de una matriz de frecuencias transformada en perfiles-fila o perfiles-columna.

Denotamos k_{ij} a la frecuencia absoluta correspondiente a la intersección de la fila i-ésima con la columna j-ésima de una tabla de frecuencias de orden $I \times J$ y $k = \sum_{i,j} k_{ij}$ a la frecuencia total. En este caso, podemos calcular las frecuencias relativas como $f_{ij} = \frac{k_{ij}}{k}$ de forma que $\sum_i \sum_j f_{ij} = 1$. Se definen las

frecuencias relativas marginales de, respectivamente, las filas y las columnas de la tabla de contingencia como $f_{i.} = \sum_j f_{ij}$ y $f_{.j} = \sum_i f_{ij}$. Entonces, tenemos:

- perfiles-fila: $\frac{f_{ij}}{f_{i.}}$ El conjunto de los perfiles-fila es una nube de I puntos sobre el espacio $\mathbb{R}^J$.
- perfiles-columna: $\frac{f_{ij}}{f_{.j}}$ El conjunto de los perfiles-columna es una nube de J puntos sobre el espacio $\mathbb{R}^I$.

Los centros de gravedad de las nubes de puntos fila y puntos columna son, respectivamente:

$$\sum_{i=1}^{I} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j} \quad \text{y} \quad \sum_{j=1}^{J} f_{.j} \frac{f_{ij}}{f_{.j}} = f_{i.} \tag{2.20}$$

La distancia entre los puntos definidos por dichos perfiles no es la euclidea habitual sino la denominada χ^2. Se diferencia de la euclídea en que se pondera por el inverso de la marginal de las columnas o filas, respectivamente:

$$d^2(i, i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \text{ (perfiles-fila)} \tag{2.21}$$

$$d^2(j, j') = \sum_{i=1}^{I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \text{ (perfiles-columna)} \tag{2.22}$$

de forma que se da mayor peso a modalidades de efectivo más débil. Esta distancia tiene la propiedad de *equivalencia distribucional*: si se amalgaman dos categorías de una misma variable con perfiles iguales, las distancias entre las modalidades no varían para ninguna de las dos variables.

Especificación del AC Definimos la matriz de frecuencias relativas $\boldsymbol{F} = \{f_{ij}\}$, y las matrices diagonales de frecuencias marginales como $\boldsymbol{D_I} = diag\{f_{i.}\}$ y $\boldsymbol{D_J} = diag\{f_{.j}\}$. De esta forma, las matrices de perfiles-fila y perfiles-columna son, respectivamente, $\boldsymbol{D_I}^{-1}\boldsymbol{F}$ y $\boldsymbol{D_J}^{-1}\boldsymbol{F'}$.

Función objetivo y obtención de factores El objetivo es maximizar la suma de las distancias al cuadrado de los puntos al origen, ponderadas por su frecuencia marginal, lo que equivale al concepto físico de inercia.

En el caso de la nube de los I puntos-fila en el espacio $\mathbb{R}^J$,

$$\max_{\boldsymbol{u}} \left\{ \sum_i f_{i.} d^2(i, O) \right\} \tag{2.23}$$

donde O es el centro de gravedad definido en la ecuación (2.20). Es decir, maximizar

$$\boldsymbol{u}'\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{F}'\boldsymbol{D}_{\boldsymbol{I}}^{-1}\boldsymbol{F}\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{u} \quad \text{sujeto a} \quad \boldsymbol{u}'\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{u} = 1 \tag{2.24}$$

donde se incluye una restricción de normalización del vector propio $\boldsymbol{u}$. La matriz a diagonalizar es

$$\boldsymbol{S} = \boldsymbol{F}'\boldsymbol{D}_{\boldsymbol{I}}^{-1}\boldsymbol{F}\boldsymbol{D}_{\boldsymbol{J}}^{-1} \tag{2.25}$$

de forma que los ejes factoriales son la solución a

$$\boldsymbol{S}\boldsymbol{u}_\alpha = \lambda_\alpha \boldsymbol{u}_\alpha \tag{2.26}$$

y las coordenadas factoriales

$$\boldsymbol{\psi}_\alpha = \boldsymbol{D}_{\boldsymbol{I}}^{-1}\boldsymbol{F}\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{u}_\alpha \tag{2.27}$$

$$\psi_{\alpha i} = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}f_{.j}} u_{\alpha j} \quad i = 1, ..., I. \tag{2.28}$$

que son centradas y de varianza λ_α.

El problema desde el punto de vista de la nube de los J puntos columna es totalmente simétrico, de forma que las matrices a diagonalizar, ejes y coordenadas factoriales son, respectivamente:

$$\boldsymbol{T} = \boldsymbol{F}\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{F}'\boldsymbol{D}_{\boldsymbol{I}}^{-1} \tag{2.29}$$

$$\boldsymbol{T}\boldsymbol{v}_\alpha = \lambda_\alpha \boldsymbol{v}_\alpha \tag{2.30}$$

$$\boldsymbol{\phi}_\alpha = \boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{F}'\boldsymbol{D}_{\boldsymbol{I}}^{-1}\boldsymbol{v}_\alpha \tag{2.31}$$

$$\phi_{\alpha j} = \sum_{i=1}^{I} \frac{f_{ij}}{f_{i.}f_{.j}} v_{\alpha i} \quad j = 1, ..., J. \tag{2.32}$$

que están también centradas y tienen varianza λ_α.

Relaciones de transición Dado que los valores propios de los análisis de las dos nubes son idénticos, de las ecuaciones (2.25)-(2.30) se extraen las siguientes relaciones de transición entre los vectores propios respectivos:

$$\boldsymbol{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{F}\boldsymbol{D}_{\boldsymbol{J}}^{-1}\boldsymbol{u}_\alpha \tag{2.33}$$

$$\boldsymbol{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{F}'\boldsymbol{D}_{\boldsymbol{I}}^{-1}\boldsymbol{v}_\alpha \tag{2.34}$$

que derivan en las siguientes coordenadas factoriales

$$\boldsymbol{\psi}_\alpha = \sqrt{\lambda_\alpha}\boldsymbol{D_I}^{-1}\boldsymbol{v}_\alpha \quad \Rightarrow \quad \psi_{\alpha i} = \frac{\sqrt{\lambda_\alpha}}{f_{i.}}v_{\alpha i} \tag{2.35}$$

$$\boldsymbol{\phi}_\alpha = \sqrt{\lambda_\alpha}\boldsymbol{D_J}^{-1}\boldsymbol{u}_\alpha \quad \Rightarrow \quad \phi_{\alpha j} = \frac{\sqrt{\lambda_\alpha}}{f_{.j}}u_{\alpha j} \tag{2.36}$$

y en las siguientes relaciones de transición cuasi-baricéntricas:

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{J}\frac{f_{ij}}{f_{i.}}\phi_{\alpha j} \tag{2.37}$$

$$\phi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{i=1}^{I}\frac{f_{ij}}{f_{.j}}\psi_{\alpha i} \tag{2.38}$$

las cuales reflejan que, salvo por el factor $\frac{1}{\sqrt{\lambda_\alpha}}$, la proyección de una modalidad-fila sobre un eje se sitúa en el baricentro de las proyecciones de las modalidades-columna sobre ese mismo eje. Es decir, la coordenada de una modalidad de una variable en un eje es la media ponderada de las proyecciones de las modalidades de la otra variable, donde las ponderaciones son las frecuencias condicionadas del perfil de la modalidad. Se ve *atraída* hacia modalidades con una frecuencia conjunta mayor.

Inercia total e inercia proyectada La inercia total se define como

$$In = \sum_{i=1}^{I} f_{i.}d^2(i,O) = \sum_{j=1}^{J} f_{.j}d^2(j,O) \tag{2.39}$$

y coincide con el estadístico χ^2 de independencia entre dos variables categóricas si se multiplica por el efectivo total de la tabla. Asimismo, se puede descomponer como la suma de los valores propios del AC, $In = \sum_{\alpha}^{\text{mín}\{I,J\}-1}\lambda_\alpha$, lo que permite computar tasas de inercia proyectada, como en ACP.

Ayudas a la interpretación

La interpretación de los ejes se realiza con indicadores numéricos como las contribuciones y los cosenos cuadrado, al igual que en el ACP (página 16). En el caso del AC éstos son:

Contribuciones Se calcula la *contribución de la categoría i al eje α* como

$$Cr_\alpha(i) = \frac{f_{i.}\psi_{\alpha i}^2}{\lambda_\alpha} \tag{2.40}$$

de forma que $\sum_{i=1}^{I} Cr_\alpha(i) = 1$ y, por tanto, las categorías con mayor contribución son más relevantes para la descripción del eje.

De forma totalmente simétrica se computa la *contribución de la modalidad j al eje α* como

$$Cr_\alpha(j) = \frac{f_{.j}\phi_{\alpha j}^2}{\lambda_\alpha} \tag{2.41}$$

que también cumple $\sum_{j=1}^{J} Cr_\alpha(j) = 1$ y se interpreta de igual forma.

Cosenos cuadrados El *coseno cuadrado de la modalidad i sobre el eje α* es (ver ecuación (2.15)):

$$Cos_\alpha^2(i) = \frac{d_\alpha^2(i,O)}{d^2(i,O)} = \frac{\psi_{\alpha i}^2}{d^2(i,O)} \tag{2.42}$$

donde la distancia χ^2 de la modalidad i al origen es, según la ecuación (2.21),

$$d^2(i,O) = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \tag{2.43}$$

y, por construcción, $\sum_\alpha Cos_\alpha^2(i) = 1$.

De forma totalmente simétrica, el *coseno cuadrado de la modalidad j sobre un eje α* es:

$$Cos_\alpha^2(j) = \frac{d_\alpha^2(j,O)}{d^2(j,O)} = \frac{\phi_{\alpha j}^2}{d^2(j,O)} \tag{2.44}$$

y $\sum_\alpha Cos_\alpha^2(j) = 1$.

Proyección de categorías suplementarias La proyección en suplementario en AC consiste en proyectar modalidades que no han intervenido en la obtención de los ejes del análisis y, al igual que en el ACP, proporcionan una ayuda adicional a la interpretación de los ejes.

Denotamos una columna (respectivamente, fila) correspondiente a una categoría suplementaria con el superíndice $^+$. De esta forma, su perfil columna es:

$$\frac{k_{ij}^+}{k_{.j}^+} \quad i = 1, ..., I \tag{2.45}$$

La proyección de esta modalidad-columna en suplementario sobre un eje α se realiza utilizando la relación de transición (2.38)

$$\phi_{\alpha j}^+ = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{I} \frac{k_{ij}^+}{k_{.j}^+} \psi_{\alpha i} \tag{2.46}$$

y de forma simétrica, la proyección de una modalidad-fila sobre un eje α es:

$$\psi_{\alpha i}^{+} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{J} \frac{k_{ij}^{+}}{k_{i.}^{+}} \phi_{\alpha j} \tag{2.47}$$

2.5. El Análisis de Correspondencias Múltiples de variables categóricas

El Análisis de Correspondencias Múltiples (ACM) es una técnica descriptiva que, a diferencia del ACP, estudia las relaciones entre varias variables categóricas y más precisamente, entre sus categorías. Su dominio de aplicación, por tanto, son tablas rectangulares que contienen por filas individuos y por columnas contienen las categorías posibles de las variables. En la práctica, frecuentemente corresponde a respuestas codificadas a las preguntas de una encuesta, según haya sido la escala elegida por los diseñadores de la encuesta.

Una tabla susceptible de ser analizada mediante un ACM contiene la información correspondiente a más de una variable categórica y es una tabla rectangular donde las filas son *individuos* u *observaciones* y las columnas son normalmente variables en codificación condensada con categorías o modalidades representadas por números consecutivos. El objetivo del ACM es encontrar las asociaciones entre variables o entre modalidades de esas mismas variables que revisten mayor importancia en términos de la inercia original. Es una extensión para el análisis de más de dos variables del Análisis de Correspondencias de una tabla que cruza dos variables categóricas.

El germen de dicha técnica se encuentra en Guttman (1941), Burt (1950) y Hayashi (1956) siendo extendido por Benzécri (1973), Escofier (1965) y Masson (1974). También ha sido desarrollado simultáneamente bajo el nombre[2] de *Homogeneity Analysis* por el equipo de Jan de Leeuw, Gifi (1990) y de *Dual Scaling* por Nishisato (1980). Una síntesis de todos ellos está en Tenenhaus & Young (1985).

El ACM consiste en un Análisis de Correspondencias de una tabla denominada *disyuntiva completa* que se obtiene de la tabla que contiene las categorías correspondientes a cada individuo para cada variable transformada en sus variables indicadoras, también llamadas variables ficticias. La tabla que contiene sólo las categorías de las variables, en *codificación condensada*, debe necesariamente transformarse en variables indicadoras para su tratamiento estadístico por ACM. A continuación se muestra lo que puede considerarse un ejemplo.

[2]La denominación de Análisis de Correspondencias Múltiples que finalmente se ha impuesto (Hwang et al. 2006) fue utilizada por primera vez en Lebart (1975).

Denotamos R a la matriz $I \times J$ de variables en su codificación condensada y r_{ij} a la categoría de la variable j escogida por el individuo i. Si k_j es el número de modalidades posibles de la variable j, entonces puede construirse una matriz Z de término general z_{ik} tal que:

$$z_{ik} = \begin{cases} 1 & \text{si el individuo } i \text{ ha escogido la modalidad } k \\ 0 & \text{en caso contrario.} \end{cases} \tag{2.48}$$

donde $k = 1, ..., K$ y $K = \sum_{j=1}^{J} k_j$. En el ejemplo siguiente se tienen $J = 2$ variables con modalidades $\{M, A\}$ para la primera pregunta y $\{N, R, P\}$ para la segunda, de forma que $k_1 = 2$ y $k_2 = 3$. Tras una transformación a valores numéricos de las modalidades, por ejemplo,$\{M, A\} \rightarrow \{1, 2\}$ y $\{N, R, P\} \rightarrow \{1, 2, 3\}$, tendríamos:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 2 \\ 2 & 3 \\ 2 & 1 \\ 1 & 1 \\ 2 & 2 \\ 2 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \implies \boldsymbol{Z} = \begin{array}{c} \begin{matrix} M & A & N & R & P \end{matrix} \\ \left[\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{array}\right] \end{array} \tag{2.49}$$

Donde Z es la yuxtaposición de $J = 2$ subtablas de variables llamadas *indicadoras* de cada una de las variables disponibles.

El ACM consiste en un AC de la tabla disyuntiva completa transformada en sus perfiles-fila y perfiles-columna[3]. Puede considerarse también como una reducción de dimensionalidad similar al ACP en donde la matriz de datos disyuntiva completa es transformada en sus perfiles-fila y perfiles-columna. Las frecuencias absolutas marginales de las filas de la matriz Z de variables indicadoras son $z_{i.} = \sum_{k=1}^{K} z_{ik} = J$ constantes e igual al número de variables, mientras que las frecuencias absolutas marginales de las columnas son $z_{.k} = \sum_{i=1}^{I} z_{ik}$ igual al número de individuos que escogen la modalidad correspondiente a la columna k (de forma que la suma de perfiles-columna para cada variable j es $\sum_{k}^{k_j} z_{.k} = I$ y que el efectivo total de la tabla es $z = \sum_i \sum_k z_{ik} = IJ$). La transformación en perfiles, conjuntamente con la ponderación de las columnas

[3]Otra presentación posible del ACM parte del análisis de la tabla de Burt $Z'Z$ que presenta resultados equivalentes. Ver, p. ej., Lebart et al. (2006, págs. 126-127).

de la tabla disyuntiva completa que se realiza, motivan que la distancia entre individuos no sea la euclidea habitual en ACP sino la distancia χ^2 propia del Análisis de Correspondencias:

$$d^2(i,i') = \frac{1}{J}\sum_{k=1}^{K}\frac{I}{z_{.k}}\left(z_{ik} - z_{i'k}\right)^2 \tag{2.50}$$

en el caso de las filas y

$$d^2(k,k') = \sum_{i=1}^{I} I\left(\frac{z_{ik}}{z_{.k}} - \frac{z_{ik'}}{z_{.k'}}\right)^2 \tag{2.51}$$

para las columnas. De esta forma, dos individuos están próximos si escogen las mismas modalidades y alejados cuanto más difieran en su elección. Asimismo, dos modalidades que han sido escogidas por los mismos individuos están a una distancia nula.

Al igual que en el caso del AC, es posible proyectar en suplementario otras variables, categóricas o continuas. Como en tal caso, suelen ser características socioeconómicas de los individuos.

Principios del ACM. Obtención de los ejes y factores

El ACM aplicado a una tabla disyuntiva completa presentada en (2.49) se basa en los mismos principios del Análisis de Correspondencias aplicado a una tabla de frecuencias. En particular:

- Se realizan las mismas transformaciones en perfiles-fila y perfiles-columna en la tabla de datos.
- Se utiliza el mismo criterio de ajuste con iguales ponderaciones por perfiles.
- Se usa también la distancia χ^2 de las ecuaciones (2.50) y (2.51).

En particular, sea $\boldsymbol{Z}$ la tabla disyuntiva completa correspondiente a una matriz de datos de variables categóricas que contiene a las variables indicadoras z_{ik} definidas en la ecuación (2.48). Denotando como

$$\boldsymbol{F} = \frac{1}{IJ}\boldsymbol{Z}$$

y siendo $\boldsymbol{D}_K$ la matriz diagonal de pesos de las columnas de término general $z_{.k}/IJ$ y $\boldsymbol{D}_I$ la matriz diagonal de pesos de los individuos de término general

$1/I$ generalmente constante, entonces los ejes principales $\boldsymbol{u}_\alpha$ se obtienen de la diagonalización de:

$$\boldsymbol{S} = \boldsymbol{F}'\boldsymbol{D}_I^{-1}\boldsymbol{F}\boldsymbol{D}_K^{-1} = \frac{1}{J}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{D}^{-1} \tag{2.52}$$

donde $\boldsymbol{D}$ es una matriz diagonal de término general $z_{.k}$ y el término general de la matriz $\boldsymbol{S}$ es

$$s_{kk'} = \frac{1}{Jz_{.k'}}\sum_{i=1}^{I} z_{ik}z_{ik'}$$

En el espacio de las columnas $\mathbb{R}^K$ el α-ésimo eje factorial $\boldsymbol{u}_\alpha$ se obtiene resolviendo

$$\frac{1}{J}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{u}_\alpha = \lambda_\alpha \boldsymbol{u}_\alpha \tag{2.53}$$

A su vez, el α-ésimo factor $\boldsymbol{\varphi}_\alpha = \boldsymbol{D}^{-1}\boldsymbol{u}_\alpha$ se obtiene de

$$\frac{1}{J}\boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\varphi}_\alpha = \lambda_\alpha \boldsymbol{\varphi}_\alpha \tag{2.54}$$

Asimismo, por simetría del análisis, puede deducirse que el α-ésimo factor en el espacio de las filas $\mathbb{R}^I$, $\boldsymbol{\psi}_\alpha$, se extrae de la relación:

$$\frac{1}{J}\boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{\psi}_\alpha = \lambda_\alpha \boldsymbol{\psi}_\alpha \tag{2.55}$$

donde $\boldsymbol{\varphi}_\alpha$ y $\boldsymbol{\psi}_\alpha$ contienen respectivamente las coordenadas de los puntos-fila (o individuos) y puntos-variable (o modalidades) sobre el α-ésimo eje factorial.

Relaciones de transición e interpretación de los factores

La simetría de los análisis de puntos-fila y puntos-columna proporciona las llamadas *relaciones de transición* entre los factores en uno y otro espacio. Estas relaciones se pueden expresar como sigue:

$$\boldsymbol{\varphi}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{\psi}_\alpha \tag{2.56}$$

$$\boldsymbol{\psi}_\alpha = \frac{1}{J\sqrt{\lambda_\alpha}}\boldsymbol{Z}\boldsymbol{\varphi}_\alpha \tag{2.57}$$

Asimismo, las coordenadas factoriales para la modalidad k y para el individuo i pueden expresarse:

$$\varphi_{\alpha k} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{i=1}^{I}\frac{z_{ik}}{z_{.k}}\psi_{\alpha i} = \frac{1}{z_{.k}\sqrt{\lambda_\alpha}}\sum_{i\in I(k)}^{I}\psi_{\alpha i} \tag{2.58}$$

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{k=1}^{K}\frac{z_{ik}}{z_{i.}}\varphi_{\alpha k} = \frac{1}{J\sqrt{\lambda_\alpha}}\sum_{k\in K(i)}^{K}\varphi_{\alpha k} \tag{2.59}$$

Estas expresiones establecen respectivamente que, salvo por el factor de escala $\frac{1}{\sqrt{\lambda_\alpha}}$, la modalidad k está en el punto medio de los individuos que la han elegido y que el individuo i está en el punto medio de las modalidades escogidas por él mismo. Estas relaciones son importantes a la hora de interpretar los resultados de un ACM de forma que:

- Si dos individuos están próximos entre sí, eso significa que se parecen en el sentido de que han escogido aproximadamente las mismas modalidades de las variables.

- Si dos modalidades correspondientes a dos variables distintas están próximas entre sí, se entiende que están asociadas en el sentido de que, dado que están en el punto medio de los individuos que las han escogido, han sido escogidas por los mismos individuos o por individuos muy parecidos.

- Si las modalidades que están próximas pertenecen a una misma variable, dado que dichas modalidades son mutuamente excluyentes y no pueden estar asociadas, eso indica semejanza entre los grupos de individuos que las han elegido.

Inercia de una modalidad, de una variable e inercia total

La definición de los ejes y factores de un ACM que se obtienen de la diagonalización de (2.52) responden, en realidad, a la maximización de la inercia total medida como

$$\max_{\boldsymbol{u}} \sum_i f_{i.} d^2(i, G) \tag{2.60}$$

esto es, la suma, ponderada por sus perfiles, de las distancias al cuadrado de todos los puntos-individuo al centro de gravedad de la nube (de puntos individuo).

Desde el punto de vista de las modalidades, el centro de gravedad G es un vector compuesto por los valores $f_{i.} = 1/I$ y la distancia cuadrática χ^2 entre la modalidad k y el centro se define:

$$d^2(k, G) = I \sum_{i=1}^{I} \left(\frac{z_{ik}}{z_{.k}} - \frac{1}{I} \right)^2 = \frac{I}{z_{.k}} - 1 \tag{2.61}$$

siendo mayor cuanto menor sea el número de individuos que la han escogido.

Inercia de una modalidad o categoría La inercia de la categoría k-ésima, de forma análoga a la descrita en (2.60) es

$$In(k) = f_{.k}\, d^2(k, G) = \frac{1}{J}\left(1 - \frac{z_{.k}}{I}\right) \tag{2.62}$$

Se puede observar que esta inercia es inversamente proporcional al número de individuos que han escogido la categoría siendo el peso $f_{.k} = \frac{z_{.k}}{IJ}$. El máximo $1/J$ se alcanza cuando el efectivo es nulo por lo que, generalmente, conviene evitar las modalidades de efectivo débil.

Inercia de una variable Para la variable j-ésima, es:

$$In(j) = \sum_{k=1}^{k_j} In(k) = \frac{1}{J}(k_j - 1) \tag{2.63}$$

siendo directamente proporcional al número de categorías asociadas a la variable. El mínimo es $1/J$ y corresponde a una división en $k_j = 2$ modalidades. Es, por tanto, deseable que el número de categorías de las variables categóricas sea similar a lo largo de toda la tabla.

Inercia total Finalmente, la inercia total es

$$In = \sum_{j} In(j) = \sum_{k=1}^{K} \frac{z_{.k}}{IJ}\, d^2(k, G) = \frac{K}{J} - 1 \tag{2.64}$$

que vale 1 cuando $K = 2J$ y todas las variables están compuestas de dos categorías. En ACM la inercia total no es una magnitud interesante, ya que únicamente depende del número de variables y de modalidades y en ningún caso de las relaciones entre las variables.

Tasas de inercia y correcciones

En los análisis factoriales es habitual utilizar tasas de inercia como medidas de la calidad de la representación de los ejes factoriales de forma similar a como se usan, por ejemplo, medidas de bondad de ajuste en el análisis de regresión. Estas tasas de inercia se definen como porcentajes de inercia proyectada (definida a partir de los valores propios obtenidos de la diagonalización correspondiente al método empleado) sobre la inercia total.

En el ACM, como se ha expuesto anteriormente, la inercia total es una magnitud de escaso significado estadístico, por lo que es natural sospechar que

las tasas de inercia mencionadas no tengan el mismo valor que en un ACP o en un AC.

Un análisis comparativo de las tasas de inercia de un AC y un ACM en el único caso en que son directamente comparables, como es aquél en que se dispone de $J = 2$ variables categóricas, permite demostrar que en ACM se obtiene un mayor número de ejes que en AC. Además, las tasas de inercia por eje son significativamente menores en el caso del ACM (con respecto al AC), incluso sobre el mismo conjunto de datos. Un ejemplo puede verse en Fernández-Aguirre & Modroño (2008, págs. 172-173).

En la práctica, el uso de tasas de inercia proyectada sobre los primeros factores de un ACM, realizado sobre una tabla disyuntiva completa, proporciona una medida excesivamente pesimista de la calidad de la representación. Esto es debido en parte a que una tabla disyuntiva completa está compuesta por variables indicadoras ortogonales entre sí dentro de cada subtabla y, por tanto, con nula relación desde el punto de vista matemático.

Existen varias medidas correctoras de las tasas de inercia proyectada en ACM. La primera es la corrección de Benzécri (1979). Esta corrección se basa en igualar las tasas de inercia del ACM a las del AC cuando se considera una tabla de dos variables categóricas, como se ha mencionado anteriormente, y ajustar los valores propios consecuentemente. De esta forma, se propone calcular tasas de inercia en ACM en base a los valores propios corregidos λ^*_α que se obtienen de:

$$\lambda^*_\alpha = \left(\frac{J}{J-1}\right)^2 \left(\lambda_\alpha - \frac{1}{J}\right)^2 \qquad \forall\, \lambda_\alpha > \frac{1}{J} \tag{2.65}$$

en vez de usar los valores propios λ_α que se obtienen directamente de la diagonalización de la matriz correspondiente de la ecuación (2.52).

La segunda corrección, propuesta por Greenacre (1993) en el marco del *Joint Correspondence Analysis*, considera que la corrección de Benzécri es, por el contrario, una medida demasiado optimista. Esto es así porque al aplicar la corrección a los valores propios mayores que el inverso del número de variables, la inercia tomada como total se reduce, de forma que se produce una sobreestimación de las tasas de inercia proyectadas sobre los factores. Este autor, propone dividir cada valor propio corregido en el sentido de Benzécri por la inercia media de los bloques, excluidos los correspondientes a la diagonal principal de la tabla de Burt:

$$\overline{\mathcal{I}} = \frac{J}{J-1}\left(\sum_\alpha \lambda^2_\alpha - \frac{K-J}{J}\right)^2 \tag{2.66}$$

de forma que la tasa de inercia corregida provendría de la expresión:

$$\tau_\alpha = \frac{\lambda^*_\alpha}{\overline{\mathcal{I}}} \tag{2.67}$$

En cualquier caso, es probable que cualquiera de las dos opciones elegidas proporcione una medida más adecuada que la proporción de inercia medida a través de los valores propios originales λ_α para evaluar la calidad de representación de los primeros factores.

Ayudas a la interpretación

Las ayudas a la interpretación se componen de contribuciones y cosenos cuadrado, tomados del Análisis de Correspondencias (AC).

Contribuciones En el caso del ACM, los individuos suelen ser anónimos en cuyo caso las contribuciones de los individuos carecen de interés, no así las de las modalidades y las de las variables.

Se calcula la *contribución de la categoría k al eje α* como

$$Cr_\alpha(k) = \frac{f_{.k}\psi^2_{\alpha k}}{\lambda_\alpha} = \frac{z_{.k}\psi^2_{\alpha k}}{IJ\lambda_\alpha} \tag{2.68}$$

de forma que $\sum_{k=1}^{K} Cr_\alpha(k) = 1$ y, por tanto, las categorías con mayor contribución son más relevantes para la descripción del eje.

Adicionalmente, es posible computar la *contribución de la variable j al eje α* como

$$Cr_\alpha(j) = \sum_{k \in K(j)} Cr_\alpha(k) \tag{2.69}$$

que se interpreta de manera similar pero respecto de la variable categórica j-ésima.

Cosenos cuadrados Al igual que en el caso de las contribuciones, en ACM, al ser los individuos anónimos, el interés se centra en las modalidades.

El *coseno cuadrado de la categoría k-ésima repecto al eje α* o contribución relativa del factor α a la posición de la modalidad k-ésima se define como

$$Cos^2_\alpha(k) = \frac{\varphi^2_{\alpha k}}{d^2(k, G)} \tag{2.70}$$

donde la distancia cuadrática al centro de gravedad de la modalidad k es la expresada en la ecuación (2.61). Se cumple que $\sum_\alpha Cos^2_\alpha(k) = 1$.

No existe una calidad de representación de una variable, puesto que no tiene un representación específica sobre un eje, más allá de la compuesta por sus modalidades.

Proyección de elementos suplementarios

En ACM los elementos suplementarios son útiles si la muestra utilizada contiene individuos que están agrupados de alguna manera en relación a las variables no consideradas como activas.

En particular, cuando se analiza una tabla disyuntiva completa, la consideración de tales elementos suplementarios permite:

- Enriquecer la interpretación de los ejes con otras variables diferentes, proyectando en el espacio de las variables los centros de gravedad de los grupos de individuos definidos por las variables suplementarias (cuando éstas sean también categóricas).
- Al proyectar las variables suplementarias en el espacio de los individuos, se consigue un efecto de *predicción* (en el sentido del análisis de regresión), donde quedarían explicadas por las variables activas. Cuando se proyectan los individuos suplementarios (en el espacio de las variables) el objetivo es la comparación con los individuos activos, más relacionado con una posible *discriminación*.

Las variables a proyectar pueden ser tanto categóricas como continuas.

Proyección de categorías suplementarias La proyección en suplementario de modalidades de una variable categórica viene dada por la relación de transición (2.58):

$$\varphi_{\alpha k^+} = \frac{1}{z_{.k^+}\sqrt{\lambda_\alpha}} \sum_{i \in I(k^+)} \psi_{\alpha i} \tag{2.71}$$

que es, simplemente, la media aritmética de las proyecciones sobre el eje α de los individuos que han elegido la modalidad suplementaria k^+, multiplicada por el factor $1/\sqrt{\lambda_\alpha}$.

Proyección de variables continuas suplementarias El cálculo de la coordenada de una variable continua suplementaria sobre un eje factorial viene dado por el coeficiente de correlación entre la variable suplementaria y el factor, como en un ACP normado. Si se elevan esas coordenadas al cuadrado, se obtienen los cosenos cuadrado.

De esta manera, la proyección de una variable sobre el plano indica la dirección hacia donde se sitúan los valores más elevados de la misma. Por ello, será conveniente que esté próxima al círculo de correlaciones de radio unitario. Sin embargo, probablemente sea preferible la transformación de la variable continua en categórica de cara a la proyección en suplementario al tener la proyección una interpretación similar a la del resto de categorías.

Valoración de la proyección en suplementario. Valores-test y Bootstrap parcial de modalidades suplementarias

Valores-test Los valores test constituyen una herramienta de ayuda a la interpretación de los ejes basada en la significación relativa de las proyecciones de las modalidades suplementarias.

Una modalidad suplementaria k^+ concierne a I_{k^+} individuos ($I_{k^+} = z_{.k^+}$). Bajo la hipótesis nula H_0 de que esos I_{k^+} individuos son extraídos al azar entre los I individuos analizados (extracción que se supone sin reposición), la media de las I_{k^+} coordenadas extraídas al azar en el conjunto finito de los I valores $\psi_{\alpha i}$ es una variable aleatoria $X_{\alpha k^+}$:

$$X_{\alpha k^+} = \frac{1}{I_{k^+}} \sum_{i \in I(k^+)} \psi_{\alpha i}$$

es decir, $X_{\alpha k^+}$ es la media aritmética de las coordenadas $\psi_{\alpha i}$ de los individuos que han elegido la modalidad k^+-ésima. Su valor medio o esperanza matemática es:

$$E(X_{\alpha k^+}) = 0$$

y la varianza:

$$Var_{H_0}(X_{\alpha k^+}) = \frac{I - I_{k^+}}{I - 1} \frac{\lambda_\alpha}{I_{k^+}}$$

La coordenada $\varphi_{\alpha k^+}$ de la modalidad suplementaria está relacionada con la variable aleatoria $X_{\alpha k^+}$ mediante la expresión:

$$\varphi_{\alpha k^+} = \frac{1}{\sqrt{\lambda_\alpha}} X_{\alpha k^+}$$

Se tiene por tanto:

$$E[\varphi_{\alpha k^+}] = 0$$

$$Var(\varphi_{\alpha k^+}) = \frac{1}{\lambda_\alpha} Var(X_{\alpha k^+}) = \frac{I - I_{k^+}}{I - 1} \frac{1}{I_{k^+}}$$

Tipificando,

$$t_{\alpha k^+} = \frac{\varphi_{\alpha k^+} - 0}{\sqrt{\frac{I - I_{k^+}}{I-1}} \frac{1}{I_{k^+}}} = \sqrt{I_j \frac{I-1}{I - I_{k^+}}}\, \varphi_{\alpha k^+}$$

La cantidad $t_{\alpha k^+}$ mide, en número de desviaciones típicas, la distancia entre la modalidad k^+ (el cuasi-baricentro de los n_j individuos) y el centro de gravedad sobre el eje factorial α. A esta cantidad se le llama "valor test". Aplicando el

teorema central del límite, su distribución converge en distribución a una ley normal de Laplace-Gauss centrada y reducida, es decir,

$$t_{\alpha k^+} \xrightarrow{d} N(0,1)$$

El valor-test es un criterio que permite apreciar rápidamente si una modalidad tiene una posición "significativa" sobre un eje. Se considera generalmente ocupando una "posición significativa" a las modalidades con valores test superiores en valor absoluto al cuantil $1-\alpha$ de la distribución normal reducida que es 1,96 para un nivel del 5 %. Cuando se dispone de un número importante de modalidades suplementarias, los valores test permiten apreciar rápidamente las modalidades útiles a la interpretación de un eje o de un plano factorial.

El cálculo simultáneo de varios valores-test o de varios umbrales de probabilidad tropieza con el escollo de la inferencia simultánea, por lo que es conveniente considerar un umbral superior al anterior. En cualquier caso, los valores test permiten sobre todo clasificar las modalidades suplementarias por orden de interés decreciente, lo que constituye una ayuda apreciable a la interpretación de resultados.

Finalmente, hay que señalar que los valores-test no tienen estrictamente sentido más que para las modalidades suplementarias. Las coordenadas sobre un eje de los individuos correspondientes a una modalidad activa no pueden ser considerados como obtenidas al azar sobre el mismo, ya que esta modalidad ha contribuido a la construcción del eje.

Validación. Bootstrap parcial La técnica de validación bootstrap se debe originalmente a Efron (1979) y consiste en simular un número elevado de veces los elementos de una muestra de tamaño I mediante muestreo aleatorio repetido con reemplazamiento. Se crean M muestras, todas ellas del mismo tamaño que la muestra original. Cada elemento de la muestra original tiene la misma probabilidad $1/I$ de aparecer en cada *extracción*. En el seno de cada muestra obtenida, o muestra bootstrap, dado que las extracciones son con reemplazamiento, un elemento dado puede aparecer varias veces o ninguna.

Las M réplicas de la muestra original se usan para estudiar la estabilidad de cualquier estadístico que pueda calcularse a través de ellas mediante la construcción de intervalos o regiones de confianza para el mismo. Efron demostró que, cuando un estadístico sigue una distribución teórica conocida, los intervalos de confianza teóricos y los intervalos bootstrap son, aproximadamente de la misma amplitud.

En análisis multivariante, el muestreo bootstrap generalmente se utiliza de forma que la muestra a replicar es la tabla original de datos y lo que se replica M veces es dicha tabla. Los elementos de la muestra, que son extraídos con

reemplazamiento cada vez en cada réplica, son generalmente las filas de la matriz que en el caso del ACM se corresponden con los individuos de la tabla de datos.

La muestra bootstrap generada puede ser utilizada para diferentes objetivos. Puede ser utilizada para analizar la estabilidad de la estructura factorial completa, que incluye los valores propios, porcentajes de inercia, factores, ejes unitarios, etc., a lo que se denomina bootstrap total. A continuación, pueden construirse intervalos o regiones de confianza para todos los elementos mencionados. Para ello, se realiza un análisis factorial completo (ACP o ACM , por ejemplo) y, según el caso, sobre cada tabla replicada m se examina la variabilidad de cada elemento de interés (valores propios, ejes,...). Este bootstrap tiene importantes complicaciones prácticas, puesto que en los análisis factoriales la orientación de los ejes es arbitraria y su sentido se intercambia constantemente de unas réplicas a otras.

Existe otro uso del bootstrap, denominado boostrap parcial, sugerido inicialmente por Chateau & Lebart (1996). En él, se conservan los elementos principales del análisis factorial, como son los valores propios y los ejes, y se proyectan en suplementario, siempre sobre esos mismos ejes, los elementos deseados a partir de las muestras replicadas. De esta manera, el estudio se reduce a la estabilidad de las configuraciones, que es lo más utilizado en la práctica. La idea subyacente es que el recálculo de los ejes que se hace en cada una de las réplicas del bootstrap total puede constituir una variabilidad excesiva en dichos ejes, dado que en una muestra bootstrap la probabilidad de que una fila dada aparezca más de una vez es probablemente excesiva, lo mismo que la de que no aparezca. En concreto, Daudin et al. (1988) estiman como 0,26 y 0,37 dichos valores.

La visualización de los elementos suplementarios así proyectados suele realizarse mediante su proyección sobre los mismos ejes principales y planos bidimensionales originales. Sobre esos planos se dibujan un número muy elevado de puntos (m puntos para cada elemento a validar), por lo que se suelen englobar encerrándolos en conjuntos convexos, que son posteriormente *pelados*, o en elipses de confianza $(1-\alpha)$, que dejan fuera una proporción determinada α de réplicas para cada elemento validable. Una referencia básica en este tipo de técnicas es Markus (1994).

Un conjunto convexo se puede definir de varias maneras. La más extendida es la siguiente: es el polígono convexo más pequeño que encierra un conjunto de puntos. La noción de *pelado* se refiere a la eliminación secuencial de los vértices de ese conjunto hasta que quede un porcentaje de puntos igual al nivel de confianza requerido. Las elipses de confianza pueden dibujarse de varias formas. La más básica se basa, por una lado, en el supuesto de distribución

normal multivariante, a partir de cuya matriz de covarianzas pueden obtenerse los ejes principales de la elipse y, por otro lado, en la igualdad distribucional de los factores considerados. En cualquier caso, en términos prácticos, la función de ambos es puramente indicativa, y no es probable obtener resultados significativamente diferentes.

A este respecto, Michailidis & de Leeuw (1998) establecen que la consecución de unos resultados de análisis de estabilidad mediante bootstrap válidos requieren un número de muestras bootstrap elevado, mayores que 1000. También recomiendan evitar el uso de categorías de efectivo débil, amalgamando categorías próximas si fuera necesario, hecho a tener en cuenta cuando los datos disponibles contienen categorías extremas de efectivo reducido. Recientemente, Lebart (2006), en base a una dilatada experiencia, afirma que el número de muestras bootstrap necesarias para que los resultados sean fiables a tan sólo unas pocas decenas, al menos para el caso del bootstrap parcial.

2.6. Análisis Factorial Múltiple

Las herramientas exploratorias de los apartados anteriores tienen en común que se trata de técnicas multivariantes de análisis de tablas rectangulares de datos $\times$ variables, pudiendo ser estas variables de diversos tipos. El análisis se centra en una reducción de la dimensionalidad de las tablas, donde las variables tienen una importancia a priori similar sobre los resultados.

El Análisis Factorial Múltiple (AFM) es una técnica conexa a las anteriores que tiene sentido cuando las variables de la tabla se estructuran en subtablas por alguna característica común a dichas subtablas. Pueden ser variables que tengan en común características sensoriales, geográficas, temporales o de otro tipo. Algunas veces esas características pueden ser tenidas en cuenta mediante una variable cualitativa que puede ser proyectada en suplementario. Sin embargo, esa proyección en suplementario obliga a que tal diferenciación no tome parte en la formación de los ejes principales, objeto de la reducción de dimensionalidad que llevan a cabo estas técnicas.

El AFM es parte de una familia de técnicas denominadas de tablas múltiples. Entre ellas tiene la virtud de equilibrar las subtablas a analizar de una determinada manera para evitar la excesiva influencia de unas subtablas sobre otras. Tanto el AFM, como otras técnicas relacionadas, permiten que la diferenciación entre subtablas forme parte de la extracción de los ejes principales de la tabla total, compuesta por la yuxtaposición de todas las subtablas, como si fuese una única tabla rectangular habitual.

El AFM será descrito con detalle en los capítulos 4 y 5 en dos contextos diferentes. En el primero de ellos se examinará su uso en un caso de dificul-

tad técnica en el que las subtablas contienen diferente número de individuos mientras que en el segundo se analizará el caso en el que las subtablas son de diferentes tipos, mezclando variables cuantitativas, cualitativas y frecuencias, donde las frecuencias provienen de las palabras empleadas en un texto que, además, aparece en varios idiomas diferentes. Es por esto que la exposición teórica del método se deja para estos capítulos, haciendo especial incidencia en cada uno de ellos en los aspectos más relevantes de la aplicación que se pretende llevar a cabo.

2.7. Clasificación sobre los factores

Las técnicas de clasificación automática o *clustering* tienen como objetivo el agrupamiento de líneas o de columnas de una tabla de datos. Generalmente, la tabla de datos está diseñada en la forma individuos $\times$ variables y en este caso el agrupamiento se produce sobre las filas, que contienen individuos. Cuando se trate de una tabla de contingencia, el agrupamiento puede producirse sobre los elementos de cualquiera de sus dos dimensiones. El número de grupos en los que se divide el total de la muestra conforma el cardinal de la partición que se obtiene, y a cada grupo generado se le suele denominar clase o cluster.

Tipos de métodos de clasificación Existen dos grandes familias o tipos de clasificación: la clasificación jerárquica basada generalmente en un algoritmo ascendente aglomerativo y la clasificación basada en un algoritmo iterativo de agregación en torno a centros móviles o K-means.

En el primer caso se genera una jerarquía de particiones en forma de árbol que, junto con la observación de los índices de nivel de la jerarquía, da una idea del número de clases homogéneas presentes en los elementos que tratamos de clasificar.

El método de agregación en torno a centros móviles puede usarse en combinación con el método jerárquico para refinar la partición obtenida por este último. Así, se consideran los centros de gravedad de las clases obtenidas mediante el método jerárquico y se reasignan algunos elementos por medio de una clasificación de tipo K-means.

En general, en la práctica se ha demostrado que la utilización conjunta de las dos grandes familias de métodos exploratorios factoriales y de clasificación produce resultados satisfactorios desde el punto de vista de la extracción del conocimiento de grandes tablas de datos.

Criterio de clasificación Existen múltiples distancias susceptibles de ser usadas en un análisis de clasificación como criterio de agrupamiento jerárquico.

Una particularmente interesante, que se basa en la descomposición de la inercia total en la suma de la inercia intra grupos y la inercia inter grupos (usando la descomposición de Huygens), es la del método de Ward generalizado:

$$\Delta In(a,b) = \frac{p_a p_b}{p_a + p_b}\, d^2(g_a, g_b) \tag{2.72}$$

donde a, b son dos elementos a agrupar (inicialmente individuos y eventualmente pasan a ser clases de los mismos), p_a, p_b sus pesos o masas y g_a, g_b los centros de gravedad de sus clases. El criterio de Ward pretende minimizar la inercia (o distancia corregida por los pesos) entre los elementos componentes de un grupo a la vez que se maximiza la inercia entre los grupos existentes. La ecuación (2.72) representa el incremento de la inercia intra-grupos al unir los grupos a y b en uno solo. La aplicación del criterio de Ward consiste en unir en cada paso de la clasificación jerárquica los dos grupos que incrementen lo menos posible la inercia intra-grupos.

Este criterio cumple la interesante propiedad de que la suma de los índices de nivel de la jerarquía es igual a la inercia total de la nube, véase Lebart et al. (2006). Si este criterio se utiliza no sobre la nube de individuos original, sino sobre la proyección de ésta sobre los factores principales de algún método factorial (i.e., clasificación sobre factores) la suma de los índices de nivel equivale a la suma de los valores propios asociados a esos mismos factores (Lebart 1994, Lebart et al. 2006) y permite elegir la partición de los individuos más adecuada en base a ellos.

Procedimiento de clasificación. Algoritmo Normalmente se parte de una clasificación inicial en k clases donde k es un número arbitrario de clases menor o igual al número de individuos disponible pero mayor al número deseable y/o lógico de clases que se prevé obtener al final. Sobre la partición inicial se conduce una clasificación jerárquica por el método de Ward (u otro) utilizando la expresión (2.72) con la construcción del correspondiente árbol de agregación. En base al dendrograma, o al histograma de índices de nivel de agregación, se selecciona una partición final allí donde el incremento de la inercia intra grupos sea *elevado*. Como último paso, y por el método de centros móviles, se reasignan los elementos de las clases a las mismas hasta llegar a la estabilidad de las clases, en la denominada fase de consolidación.

Descripción de las clases La descripción estadística de las clases obtenidas en la clasificación es muy importante para la interpretación de los clusters. Si las variables sobre las que se realiza la clasificación son continuas, el objetivo es comparar la media total de una variable X, $\overline{X}$, con la media de la variable

dentro de la clase k, $\overline{X}_k$. Se forma para ello el siguiente valor test:

$$t_k(X) = \frac{\overline{X}_k - \overline{X}}{s_k(X)} \quad \text{con} \quad s_k^2(X) = \frac{(n-n_k)s^2(X)}{(n-1)n_k} \tag{2.73}$$

En el caso de que la variable X sea suplementaria, y bajo la hipótesis nula de que los individuos en la clase k han sido asignados a esa clase de forma completamente aleatoria, la media de la clase $\overline{X}_k$ tiene como media y varianza totales a los valores de $\overline{X}$ y $s^2(X)$ respectivamente, y el valor test sigue asintóticamente la ley normal reducida.

Si la variable X es activa y no suplementaria, lo anterior deja de ser cierto, y los valores test sólo pueden usarse como medidas de similaridad entre variables y clases.

En el caso de que las variables sean categóricas, el valor test se obtiene de forma aproximada a partir de la diferencia entre el porcentaje de individuos de una clase k que presenta una modalidad j y el porcentaje de individuos del total que presentan esa modalidad j. Bajo la hipótesis de que son iguales, el valor test sigue una ley hipergeométrica que en la práctica se aproxima por la normal reducida, lo que permite ordenar las modalidades características de cada clase. Como en el caso de las variables continuas, esta interpretación sólo es cierta para las variables suplementarias.

Complementariedad de los análisis factoriales y de clasificación Los análisis factoriales de las secciones 2.3-2.5 anteriores se ven frecuentemente complementados con una posterior clasificación sobre los principales factores.

Los análisis factoriales por sí solos plantean algunas debilidades. Entre ellas, la dificultad de interpretación de ejes de orden superior a los dos o tres primeros cuando hay un número de ejes retenidos claramente superior, la existencia de algunos individuos o perfiles fila excesivamente influyentes en la determinación de los ejes principales o la dificultad de interpretar planos principales con un número muy elevado de puntos.

Si el análisis factorial utilizado es completado con una clasificación sobre los factores principales retenidos, el resultado puede ayudar a superar las debilidades anteriores. Por ejemplo, es posible que alguna clase resultante resulte característica de algún factor de orden superior de interpretación de otra manera difícil. Además, los algoritmos de clasificación son poco sensibles a observaciones aisladas. Finalmente, sobre la partición obtenida se puede obtener una descripción automática de las clases que facilita la interpretación de los ejes sobre los cuales están mejor representadas, incluso aunque se trate de una clase muy numerosa (que se suele representar gráficamente por un sólo punto, su centro de gravedad).

Así, el procedimiento combinado de análisis factorial-clasificación suele constar de las fases siguientes. Un análisis factorial que reduce el número de variables y las cambia por factores, una clasificación sobre los principales factores retenidos, que no deben ser ni excesivos ni insuficientes, la descripción de las clases obtenidas a partir de las variables disponibles, suplementarias principalmente, y el posicionamiento de las clases en los planos factoriales principales. Este permite evaluar la densidad interna de las clases, la distancia entre ellas e incluso dibujar trayectorias cuando las clases tienen una interpretación ordinal.

2.8. Herramientas confirmatorias o predictivas de data mining para el análisis de grandes encuestas

El análisis de encuestas no tiene por qué ser realizado o enfocado necesariamente mediante herramientas de análisis de tipo exploratorio, en el que típicamente la relación entre las variables es considerada de manera simétrica. Técnicas de corte inferencial o confirmatorio (Lebart et al. 2006), también denominadas como de modelización predictiva por Hand et al. (2001), permiten formular hipótesis, validarlas o no, extrapolar los resultados a la población a la que la muestra pertenece y predecir. Todo ello haciendo uso de elementos basados en la Estadística Inferencial, como el Análisis de Regresión, Regresión Logística, Análisis Discriminante, Árboles de Decisión y otros. Como dicen, entre otros, estos autores, no hay que entender las técnicas exploratorias y confirmatorias como técnicas alternativas excluyentes, sino más bien como complementarias. Las técnicas exploratorias tienen gran importancia en un análisis preliminar de depuración de los datos, eliminando, por ejemplo, variables redundantes o de escasa importancia.

En este capítulo se van a considerar dos técnicas de análisis confirmatorio o predictivo: el *PLS path modelling* o modelización en Mínimos Cuadrados Parciales y los modelos Logit. Los modelos de mínimos cuadrados parciales presentan y tratan de estimar relaciones entre variables latentes, no observadas, que se obtienen, bien a partir de grupos de variables observadas, por ejemplo, a partir de las preguntas de una encuesta, o bien a partir de otras variables latentes. Dichas variables latentes recogen variables que pueden ser difíciles de medir, y son de uso generalizado en Sociología, Psicología y Marketing. Por otro lado, los modelos Logit son un tipo de modelos de elección discreta, que tratan de establecer la relación de dependencia de una variable discreta respecto de otras variables explicativas. Estos modelos se han utilizado con éxito en Bioestadística y en Economía, en particular, para modelizar

y analizar las decisiones de los consumidores tanto en Microeconometría como en Marketing.

2.9. PLS path modelling

La técnica de análisis conocida como *Partial Least Squares (PLS) path modelling* se debe a Wold (1979). Podría traducirse al castellano como Modelización de sendas por Mínimos Cuadrados Parciales. El algoritmo de estimación correspondiente apareció en Wold (1982, 1985*a*). Es una técnica de modelización de ecuaciones estructurales (Jöreskog 1973) de relaciones entre variables latentes, que suelen presentarse gráficamente mediante *path diagrams* o diagramas de flechas.

Las variables latentes son definidas en Estadística como variables no directamente observables, sino construidas a partir de otras variables que son observadas y medidas de una forma directa. Se utilizan cuando existen variables que no pueden medirse de forma directa (como la inteligencia) o variables que corresponden a conceptos abstractos como la calidad de vida, la confianza de los consumidores o la actitud hacia una cierta marca comercial. Generalmente las variables observables (o manifiestas, en la terminología PLS) que se usan para construir dichas variables latentes son abundantes, de manera que en el proceso de construcción de estas últimas se da un proceso evidente de reducción de dimensionalidad de los datos. A las variables latentes también se les suele denominar constructos (del inglés, *constructs*).

El término de modelización de ecuaciones estructurales generalmente se refiere a los denominados métodos de la matriz de covarianzas de Jöreskog (1973) o *Structural Equation Modelling* (SEM). Estos métodos comienzan mediante la especificación de un *path model* donde se definen las variables latentes a partir de las variables observables y las múltiples relaciones entre las variables latentes. A partir de ahí se establece un modelo que consiste en un sistema de ecuaciones entre las variables latentes cuya estimación se obtiene a partir de la minimización de una función de discrepancia. Ésta se establece entre la matriz de covarianzas muestrales de las variables observadas y la de las covarianzas estimadas a través del modelo, es decir, teniendo en cuenta las restricciones impuestas por las relaciones entre las variables latentes del modelo. Típicamente esta matriz se obtiene por un método de estimación máximo-verosímil. Esta metodología es de tipo confirmatorio y requiere supuestos distribucionales (normalidad multivariante de las variables observables, para la estimación por máxima verosimilitud) e independencia entre las observaciones. Este tipo de técnicas a veces son más conocidas por el nombre del software utilizado; los más conocidos son Lisrel, Eqs, Amos, Sepath, M-plus y Calis (SAS).

El método PLS path modelling es también una modelización en variables latentes obtenidas a partir de variables indicadoras observables. Como características más ventajosas caben destacar la escasez de requerimientos sobre las escalas de medida, el tamaño muestral y la distribución de los términos de error (Wold 1985*b*). Tiene ventajas sobre los modelos estructurales de estimación basada en la matriz de covarianzas, que en ocasiones proporcionan soluciones no admisibles y presentan dificultades de identificación (Fornell & Bookstein 1982). Una comparación entre ambos métodos, enfocada a la práctica, mostrando situaciones en que se espera obtener resultados parecidos con los dos métodos puede encontrarse en Tenenhaus et al. (2005).

2.9.1. Especificación del modelo

Un modelo PLS está definido por dos sistemas de ecuaciones. El primero es un sistema estructural similar al de los métodos SEM, que se denomina modelo estructural o modelo interno en PLS. El modelo interno especifica las relaciones existentes (o que se quieren comprobar si existen) entre las variables latentes

$$\boldsymbol{\eta} = \boldsymbol{\beta\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\nu} \tag{2.74}$$

donde $\boldsymbol{\eta}$ es un vector de variables latentes endógenas, $\boldsymbol{\xi}$ un vector de variables latentes exógenas y $\boldsymbol{\nu}$ un vector de términos de error. $\boldsymbol{\beta}$ y Γ son dos matrices de parámetros (o *path coefficients*), respectivamente. Esta forma estructural puede representarse mediante un diagrama de sendas o flechas *(path diagram)*. El modelo es una cadena causal, de manera que las relaciones entre las variables latentes son unidireccionales, nunca de ida y vuelta.

El sistema estructural (2.74) puede escribirse de forma reducida así:

$$\boldsymbol{\eta} = (\boldsymbol{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma\xi} + (\boldsymbol{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\nu} = \boldsymbol{\beta^*\xi} + \boldsymbol{\nu^*} \tag{2.75}$$

donde β^* contienen los efectos finales de las variables latentes exógenas. Se supone que $E(\nu_j|\eta_i, \xi_h) = Cov(\nu_j, \eta_i) = Cov(\nu_j, \xi_h) = 0$.

El segundo sistema de ecuaciones, que veremos a continuación, contiene ecuaciones de medida y se denomina modelo de medida o modelo externo en la metodología PLS. El modelo externo contiene las relaciones de las variables observadas, indicadoras o manifiestas con las variables latentes.

Para el modelo de medida existen varias posibles especificaciones atendiendo a la relación entre las variables manifiestas y las variables latentes:

1. *Reflexiva.* En este caso, cada variable manifiesta se relaciona con una variable latente mediante una relación del tipo:

$$\boldsymbol{x} = \boldsymbol{\Pi_x\xi} + \boldsymbol{\varepsilon_x} \tag{2.76}$$

$$\boldsymbol{y} = \boldsymbol{\Pi_y\eta} + \boldsymbol{\varepsilon_y} \tag{2.77}$$

donde las variables latentes están tipificadas (o al menos con desviación típica 1, no necesariamente centradas). Los errores de medida ε tienen media cero y están incorrelacionados con las variables latentes a las que corresponden. Los coeficientes π_h no son más que coeficientes de regresión.

Esta especificación no requiere necesariamente que exista una sola variable manifiesta por cada variable latente. Si existen varias variables manifiestas para una sola latente, éstas deben medir el mismo fenómeno. Tal bloque de variables indicadoras debe de ser unidimensional, en el sentido, por ejemplo, del Análisis Factorial. Cuando esta característica se cumple, se dice que el bloque de variables indicadoras es consistente internamente. Este tipo de relación es frecuente cuando se miden variables actitudinales. También se supone que las variables indicadoras están correlacionadas positivamente; si no es así, se ajustan en ese sentido, sin pérdida de generalidad.

Existen varias maneras de comprobar que un bloque de variables manifiestas es unidimensional (Tenenhaus et al. 2005):

- Mediante un análisis de componentes principales del bloque. El bloque es unidimensional si el primer valor propio (de un análisis normado) es mayor que 1 y el segundo menor que 1.
- El coeficiente α de Cronbach. Si tenemos $x_h, h = 1, ..., p$ variables manifiestas para una misma variable latente, entonces,

$$\alpha = \frac{\sum_{h \neq h'} cov(x_h, x_{h'})}{var(\sum_h x_h)} \times \frac{p}{p-1} \tag{2.78}$$

  de forma que un bloque se considera unidimensional si $\alpha > 0{,}7$.
- El coeficiente ρ de Dillon-Goldstein. La idea se basa en que basta con que la relación entre las variables latentes y sus indicadoras sea alta. Dada la expresión (2.76) y que las variables están tipificadas, basta con que los coeficientes π_h de $\mathbf{\Pi}$ sean grandes.

  Si la relación entre una variable manifiesta y su latente se escribe así:

$$x_h = \pi_{ho} + \pi_h \xi + \varepsilon_h \tag{2.79}$$

  entonces, bajo independencia de los errores ε_h,

$$Var\left(\sum_{h=1}^{p} x_h\right) = \left(\sum_{h=1}^{p} \pi_h\right)^2 Var(\xi) + \sum_{h=1}^{p} Var(\varepsilon_h) \tag{2.80}$$

donde se considera un bloque como unidimensional si $\left(\sum_{h=1}^{p} \pi_h\right)^2$ es grande. Entonces el coeficiente ρ de Dillon-Goldstein es:

$$\rho = \frac{\left(\sum_{h=1}^{p} \pi_h\right)^2 Var(\xi)}{\left(\sum_{h=1}^{p} \pi_h\right)^2 Var(\xi) + \sum_{h=1}^{p} Var(\varepsilon_h)} \tag{2.81}$$

Para estimar el coeficiente ρ se usa como estimación de ξ la primera componente principal estandarizada del bloque de variables manifiestas, F_1 o t_1, estimando π_h como $cor(x_h, t_1)$ y $Var(\varepsilon_h)$ como $(1 - cor^2(x_h, t_1))$. Se considera un bloque como unidimensional si la estimación $\hat{\rho} > 0{,}7$. Según Chin (1998), este indicador proporciona una medida más precisa que el coeficiente α de Cronbach de la consistencia interna del bloque de variables manifiestas.

Cualquiera que sea la medida de unidimensionalidad utilizada, si los datos no se ajustan a ella, es necesario prescindir de las variables indicadoras que sea necesario o utilizar una especificación diferente para el modelo de medida.

2. *Formativa* En este caso, las variables latentes se escriben en función de las variables manifiestas así:

$$\boldsymbol{\xi} = \boldsymbol{\Pi_{\xi} x} + \boldsymbol{\delta_{\xi}} \tag{2.82}$$
$$\boldsymbol{\eta} = \boldsymbol{\Pi_{\eta} y} + \boldsymbol{\delta_{\eta}} \tag{2.83}$$

donde los bloques de variables manifiestas que abarcan cada variable latente sí pueden ser multidimensionales. Los signos de los coeficientes que relacionan variables latentes y manifiestas pueden ser distintos, aunque deben de ser razonables como lo son, por ejemplo, en regresión múltiple. Se supone que $E(\xi|x) = \Pi_{\xi} x$ y que $E(\eta|y) = \Pi_{\eta} y$.

3. *MIMIC* o mixta. Es una combinación de las dos anteriores, de forma que algunas variables latentes se forman de manera reflexiva y otras de manera formativa. Por ejemplo, en el caso de las latentes exógenas,

$$\boldsymbol{x_1} = \boldsymbol{\Pi_{x1} \xi_1} + \boldsymbol{\varepsilon_{x1}} \tag{2.84}$$
$$\boldsymbol{\xi_2} = \boldsymbol{\Pi_{\xi 2} x_2} + \boldsymbol{\delta_{\xi 2}} \tag{2.85}$$

de forma que el vector de variables latentes es $\boldsymbol{\xi} = (\boldsymbol{\xi_1 \xi_2})'$, donde $\boldsymbol{\xi_1}$ se forma de manera reflexiva y $\boldsymbol{\xi_2}$ de manera formativa.

Normalización de las variables latentes Las variables latentes se construyen de manera que están normalizadas. Habitualmente se normalizan de manera que están tipificadas, aunque es posible utilizar otra medida de normalización, como en Fornell (1992).

2.9.2. El algoritmo de estimación

El algoritmo de estimación de modelos de mínimos cuadrados parciales ha sido desarrollado principalmente por Wold (1982, 1985*a*) y ampliado por Lohmöller (1989), quien también desarrolló el primer software que existió para estimar este tipo de modelos, de nombre *LVPLS*. Hoy existen otros muchos, como PLSX, PLS-Graph, Smart-PLS y módulos añadidos en Spad o SPSS.

Estimación de las variables latentes

Existen dos maneras de obtener o estimar las variables latentes según que dichas variables se consideren como exógenas o como endógenas (ver ecuación (2.74)). Respectivamente se les denomina estimación externa (para las variables endógenas y exógenas) e interna (para las endógenas).

Estimación interna Una variable latente estandarizada endógena se estima internamente en función de las otras latentes de las que depende así:

$$\hat{\eta}_j = \sum_{j':\eta_{j'} \text{ conectada a } \eta_j} \hat{\beta}_{j'} \hat{\eta}_{j'} \quad + \sum_{j':\xi_{j'} \text{ conectada a } \eta_j} \hat{\gamma}_{j'} \hat{\xi}_{j'} \tag{2.86}$$

Los pesos internos $\hat{\beta}_{j'}$ y $\hat{\gamma}_{j'}$ pueden estimarse de varias maneras:

1. Centroide. Los pesos internos $\hat{\beta}_{j'}$ y $\hat{\gamma}_{j'}$ son iguales a los signos de los coeficientes de correlación entre $\hat{\eta}_j$ y, respectivamente, $\hat{\eta}_{j'}$ y $\hat{\xi}_{j'}$. El aparente problema de la inestabilidad del signo cuando el coeficiente de correlación es cercano a cero no presenta problemas en la práctica (Tenenhaus et al. 2005).

2. Factorial. Los pesos internos son iguales a los coeficientes de correlación simples entre las variables ficticias involucradas.

3. Estructural (o *Path weighting*). Las variables latentes conectadas a una latente determinada ξ_j se clasifican en predecesoras (causantes de aquélla) y sucesoras (causadas por aquélla). Para una variable predecesora, su peso interno es igual al coeficiente de regresión múltiple de la regresión de la variable $\hat{\xi}_j$ sobre todas las estimaciones de las latentes predecesoras,

incluida ella misma. Para una variable sucesora, el peso interno es el coeficiente de correlación simple entre sucesora y sucedida.

Según Tenenhaus et al. (2005), aun respondiendo a diferentes consideraciones teóricas, la elección del método de estimación de los pesos internos no introduce diferencias significativas en los resultados.

Estimación externa Una variable latente estandarizada se estima a partir de las variables manifiestas que la componen mediante

$$\hat{\xi}_j = \sum_h w_{jh}(x_{jh} - \overline{x}_{jh}) \tag{2.87}$$

de forma que la variable ξ_j queda estandarizada con signo arbitrario, que debe de tenerse en cuenta, dado el significado de las variables. Es habitual reescalar $\hat{\xi}_j$ de forma que pertenezca al intervalo $[0, 100]$. Los pesos w_{jh} se denominan pesos externos y existen varias maneras de obtenerlos:

1. *Modo A:* En este modo, el peso es el coeficiente de regresión en la regresión simple de x_{jh} sobre $\hat{\xi}_j$ (estandarizada), es decir,

$$w_{jh} = cov(x_{jh}, \hat{\xi}_j) \tag{2.88}$$

   Este modo es más apropiado para un bloque de variables manifiestas correspondiente a un modelo de medida reflexivo, es decir, un bloque unidimensional. Suele usarse más para variables latentes endógenas.

2. *Modo B*: En este caso, el vector $\boldsymbol{w}$ de pesos w_{jh} se toma igual al vector de coeficientes de regresión múltiple de $\hat{\xi}_j$ sobre las variables manifiestas centradas $x_{jh} - \overline{x}_{jh}$. Este modo es más acorde con un modelo de medida de las variables latentes de tipo formativo. Suele usarse en mayor medida para variables latentes exógenas. Cuando el bloque de variables manifiestas contiene variables con alta colinealidad, puede usarse la regresión PLS en lugar de la regresión de mínimos cuadrados habitual. La regresión PLS consiste en una extensión de la regresión en componentes principales cuando existen varias variables respuesta (Wold 1982).

3. *Modo C*: En este modo, debido a Lohmöller (1989), todos los pesos son iguales en valor absoluto pero con signo igual al de las correlaciones entre las variables manifiestas y la latente correspondiente:

$$\text{sign}(w_{jh}) = \text{sign}(cov(x_{jh}, \hat{\xi}_j)) \tag{2.89}$$

   Los pesos se normalizan de forma que la variable latente queda normalizada. Este es un caso particular del modo B, que resulta muy intuitivo para muchos usuarios del método.

Finalmente, el algoritmo de estimación de modelos PLS consiste en, a partir de unos valores iniciales de las variables latentes, ir alternando entre las ecuaciones correspondientes a la estimación externa (2.87) e interna (2.86) hasta que se llega a la convergencia. La estimación de los pesos internos en (2.86) se realiza generalmente mediante regresión mínimo cuadrática ordinaria o, en caso necesario, más raramente mediante regresión PLS.

Elección de valores iniciales para los pesos externos Lohmöller (1989) propone como pesos iniciales para las variables indicadoras de un mismo bloque al valor uno para todas ellas excepto para la última, con peso -1. Esta opción genera algunos problemas de signo, generando signos negativos en algunos bloques de pocas variables. Tenenhaus et al. (2005) proponen que ésta no sea la elección estándar y usar como pesos iniciales los elementos del primer vector propio de un ACP del bloque con mayoría de signos positivos. Si hay empate de signos, proponen asignar el signo positivo a la variable con mayor correlación en términos absolutos y ajustar el resto.

Valores perdidos

Lohmöller (1989) propuso una combinación de imputación mediante la sustitución de valores ausentes por la media en la estimación externa de variables latentes y de simple eliminación de las observaciones afectadas en la estimación de los pesos externos. Más detalles pueden obtenerse en Tenenhaus et al. (2005), pero esta elección parece dar resultados bastante robustos. En principio, no hay motivo por el que no se puedan utilizar otros métodos de imputación disponibles en la literatura.

2.9.3. Validación del modelo

Un modelo PLS no presenta supuestos distribucionales para la estimación de los parámetros, por lo que las técnicas paramétricas habituales no son de utilidad. Esto afecta fundamentalmente a las medidas de calidad del modelo, como son las medidas de bondad de ajuste a un modelo de regresión, y a la realización de contrastes de hipótesis, principalmente de significación, sobre los parámetros del modelo estructural.

Medidas de calidad

Las medidas de calidad están orientadas a la predicción y son de corte no paramétrico. Un modelo PLS puede ser validado desde varios puntos de vista.

Puede valorarse la calidad del modelo de medida, el modelo estructural y cada ecuación de regresión estructural.

Calidad del modelo de medida Es posible calcular un *índice de comunalidad* para medir la calidad de medida de cada bloque. Se obtiene, para el bloque de variables manifiestas j-ésimo, mediante

$$\text{comunalidad}_j = \frac{1}{p_j}\sum_{h=1}^{p_j}\text{cor}^2(x_{jh}, \hat{\xi}_j) \tag{2.90}$$

donde p_j es el número de variables indicadoras del bloque j. A partir de los J bloques de variables indicadoras existentes, puede calcularse la *comunalidad media* del modelo:

$$\overline{\text{comunalidad}} = \frac{1}{p}\sum_{j=1}^{J} p_j \text{ comunalidad}_j \tag{2.91}$$

donde p es el número total de variables indicadoras de todos los bloques.

Otras medidas de calidad del modelo de medida son la *Fiabilidad compuesta* y la *Varianza media extraída*, ambas válidas para bloques unidimensionales (estimados por el modo A). La primera está relacionada con el α de Cronbach (ecuación (2.78)) y la segunda con la comunalidad media de la ecuación (2.91) anterior.

Calidad de las ecuaciones estructurales Una medida de calidad para una ecuación estructural en la que sólo aparecen variables latentes, una como endógena y el resto como explicativas, es, simplemente el coeficiente de determinación R^2. La interpretación es similar a la de la regresión lineal habitual. Chin (1998) propone el uso de estadísticos F en función de los R^2 de modelo restringido y no restringido para determinar si el efecto de una variable latente particular sobre la latente dependiente es sustancial (evitando el término significativo, asociado a estadísticos con distribución conocida). Se usarían como umbrales los valores 0,02, 0,15 y 0,35 para indicar que el efecto de la variable independiente es pequeño, medio o grande.

Calidad del modelo estructural La calidad del modelo estructural debería de tener en cuenta tanto la parte estructural como el modelo de medida de las variables latentes implicadas. Por un lado, puede calcularse el *índice de redundancia* para un bloque correspondiente a la variable latente endógena j como

$$\text{redundancia}_j = \text{ comunalidad}_j \times R^2(\hat{\xi}_j, \{\hat{\xi}_{j'} | \hat{\xi}_{j'} \text{ explican } \hat{\xi}_j\}) \tag{2.92}$$

De manera similar a como puede hacerse para la comunalidad, puede también calcularse un índice de *redundancia media*.

Finalmente, Amato et al. (2005) proponen como criterio de bondad de ajuste global la media geométrica de la comunalidad media y el R^2 medio:

$$\text{GoF} = \sqrt{\overline{\text{comunalidad}} \times \overline{R^2}} \tag{2.93}$$

Técnicas de validación cruzada para la medición de la calidad Es posible tener una medida de la variabilidad de los índices de comunalidad y de redundancia mediante validación cruzada.

El procedimiento para el cálculo del índice de comunalidad obtenido mediante validación cruzada se puede consultar en Tenenhaus et al. (2005) o Chin (1998). Mediante una técnica de *jacknife*, se omiten sucesivamente un número G de veces[4] varios elementos de todas las variables indicadoras (pueden corresponder a diferentes filas de las distintas variables indicadoras dentro de la misma iteración) que forman una latente, formando G submuestras. En la práctica se suele omitir un elemento cada vez, esto es, se usan submuestras de $N-1$ observaciones para muestras originales de tamaño N. Utilizando los procedimientos establecidos para el tratamiento de valores ausentes, se trata de estimar la variable latente y, con ella, predecir las variables indicadoras que la forman (por ejemplo, dando la vuelta a la ecuación (2.87)). Una vez se tienen las variables indicadoras y sus predicciones, se calculan los errores de predicción y se descompone la varianza de éstos en la forma habitual para cada una de las G submuestras. Se suman todas las G sumas de cuadrados entre sí (totales y explicadas o residuales cada una por su lado) para dar lugar a un pseudo coeficiente de determinación

$$H_j^2 = 1 - \frac{\sum_g^G SRC_g}{\sum_g^G STC_g} \tag{2.94}$$

que es el índice de comunalidad obtenido por validación cruzada (o aproximación de *blindfolding*) para las variables del bloque j.

Para la obtención del índice de redundancia mediante validación cruzada el procedimiento es esencialmente el mismo salvo porque la estimación de la variable latente endógena se produce a partir del modelo interno y no sólo a partir de las indicadoras que la componen. El procedimiento jackknife produce G muestras y G sumas de cuadrados con las que se obtiene un pseudo coeficiente de determinación de construcción idéntica a la ecuación (2.94) que es precisamente el índice que se busca y que también recibe el nombre de Q^2 de

[4] H. Wold propuso usar $G = 7$ submuestras

Stone-Geisser (ver Tenenhaus et al. (2005), Chin (1998)). Este índice es un indicador de cómo los valores observados pueden ser reconstruídos a través del modelo y de sus parámetros estimados, de forma que $Q^2 > 0$ indica que el modelo tiene una buena capacidad predictiva y $Q^2 < 0$ que no la tiene.

Parámetros del modelo

En un modelo PLS existen muchos parámetros para los que, dada la no existencia de supuestos distribucionales, es conveniente disponer de técnicas de validación cruzada para valorar su estabilidad y, si acaso, realizar contrastes de hipótesis. Los parámetros de mayor interés son los coeficientes del modelo interno, aunque también puede ser interesante valorar la variabilidad de pesos, cargas factoriales o las correlaciones entre variables latentes.

En los modelos PLS se han usado principalmente dos técnicas de validación cruzada: *jackknife* y *bootstrap*. La técnica jackknife ha sido explicada brevemente en el apartado anterior para el cómputo de índices de calidad. De manera similar pueden calcularse medias y desviaciones típicas de coeficientes estimados (u otros parámetros) sobre las G submuestras de forma que se obtienen estadísticos t (similares a los t de Student) útiles para obtener intervalos de confianza robustos. Los detalles pueden obtenerse en Chin (1998). Según Tenenhaus et al. (2005), este procedimiento tiende a proporcionar desviaciones estándar pequeñas, lo que conduce sistemáticamente a la consideración de parámetros como significativos.

La alternativa a jackknifing, más reciente y popular, es el bootstrap (Efron & Tibshirani 1993). En bootstrap se obtienen muestras del mismo tamaño que la muestra original mediante muestreo con reemplazamiento de las unidades de la muestra original. El número de muestras bootstrap es una elección del analista, pero suele ser razonablemente alto (en muchos paquetes estadísticos que implementan esta técnica se elige el valor 100 por defecto). Al igual que en el jackknife, con las muestras obtenidas se obtiene un estadístico t a partir de la media y desviación típica muestrales de los parámetros estimados en el número de replicaciones elegido, con distribución asintóticamente normal. De esta manera se obtienen los intervalos de confianza correspondientes.

La elección de un método u otro depende de consideraciones de tiempo de cálculo y eficiencia. Por un lado el jackknife suele ser más rápido debido a que utiliza menos muestras que el bootstrap. Por otro lado, el jackknife se considera menos eficiente que el bootstrap dado que puede considerarse como una aproximación a éste (Efron & Tibshirani 1993).

Consideraciones particulares para los modelos PLS En los modelos PLS las variables latentes quedan completamente definidas excepto por el

signo, dado que tanto $\hat{\xi}_j = \sum_h w_{jh}(x_{jh} - \overline{x}_{jh})$ como su opuesto $-\hat{\xi}_j$ son soluciones igualmente válidas. Wold (1985*b*) propuso elegir la solución con una mayoría de signos positivos entre las variables manifiestas x_{jh} y la latente $\hat{\xi}_j$. El problema de esta elección es que cuando se realiza un remuestreo, en las distintas submuestras (o en las muestras bootstrap) se producen cambios arbitrarios de signos en los pesos externos, lo cual afecta a los pesos $\hat{\pi}$ y a los coeficientes $\hat{\beta}$ y $\hat{\gamma}$ del modelo interno. Cuando esto se produce, los errores estándar de las estimaciones se incrementan, generando estadísticos t anormalmente bajos.

Para corregir el problema de los signos entre distintas muestras, Chin (2001) en el software *PLS-Graph* muestra tres posibilidades de actuación:

- *Estándar*. No se hace nada para corregir el problema de los signos, con el consiguiente problema mencionado en los estadísticos t.

- *Cambios individuales de signos*. Para cada submuestra se iguala el signo de cada peso externo con el de la muestra original. Esto puede generar problemas de coherencia global al nivel de la submuestra considerada, salvo que todos los signos de un bloque sean iguales en la muestra original.

- *Cambios en los constructos (o variables latentes)*. En modo B, usar pesos externos para comparar estimaciones de latentes de la muestra original con las submuestras puede inducir a error si hay multicolinealidad acusada entre las variables indicadoras de la latente.

  Se considera mejor usar las cargas factoriales, y comparar las cargas de la latente en cada submuestra con la de la muestra original. Si O denota la muestra original y S una submuestra, se recomienda invertir el signo de las cargas (y de los pesos) si las cargas estimadas son tales que

$$\left|\sum_h (L_h^O - L_h^S)\right| > \left|\sum_h (L_h^O + L_h^S)\right| \tag{2.95}$$

  donde h es un índice que corre sobre todas las variables manifiestas del bloque correspondiente a la latente considerada.

2.10. Modelos Logit

Los modelos logit forman parte de una clase de modelos de corte inferencial más amplia, como son los modelos de variables dependientes discretas. En realidad son modelos de regresión en los cuales la variable endógena o dependiente es discreta, y en el caso concreto de los de elección binaria, con la

particularidad de que, además, sólo puede tomar dos valores, que normalmente son 0 y 1. A este tipo de modelos se les denomina modelos de elección binaria. Los modelos logit son una subclase dentro de estos modelos de elección binaria, y su estudio también se denomina Análisis de Regresión Logística. Es una metodología largamente desarrollada y establecida, algunas de cuyas referencias principales son Amemiya (1981), McFadden (1984), Maddala (1983) y Dhrymes (1984).

2.10.1. Especificación

Se considera que la variable endógena Y_i es una variable aleatoria que sólo puede tomar dos valores, o presentar dos estados, codificados como $Y_i = 0$ e $Y_i = 1$ (a veces asociados con la idea de *fracaso* y *éxito*, respectivamente). Además, se supone que el resultado de esta variable se obtiene a partir de una distribución de probabilidad binaria que es función de unas variables, explicativas o independientes $\boldsymbol{x_i}$, relación que se establece a través de unos parámetros ($\boldsymbol{\beta}$), de forma que

$$\begin{aligned} Prob(Y_i = 1) &= F(\boldsymbol{x}_i, \boldsymbol{\beta}) \\ Prob(Y_i = 0) &= 1 - F(\boldsymbol{x}_i, \boldsymbol{\beta}) \end{aligned} \tag{2.96}$$

El dilema de los modelos de elección binaria se presenta en la elección de la función F. Generalmente no es adecuada una especificación lineal ($F(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{\beta'x}_i$) por varias razones; la principal es que la función F es una probabilidad, pero no es posible generalmente restringir la combinación lineal $\boldsymbol{\beta'x}_i$ al intervalo $[0, 1]$.

Para resolver el problema del recorrido de la función $F(\boldsymbol{x}_i, \boldsymbol{\beta})$ es necesario que, para un vector de regresores dado $\boldsymbol{x}_i$, se cumpla que

$$\begin{aligned} \lim_{\boldsymbol{\beta'x}_i \to +\infty} Prob(Y_i = 1) = 1 \\ \lim_{\boldsymbol{\beta'x}_i \to -\infty} Prob(Y_i = 1) = 0 \end{aligned} \tag{2.97}$$

Existen varias funciones (Maddala 1983) que cumplen las condiciones (2.97). Las más utilizadas son la función de distribución de la distribución normal tipificada $\Phi(\boldsymbol{\beta'x}_i)$, que da lugar a los modelos *probit* y la distribución logística, en la que

$$\begin{aligned} Prob(Y_i = 1) = F(\boldsymbol{x}_i, \boldsymbol{\beta}) &= \frac{e^{\boldsymbol{\beta'x}_i}}{1 + e^{\boldsymbol{\beta'x}_i}} \\ &= \Lambda(\boldsymbol{\beta'x}_i), \end{aligned} \tag{2.98}$$

que da lugar a los modelos logit. El uso de la función logística en este contexto se debe principalmente al impulso de Berkson (1944). En cualquier caso, la elección de una u otra función es arbitraria; analíticamente, la función logística presenta unas colas más gruesas que la normal, pero en la práctica no suelen proporcionar resultados muy dispares. Ver Amemiya (1981).

Interpretación de los coeficientes $\boldsymbol{\beta}$ en un modelo logit

Un modelo de elección binaria, y el logit en particular, es en realidad un modelo no lineal, por lo que la interpretación de los coeficientes no se corresponde exactamente con la de un modelo de regresión lineal. En particular, dada la distribución de Y_i en las ecuaciones (2.96) se tiene que

$$E(Y_i|\boldsymbol{x}_i) = 0[1 - F(\boldsymbol{\beta'x}_i)] + 1[F(\boldsymbol{\beta'x}_i)] = F(\boldsymbol{\beta'x}_i) \tag{2.99}$$

de forma que los coeficientes $\boldsymbol{\beta}$ no tienen la interpretación habitual de los parámetros de un modelo lineal, ya que

$$\frac{\partial E(Y_i|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} = \left\{ \frac{dF(\boldsymbol{\beta'x}_i)}{d(\boldsymbol{\beta'x}_i)} \right\} \boldsymbol{\beta} = f(\boldsymbol{\beta'x}_i)\boldsymbol{\beta} \tag{2.100}$$

donde $f()$ es la función de densidad (o derivada) de la función $F()$. En el caso de la función logística,

$$\frac{\partial E(Y_i|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} = \frac{d\Lambda(\boldsymbol{\beta'x}_i)}{d(\boldsymbol{\beta'x}_i)}\boldsymbol{\beta} = \Lambda(\boldsymbol{\beta'x}_i)[1 - \Lambda(\boldsymbol{\beta'x}_i)]\boldsymbol{\beta} \tag{2.101}$$

ya que

$$\frac{d\Lambda(\boldsymbol{\beta'x}_i)}{d(\boldsymbol{\beta'x}_i)} = \frac{e^{\boldsymbol{\beta'x}_i}}{(1 + e^{\boldsymbol{\beta'x}_i})^2} \tag{2.102}$$

La ecuación (2.101) muestra que los efectos de los cambios en $\boldsymbol{x}_i$ sobre la $Prob(Y_i = 1)$ y la $Prob(Y_i = 0)$ no son constantes sino que cambian con los valores de $\boldsymbol{x}_i$. En la interpretación de la estimación de un modelo, se acompañan las estimaciones de los coeficientes β de los efectos marginales de (2.101), evaluados en uno o varios conjuntos de valores para $\boldsymbol{x}_i$. Una elección natural es el vector de medias de $\boldsymbol{x}$.

2.10.2. Estimación e inferencia

Los modelos de elección binaria se estiman generalmente por el método de máxima verosimilitud (MV). Pueden existir otras alternativas, como el método de la mínima χ^2, original de Berkson (1944). Este método presenta la desventaja de que sólo es útil si existen muchas observaciones por celda, es decir,

muchas observaciones de Y para idénticos valores en las variables explicativas $\boldsymbol{x}$. Una discusión sobre la comparación entre los dos métodos mencionados puede consultarse en Amemiya (1985) incluyendo varias referencias externas.

La estimación máximo verosímil se obtiene como sigue. Cada observación y_i de la variable dependiente Y_i es considerada como el resultado de una variable aleatoria con distribución binaria o de Bernoulli. Para una muestra de n observaciones, la probabilidad conjunta y, por tanto, su función de verosimilitud, es

$$Prob(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \prod_{y_i=0} [1 - F(\boldsymbol{\beta'} \boldsymbol{x}_i)] \prod_{y_i=1} [F(\boldsymbol{\beta'} \boldsymbol{x}_i)] \quad (2.103)$$

reescribible como la función de verosimilitud

$$L = \prod_{i=1}^{n} [F(\boldsymbol{\beta'} \boldsymbol{x}_i)]^{y_i} [1 - F(\boldsymbol{\beta'} \boldsymbol{x}_i)]^{1-y_i} \quad (2.104)$$

y su logaritmo

$$\ln L = \sum_{i=1}^{n} \{y_i \ln F(\boldsymbol{\beta'} \boldsymbol{x}_i) + (1 - y_i) \ln[1 - F(\boldsymbol{\beta'} \boldsymbol{x}_i)]\} \quad (2.105)$$

Las condiciones de primer orden (CPO) para el máximo de esta función son

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{1 - F_i} \right] \boldsymbol{x}_i = \mathbf{0} \quad (2.106)$$

donde se denota, como simplificación, a $F(\boldsymbol{\beta'} \boldsymbol{x}_i)$ como F_i, siendo f_i su derivada. La expresión (2.106) es válida tanto para un modelo logit como para un probit y es claramente no lineal, por lo que su solución requiere de algún método numérico de estimación.

En el caso del logit, insertando (2.98) y (2.101) en la CPO de (2.106) las CPO pueden escribirse:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \Lambda_i) \boldsymbol{x}_i = \mathbf{0} \quad (2.107)$$

que puede solucionarse por un método de estimación no lineal. Las derivadas segundas (o condiciones de segundo orden) resultan en el hessiano

$$\boldsymbol{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta'}} = -\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) \boldsymbol{x}_i \boldsymbol{x}_i', \quad (2.108)$$

una matriz semidefinida negativa, por lo que cumple la condición de máximo.

Matriz de covarianzas asintótica del estimador

La matriz de covarianzas asintótica se puede estimar como la inversa del hessiano (ecuación 2.108) evaluada en el estimador MV. Otra opción es el estimador BHHH (Berndt et al. 1974), que en el caso del modelo logit queda

$$\boldsymbol{B} = \sum_{i=1}^{n} (y_i - \Lambda_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i' \tag{2.109}$$

Estimaciones de las probabilidades y de los efectos marginales

Las probabilidades estimadas por el modelo logit $\hat{F}_i = F(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)$ y los efectos marginales estimados $\hat{f}_i \hat{\boldsymbol{\beta}} = f(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\hat{\boldsymbol{\beta}}$ son funciones no lineales de las estimaciones obtenidas.

Las desviaciones típicas de ambos pueden aproximarse por el método delta. Así, si denotamos por $AVar$ la varianza o matriz de covarianzas asintótica se tiene, para el caso de las probabilidades $\hat{F}_i$

$$AVar(\hat{F}_i) = \left[\frac{\partial \hat{F}_i}{\partial \hat{\boldsymbol{\beta}}}\right]' \boldsymbol{V} \left[\frac{\partial \hat{F}_i}{\partial \hat{\boldsymbol{\beta}}}\right] \tag{2.110}$$

y

$$\boldsymbol{V} = AVar(\hat{\boldsymbol{\beta}}) \tag{2.111}$$

$$\left[\frac{\partial \hat{F}_i}{\partial \hat{\boldsymbol{\beta}}}\right] = \left[\frac{\partial F(\boldsymbol{x_i}' \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right] = \hat{f}_i \boldsymbol{x_i} \tag{2.112}$$

de forma que

$$AVar(\hat{F}_i) = \hat{f}_i^2 \boldsymbol{x_i}' \boldsymbol{V} \boldsymbol{x_i} \tag{2.113}$$

que depende del vector $\boldsymbol{x_i}$ utilizado.

En el caso de los efectos marginales $\hat{f}_i \hat{\boldsymbol{\beta}}$, denotados por $\hat{\boldsymbol{\gamma}}_i$,

$$AVar(\hat{\boldsymbol{\gamma}}_i) = \left[\frac{\partial \hat{\boldsymbol{\gamma}}_i}{\partial \hat{\boldsymbol{\beta}}}\right]' \boldsymbol{V} \left[\frac{\partial \hat{\boldsymbol{\gamma}}_i}{\partial \hat{\boldsymbol{\beta}}}\right] \tag{2.114}$$

$$= [\Lambda_i(1-\Lambda_i)]^2 [I + (1-2\Lambda_i)\boldsymbol{\beta} \boldsymbol{x_i}'] \boldsymbol{V} [I + (1-2\Lambda_i)\boldsymbol{x_i} \boldsymbol{\beta}'] \tag{2.115}$$

Finalmente, con las matrices de covarianzas expuestas es posible realizar contrastes de hipótesis asintóticamente válidos mediante los estadísticos de Wald, de la razón de las verosimilitudes o de los multiplicadores de Lagrange.

2.10.3. Bondad del ajuste

Los modelos logit (en general, los modelos de elección discreta) no son modelos lineales, por lo que no es posible obtener una descomposición de sumas de cuadrados y un coeficiente de determinación comprendido entre 0 y 1 e interpretable de la misma manera que en un modelo de regresión lineal.

Existen diversas aproximaciones para tener una medida de la bondad del ajuste, varias de ellas basadas en la comparación de la verosimilitud evaluada en el máximo (en el estimador MV) y la evaluada en el modelo que podríamos llamar trivial, es decir, aquél en que todos los coeficientes β_i son cero, salvo la constante.

La primera aproximación es el índice del cociente de verosimilitudes,

$$R^2_{ICV} = 1 - \frac{\ln L_0}{\ln L} \tag{2.116}$$

donde L_0 es la función de verosimilitud de un modelo cuya única variable explicativa es la constante y L la función de verosimilitud evaluada en el estimador MV. El problema es que aunque teóricamente $0 \leq R^2_{ICV} \leq 1$, en la práctica sólo puede suceder que $R^2_{ICV} = 1$ si $\boldsymbol{\beta' x} = \pm\infty$ lo que no es posible ni deseable. Generalmente esta medida infravalora el ajuste.

Otra posibilidad es, por ejemplo, el coeficiente de Cox y Snell

$$R^2_{CS} = 1 - \left(\frac{L_0}{L}\right)^{2/n} \tag{2.117}$$

que tampoco puede alcanzar el valor máximo 1. Una modificación de este coeficiente, que tiene a 1 como valor máximo, es el R^2 de Nagelkerke:

$$R^2_L = \frac{R^2_{CS}}{(R^2_{CS})_{MAX}} = \frac{1 - \left(\frac{L_0}{L}\right)^{2/n}}{1 - (L_0)^{2/n}} \tag{2.118}$$

2.11. Conclusiones

En este capítulo se han presentado brevemente las técnicas que constituyen la metodología base para el análisis multivariante exploratorio de grandes volúmenes de datos, ya sean de naturaleza cuantitativa, cualitativa o, como se verá en el capítulo 5, de naturaleza textual.

El enfoque de estas técnicas está asociado a la estructura de tabla única como activa. Sin embargo, los objetivos del análisis, sobre todo desde el punto de vista de la interpretación de los resultados, se pueden enriquecer permitiendo

la intervención de otras tablas de datos como elementos suplementarios (tanto individuos como variables e incluso de distinta naturaleza a la información analizada).

El principal objetivo ha sido poner de manifiesto la idoneidad de estas técnicas en las diferentes etapas de la Minería de Datos. Dejar constancia de que constituyen herramientas muy útiles en el proceso de extracción del conocimiento implícito que subyace en las bases de datos y su transformación en conocimiento explícito. Los indicadores numéricos y visuales habituales en la metodología exploratoria multivariante proporcionan, por un lado, ayuda para revisar la calidad del conjunto de datos y poder realizar una primera selección de los mismos, en caso de ser necesaria la eliminación de posibles *elementos ruidosos*. Por otro lado, son adecuados para localizar relaciones/asociaciones ocultas y determinar ciertos patrones de comportamiento que permitan guiar la toma de decisiones, objetivo último de muchos estudios. Todo ello, quedará claramente ilustrado y completado en los trabajos empíricos presentados en los siguientes capítulos de esta tesis.

CAPÍTULO 3

Aplicación de las técnicas multivariantes a una encuesta on-line: enfoque desde el Data Mining

3.1. Introducción

El objetivo de este capítulo es el análisis de los datos provenientes de una encuesta mediante técnicas de *data mining* o minería de datos. Los datos obtenidos a partir de la encuesta tratada tienen la particularidad de ser bastante abundantes, aunque no tanto como en algunos de los entornos donde hoy día se usa más frecuentemente el término *data mining*, que están fundamentalmente asociados al ámbito empresarial y al marketing. En ellos, los datos se generan a menudo en grandes cantidades y con una periodicidad casi continua, en tiempo real. Estos datos pueden ser tratados tanto desde un punto de vista puramente descriptivo o exploratorio como inferencial o confirmatorio. En el inferencial se trata de extrapolar los resultados obtenidos de una muestra a la población a la que pertenece, generalmente mediante el establecimiento de algún tipo de modelo. En este capítulo se van a utilizar las técnicas de análisis de uno y otro tipo descritas en el capítulo 2 y se van a aplicar a los datos obtenidos mediante una encuesta on-line.

3.2. Análisis de una encuesta on-line sobre satisfacción y productos corporativos

Los datos objeto de análisis corresponden a una encuesta on-line realizada en el año 2005 a estudiantes, personal docente e investigador (PDI) y de administración y servicios (PAS) de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU).

El objetivo de la encuesta era doble: por un lado existía el interés de conocer el grado de satisfacción percibida sobre la universidad por los miembros integrantes de la misma en ese momento. Por otro, como parte de un proyecto de mejora de imagen de la misma, se trataba de evaluar la aceptación que tendría la apertura de una tienda corporativa donde los artículos a vender en ella son los clásicos objetos de tienda universitaria, con el anagrama de la universidad visible en un lugar preferente. Se disponía al respecto de algunos objetos ya diseñados de esa forma, susceptibles de formar parte del catálogo de la tienda. Como ambos elementos, el grado de satisfacción y la imagen, por un lado, y la aceptación de la tienda corporativa en sí, están claramente relacionados, se decidió implementar una encuesta que proporcionase información sobre ambos. El informe final con resultados principales puede consultarse en Fernández et al. (2005) y una reproducción de la encuesta se encuentra en el Apéndice A de esta tesis.

3.3. Descripción de las variables y de los datos obtenidos

En esta encuesta se realizó un muestreo estratificado con afijación proporcional por campus, sexo y edad, tanto para el personal de administración y servicios (PAS) como para el profesorado y personal investigador (PDI), y por campus, género y ciclo para los alumnos. El universo estaba compuesto por 48.995 alumnos, 1.128 PAS y 3.982 PDI.

En concreto, se enviaron 2.289 invitaciones a realizar la encuesta a los alumnos, 768 al personal de administración y servicios y 1.499 a los profesores. Las respuestas obtenidas fueron 547 para los alumnos (23.9 % de tasa de respuesta), 444 para el PAS (57.81 %) y 754 para el PDI (50.30 %), tasas ciertamente elevadas para una encuesta de esta naturaleza.

Los datos fueron obtenidos mediante una encuesta on-line que estuvo disponible para su realización durante un mes, entre febrero y marzo de 2005. Los detalles del muestreo estratificado pueden consultarse en la Tabla 3.1.

	Estudiantes	PAS	PDI
Población	48995	1128	3982
Tamaño muestral	2289	768	1499
Tasa de respuesta (%)	547 (23,9)	444 (57,81)	754 (50,30)
Error muestral	0,042	0,036	0,032
Nivel de confianza	0,95	0,95	0,95

Tabla 3.1: Características del muestreo realizado en la encuesta on-line.

Los datos disponibles corresponden, en primer lugar, a las respuestas asociadas a un cierto número de preguntas cerradas sobre la imagen de la universidad y sobre una valoración de los productos a vender en la tienda universitaria. Además, se tienen datos de tipo textual sobre dos preguntas abiertas y, finalmente, datos categóricos sobre algunas variables de caracterización de los encuestados.

La encuesta tiene la particularidad de que se planteó en las dos lenguas oficiales (Euskera y Castellano) de la Comunidad Autónoma Vasca (CAV), entorno de referencia para la Universidad del País Vasco/Euskal Herriko Unibertsitatea. La elección del idioma era libre para el encuestado, resultando 304 respuestas en Euskera y 1243 en Castellano.

Para este trabajo, seleccionamos las respuestas correspondientes a las siguientes preguntas:

Preguntas cerradas Las respuestas a estas preguntas dan lugar a variables categóricas. En primer lugar, consideramos las siguientes preguntas genéricas:

1. *En general, estoy satisfecho de pertenecer a la UPV/EHU.* Esta pregunta podía responderse en una clásica escala de 1 a 5. Etiqueta: `Satis`.

2. *¿Estarías interesado en comprar un producto con el logotipo de la UPV/EHU (tal como pañuelos, vacíabolsillos, camisetas, relojes, tazas, ...) para uso personal o para regalo?* Esta pregunta era de respuesta binaria, las respuestas posibles se reducen a dos: Sí o No. Etiqueta: `BuyLogo`.

Además de las preguntas anteriores, seleccionamos las respuestas a la siguiente pregunta:

> *A continuación, mostramos una serie de artículos con el logotipo de la UPV/EHU. Para cada uno de ellos, le pedimos que valore como muy poco probable, poco probable, probable y muy probable la probabilidad de que adquiera alguno de estos productos, ya sea para uso personal o para regalar.*

En este punto se mostraba, para cada uno de un total de 26 productos disponibles, una foto a color y una alternativa de elección múltiple en escala 1 a 4 como se indica en la pregunta.

Finalmente, escogemos variables resultantes de puntuaciones otorgadas por los encuestados a ciertas características generales que serían deseables para los productos susceptibles de ser vendidos en la tienda. Dichas puntuaciones se obtienen mediante una escala ordinal del 1 (poco deseable) al 7 (muy deseable) y las características son las siguientes:

- *Original*
- *Audaz*
- *Práctico*
- *Moderno*
- *Tradicional*
- *Artístico*
- *Elegante*
- *Serio.*

Estas características se han elegido por ser relevantes para los artículos en cuestión y son términos de inspiración semiométrica[1](Lebart et al. 2003).

Preguntas cerradas de caracterización Finalmente se preguntaba sobre aspectos personales de caracterización.

1. *Género.* Masculino o Femenino.
2. *Edad.* Categorizada en base a los intervalos siguientes:
   a) 18-22
   b) 23-29
   c) 30-44

[1]La semiometría es una rama de la investigación de encuestas que trata de descubrir lo que los encuestados realmente desean o piensan que puede diferir de lo que declaran desear u opinar. Para ello usan un conjunto restringido de términos (unos 200) que se interpretan como la ruta a un mapa del subconsciente, organizado en base a unos ejes que representan al Individuo, la Comunidad, el Deber y el Placer, ver Lebart et al. (2003). Este mapa es estable en el tiempo, por países, por género y edad.

d) 45 o más.

Además, se disponía de la siguiente información sobre los individuos encuestados, dado que habían sido preseleccionados con anterioridad:

3. *Vinculación* con la universidad. Estudiantes, PAS o PDI.

4. *Campus* de estudio o de destino laboral. La UPV/EHU tiene centros en campus de los tres territorios históricos de la CAV y se dispone de información del territorio donde está el campus lugar de estudio o trabajo:

   - Araba/Álava
   - Gipuzkoa
   - Bizkaia

Las variables seleccionadas en este capítulo provienen de preguntas únicamente cerradas. Las variables así obtenidas pueden ser consideradas como continuas (al menos, las que tienen un carácter ordinal) o como categóricas. Según cómo se traten este tipo de variables, y acorde con las clasificaciones de las técnicas de data mining realizadas en la sección 2.1.1, vamos a tratar de extraer la información contenida en los datos mediante técnicas exploratorias multivariantes y mediante técnicas confirmatorias o predictivas.

3.4. Técnicas exploratorias multivariantes

Las técnicas exploratorias multivariantes que aplicamos en esta sección tienen en común que tratan a todas las variables del conjunto de datos por igual, desde el punto de vista tanto analítico como interpretativo. En este sentido, no se considera a ninguna variable dependiente de ninguna otra.

Vamos a seguir un proceso en dos etapas, como el sugerido en Lebart (1994) o en Lebart et al. (1998), combinando métodos de ejes principales y de clasificación (ver, por ejemplo, Greenacre (1987)). Los métodos de ejes principales se usan como un paso previo que reduce la dimensionalidad de la tabla de datos original mientras que los de clasificación utiliza un número reducido de ejes principales, que son por construcción variables cuantitativas continuas, reteniendo un porcentaje significativo de la variabilidad original. La clasificación se lleva a cabo mediante un algoritmo apropiado (Nakache & Confais 2005) y las clases de individuos son descritas a partir de las características particularmente diferentes (por exceso o por defecto) de los individuos de cada clase con respecto al colectivo general.

3.4.1. ACP de variables cuantitativas y clasificación

Inicialmente escogemos las variables correspondientes a las puntuaciones otorgadas por los encuestados a las características deseables para los productos susceptibles de ser vendidos en la tienda, con arreglo al vocabulario semiométrico. Dichos términos están en la página 60.

Las variables de corte sociodemográfico como género o edad son consideradas como ilustrativas, es decir, no influyen en el cómputo de los ejes principales, al igual que las variables relacionadas con la satisfacción hacia la institución o el interés de compra.

La tabla de datos que se utiliza para el análisis es una tabla de 1742 individuos $\times$ 8 variables activas. El ACP da lugar a los valores propios y tasas de inercia de la tabla 3.2. Esta tabla sugiere que el número de ejes a retener puede ser 2 o 3, sólo dos en el caso de aplicar estrictamente la convención de mantener ejes con valor propio mayor que 1 en el análisis normado. Es probablemente conveniente incluir también el tercer eje, con un valor propio cercano a ese valor.

	valor propio	% inercia	% inercia acumulada
1	2,4983	31,23	31,23
2	1,8593	23,24	54,47
3	0,9083	11,35	65,82
4	0,7013	8,77	74,59
5	0,6563	8,20	82,79
6	0,5307	6,63	89,43
7	0,4581	5,73	95,15
8	0,3879	4,85	100,00

Tabla 3.2: Valores propios del ACP normado sobre 8 características.

Como ayuda a la interpretación de los ejes puede usarse la tabla 3.3 en la que se muestran las correlaciones entre las variables y los ejes. Apoyando la consideración del eje 3 está la alta correlación de la variable Práctico con el mismo, mayor que la que existe con cualquier otro eje.

La tabla 3.3, junto con las figuras 3.1 y 3.3, permiten la interpretación del ACP. El signo positivo de todas las correlaciones entre las variables activas y el primer eje es indicio de que un gran número de encuestados tiende a otorgar puntuaciones de similar magnitud (sean altas o bajas) sobre todas las variables, no diferenciando entre ellas, en lo que se suele denominar en ACP como efecto talla. El segundo eje contrapone *Tradicional, Elegante* y *Serio* a *Audaz, Original* y, en menor medida, *Moderno*. Por lo tanto, distingue entre individuos que valoran productos que podríamos denominar clásicos de otros

Variable	eje 1	eje 2	eje 3	eje 4	eje 5
Original	0,73	-0,35	0,03	-0,15	0,11
Audaz	0,63	-0,42	-0,13	-0,44	0,13
Práctico	0,48	-0,04	0,81	0,22	-0,15
Tradicional	0,11	0,71	0,26	-0,57	-0,14
Artístico	0,67	0,15	-0,34	0,08	-0,58
Elegante	0,54	0,61	-0,20	0,26	0,03
Serio	0,25	0,78	-0,04	0,06	0,41
Moderno	0,72	-0,23	-0,01	0,19	0,28

Tabla 3.3: Correlaciones entre factores y variables activas.

que valoran productos de corte que podría catalogarse como de innovador, o a la moda. Finalmente, el tercer eje contrapone individuos que valoran mucho el término *Práctico* frente al resto, sobre todo, *Artístico* y *Elegante*, poniendo de relieve la distancia entre practicidad y diseño.

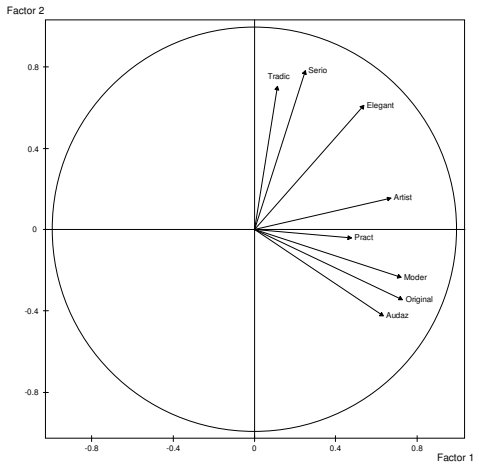


Figura 3.1: Plano principal (1,2) del ACP de la encuesta. Variables activas: valoraciones características deseables.

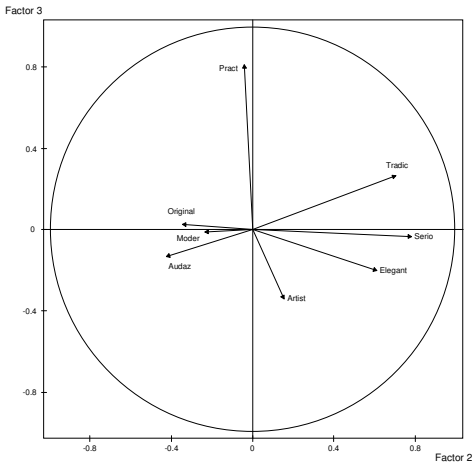


Figura 3.2: Plano (2,3) del ACP de la encuesta. Variables activas: valoraciones características deseables.

En la tabla 3.4 se recogen los valores-test correspondientes a las categorías de las variables suplementarias utilizadas. Los estadísticos a partir de los cuales se obtienen dichos valores test convergen a una distribución normal reducida y se utilizan para contrastar si una categoría se proyecta o no sobre el origen de coordenadas del eje y, por tanto, no se aleja del comportamiento medio,

ver Lebart et al. (2006). Esto es, para un nivel de significación del 5 %, estos valores deberían de exceder en valor absoluto de 1,96 para que las categorías correspondientes fuesen consideradas como características de ese eje.

Variable	Categoría	Efectivo	Eje 1	Eje 2	Eje 3
Estamento					
	Estam=ALUMNOS	547	-2,47	-7,81	4,43
	Estam=PAS	443	2,57	0,45	-1,71
	Estam=PROFESORES	752	0,06	6,93	-2,65
Sexo					
	Sexo=HOMBRE	803	-1,84	3,30	-0,20
	Sexo=MUJER	939	1,84	-3,30	0,20
Satisfecho					
	Satis=1	20	-2,82	-1,58	0,59
	Satis=2	72	-1,66	-2,36	-1,37
	Satis=3	336	-4,82	-3,18	-0,38
	Satis=4	743	0,10	0,17	1,22
	Satis=5	556	5,46	4,18	-0,79
	missing category	15	-0,63	-1,49	1,36
Interés Compra					
	Inter=1	1076	7,10	5,07	-0,30
	Inter=2	636	-7,04	-5,32	0,48
	missing category	30	-0,46	0,74	-0,63
Edad					
	Edad=1	307	-0,77	-6,21	4,23
	Edad=2	228	-1,47	-4,35	3,53
	Edad=3	649	-2,32	1,60	-0,45
	Edad=4	527	4,29	6,19	-5,12
	missing category	31	-0,43	1,64	-1,76

Tabla 3.4: Valores test de las categorías suplementarias del ACP.

La interpretación de los ejes en base a las variables ilustrativas puede hacerse mediante los valores test antes mencionados junto con la visualización de las proyecciones de las categorías de los planos correspondientes a los ejes retenidos, que están en las figuras 3.3 y 3.4.

El primer eje (Figura 3.3) proyecta en su lado derecho categorías elegidas por individuos que manifiestan un interés inicial de compra, antes de visualizar los productos (Inter=1) y de mayor edad (Edad=4, 45 años o más), además de ser los más satisfechos con la institución (Satis=5). En el lado izquierdo

aparecen categorías que representan una falta de interés inicial de compra (Inter=2), una satisfacción media o muy baja (Satis=3,Satis=1 aunque con escaso efectivo) y, en menor medida, de edad intermedia (Edad=3, 35-44 años), y alumnos. Todas estás categorías tienen valores test claramente significativos al 5 % (ver Tabla 3.4) e indican que este eje liga fuertemente la propensión de compra con la satisfacción con la institución y con la edad. No obstante, las categorías de edad y estamento (alumnos) parecen estar mejor representados en un segundo eje, dado que proporcionan mayores valores test en los ejes siguientes.

En cuanto al eje 2 (Figuras 3.3 y 3.4), se observa cómo en la parte positiva aparecen personas de mayor edad (Edad=4), interés previo de compra (Inter=1), satisfechos (Satis=5), hombres y profesores. En cambio, en el lado opuesto, se encuentran alumnos, sin interés previo de compra, de menor edad (Edad=1, Edad=2, entre 18 y 29 años) y en menor medida, de satisfacción media o baja y mujeres. Dados los valores test de la tabla 3.4, este sugiere la relación fuerte existente entre estamento e interés previo de compra, que es elevado entre los profesores y bajo en los alumnos, situándose el personal de administración y servicios (PAS) en una posición intermedia. En ambas figuras se han dibujado las trayectorias de las variables ordinales Edad e Interés para ayudar en su interpretación.

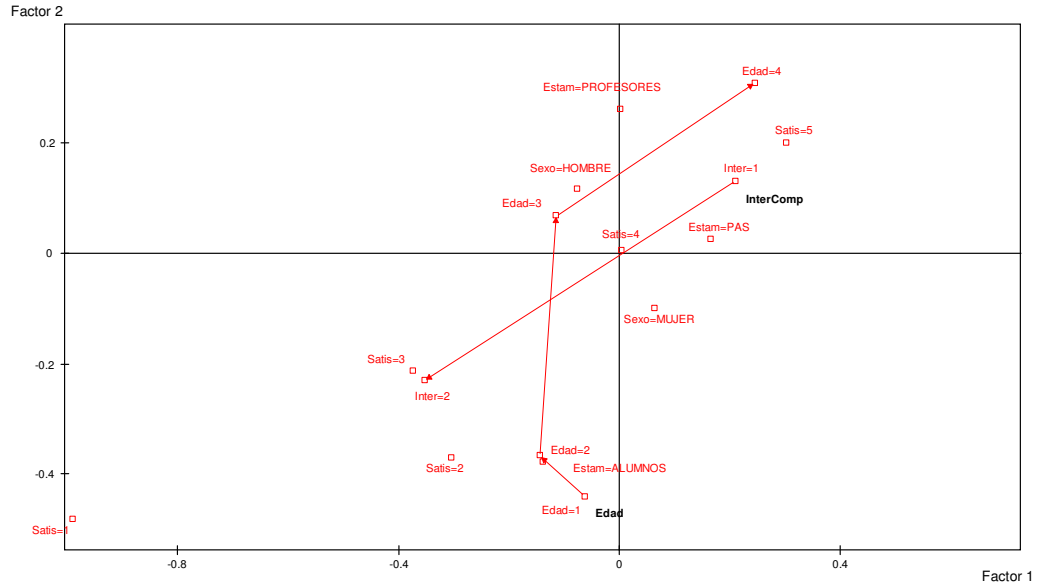


Figura 3.3: Plano principal (1,2) del ACP de la encuesta. Variables ilustrativas.

El eje 3 (Figura 3.4) incide principalmente en la relación entre Estamento

y Edad (mayores valores test en categorías procedentes de estas variables) que por su obviedad es de escasa utilidad práctica.

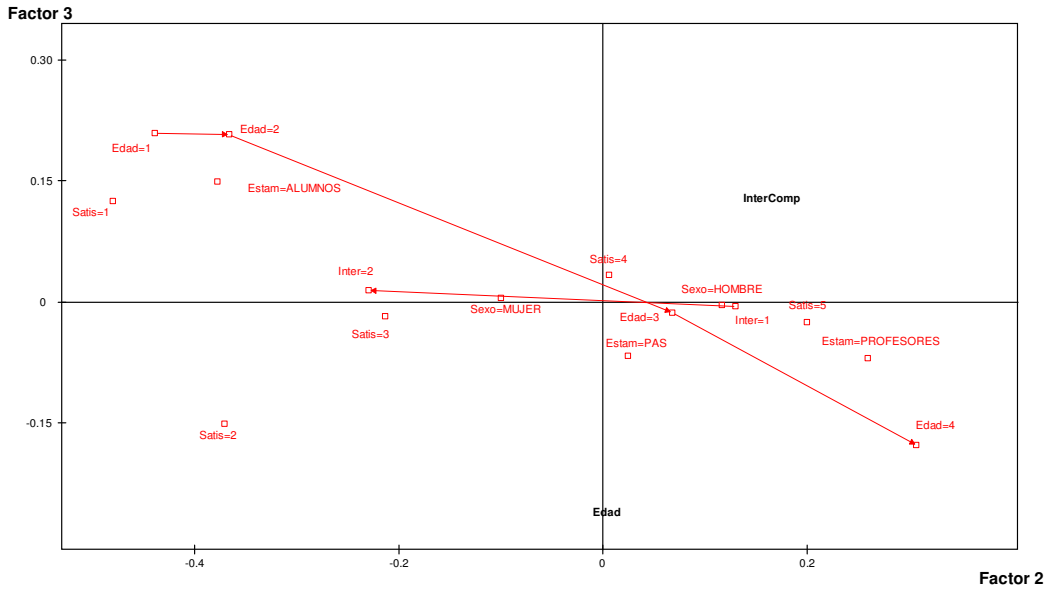


Figura 3.4: Plano (2,3) del ACP de la encuesta. Variables ilustrativas.

En resumen, los planos factoriales obtenidos ponen de manifiesto que la mayor variabilidad presente en los datos de la encuesta (eje 1) corresponde a la asociación entre individuos que valoran positivamente la elegancia y, en menor medida, la seriedad y lo *Artístico* de los productos con los profesores de mayor edad y que son los más satisfechos con la institución y más proclives a comprar productos corporativos. Esta asociación está opuesta a individuos que valoran más pobremente dichas características asociadas con la elegancia y más caracterizados por ser jóvenes, de satisfacción media o baja y con escaso interés previo en la compra de productos.

Existe, sin embargo, otra dimensión de importancia, correspondiente al segundo eje, que asocia elevadas valoraciones de los términos de tradición, seriedad y elegancia a profesores, de elevado interés de compra y mayor edad y con una mayor presencia de hombres. También, en contraposición al grupo descrito anteriormente, se tendrían bajas valoraciones en las características descritas anteriormente, mientras que relativamente elevadas en términos como *Original* o *Audaz* y claramente asociadas con alumnos, de menor edad lógicamente que el grupo anterior, de un escaso interés inicial y satisfacción con la institución, y con una mayor presencia de mujeres.

Una tercera dimensión reflejaría elevadas valoraciones del término *Prácti-*

co que estarían asociadas casi en exclusiva a alumnos e individuos de menor edad que, a nivel interpretativo, podría estar ya recogida en las dimensiones anteriores.

Con respecto a la clasificación, ésta se ha realizado sobre los cinco primeros ejes del ACP, que recogen un 82,79 % de la inercia original, y sobre ellos se ha realizado una clasificación mixta que resulta en una partición en 3 clases. La localización de los centros de las clases en el plano principal conformado por los dos primeros ejes aparece en la Figura 3.5, estando los centros representados por círculos de diámetro proporcional a su tamaño. La descripción de estas clases relativa a las variables activas continuas del ACP se encuentra en la Tabla 3.5 y relativa a las modalidades de las variables suplementarias en la Tabla 3.6.

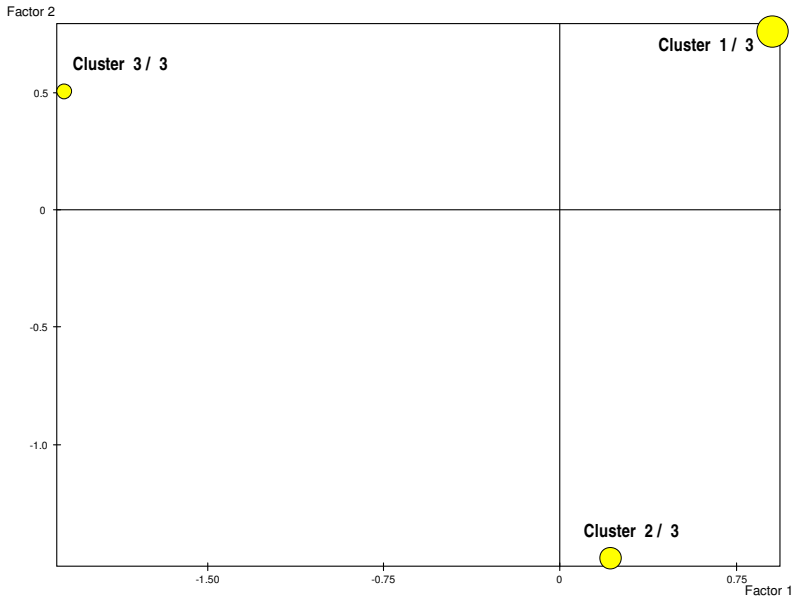


Figura 3.5: Plano principal (1,2) del ACP de la encuesta. Posición de los centros de los clusters y tamaño relativo para una partición en 3 clases.

Parece claro que el primer cluster se sitúa del lado de los individuos que valoran más las características de *Elegante, Serio, Tradicional* y *Artístico*, que se corresponden con la interpretación del lado positivo del primer y del segundo eje, y que son las características más relevantes de dicho cluster en la descripción de la Tabla 3.5. Es el cluster de mayor tamaño, un 46,04 % y se caracteriza por las modalidades ilustrativas de individuos de mayor edad (Edad=4) e interés de compra, muy satisfechos, profesores y miembros del PAS.

El centro del segundo cluster queda colocado en el cuarto cuadrante del plano principal (Figura 3.5) fuertemente asociado a la zona negativa del segundo eje. Contiene al 31,11 % de la muestra y se caracteriza por puntuaciones altas en los términos *Original, Audaz* y *Moderno*, y por encuestados que son elegidos principalmente por alumnos y de edades menores a 29, que manifiestan un interés de compra tendente al no y a satisfacción media.

Finalmente, el tercer cluster sitúa su centro en el segundo cuadrante, más vinculado a la parte negativa del primer eje. Se caracteriza por puntuaciones bajas o muy bajas en los términos *Original, Moderno, Audaz, Artístico, Práctico* y *Elegante* y por escaso interés previo de compra y satisfacción media. La caracterización es mucho más débil que en el caso de los otros dos clusters. Es el más pequeño, con un 22,85 % del total.

En resumen, el ACP sobre las valoraciones de los términos semiométricos referidos a los artículos de la tienda corporativa, complementado con una clasificación sobre los factores, nos ha proporcionado una tipificación de los individuos encuestados basada en una reducción de dimensionalidad sobre la distancia euclidea de unas variables de puntuaciones tomadas como continuas. Asimismo, se obtienen clusters relacionados con las categorías obtenidas de las respuestas a otras preguntas de la encuesta o de características personales de los individuos. La tipificación pone de relieve que los individuos más interesados son de mayor edad y están más vinculados a la universidad, al formar parte del sector docente y al mismo tiempo prefieren artículos de índole más serio y elegante, a diferencia de los estudiantes jóvenes que se diferencian claramente del grupo anterior. El hecho de que sean un grupo inferior en número puede ser debido exclusivamente a que representen un número menor en la encuesta, no a que exista un menor número de estudiantes en la universidad, lo que evidentemente no es cierto.

Variables características	Media Cluster	Media Total	Desviación Cluster	Desviación Total	Valor test	Prob.
Cluster 1 / 3 (Efectivo = 802.00 - Porcentaje = 46.04)						
Elegant	5,946	4,850	0,930	1,625	25,55	0,000
Serio	4,720	3,738	1,313	1,644	22,65	0,000
Tradic	4,227	3,488	1,417	1,593	17,53	0,000
Artist	5,545	4,823	1,222	1,631	16,78	0,000
Moder	5,696	5,299	1,098	1,441	10,41	0,000
Original	5,902	5,530	1,103	1,521	9,29	0,000
Pract	6,319	6,002	0,961	1,326	9,08	0,000
Audaz	4,619	4,372	1,427	1,641	5,71	0,000
Cluster 2 / 3 (Efectivo = 542.00 - Porcentaje = 31.11)						
Original	6,303	5,530	0,903	1,521	14,26	0,000
Audaz	5,165	4,372	1,377	1,641	13,56	0,000
Moder	5,839	5,299	1,128	1,441	10,51	0,000
Pract	6,228	6,002	1,154	1,326	4,76	0,000
Artist	4,670	4,823	1,649	1,631	-2,64	0,004
Elegant	3,841	4,850	1,478	1,625	-17,43	0,000
Tradic	2,364	3,488	1,175	1,593	-19,82	0,000
Serio	2,359	3,738	1,057	1,644	-23,56	0,000
Cluster 3 / 3 (Efectivo = 398.00 - Porcentaje =22.85)						
Elegant	4,065	4,850	1,547	1,625	-10,98	0,000
Pract	5,068	6,002	1,691	1,326	-15,99	0,000
Artist	3,611	4,823	1,529	1,631	-16,91	0,000
Audaz	2,809	4,372	1,281	1,641	-21,67	0,000
Moder	3,774	5,299	1,368	1,441	-23,93	0,000
Original	3,744	5,530	1,499	1,521	-26,70	0,000

Tabla 3.5: Caracterización de la clasificación sobre los 3 primeros factores del ACP de las preguntas semiométricas.

Variable	Categoría	% categ. en grupo	% categ. en muestra	% grupo en categ.	Valor Test	Prob.	Peso
Cluster 1 / 3 (Efectivo = 802.00 - Porcentaje = 46.04)							
Edad	Edad=4	38,15	30,25	58,06	6,58	0,000	527
InterComp	Inter=1	69,08	61,77	51,49	5,77	0,000	1076
Satisf	Satis=5	38,90	31,92	56,12	5,72	0,000	556
Estam	PROFESORES	48,01	43,17	51,20	3,72	0,000	752
Estam	PAS	29,18	25,43	52,82	3,26	0,001	443
Estudios	Estud=4	68,70	65,10	48,59	2,87	0,002	1134
Edad	Edad=2	10,60	13,09	37,28	-2,79	0,003	228
Estudios	Estud=3	29,18	32,95	40,77	-3,05	0,001	574
Edad	Edad=1	12,59	17,62	32,90	-5,08	0,000	307
Satisf	Satis=3	13,34	19,29	31,85	-5,81	0,000	336
InterComp	Inter=2	28,80	36,51	36,32	-6,15	0,000	636
Estam	ALUMNOS	22,82	31,40	33,46	-7,13	0,000	547
Cluster 2 / 3 (Efectivo = 542.00 - Porcentaje = 31.11)							
Estam	ALUMNOS	42,62	31,40	42,23	6,64	0,000	547
Edad	Edad=1	25,46	17,62	44,95	5,57	0,000	307
Estudios	Estud=3	40,41	32,95	38,15	4,36	0,000	574
Edad	Edad=2	17,34	13,09	41,23	3,40	0,000	228
Satisf	Satis=3	23,62	19,29	38,10	2,98	0,001	336
InterComp	Inter=2	41,70	36,51	35,53	2,96	0,002	636
InterComp	Inter=1	57,38	61,77	28,90	-2,47	0,007	1076
Satisf	Satis=5	27,12	31,92	26,44	-2,85	0,002	556
Estudios	Estud=4	57,93	65,10	27,69	-4,14	0,000	1134
Estam	PROFESORES	33,76	43,17	24,34	-5,31	0,000	752
Edad	Edad=4	21,22	30,25	21,82	-5,57	0,000	527
Cluster 3 / 3 (Efectivo = 398.00 - Porcentaje =22.85)							
InterComp	Inter=2	44,97	36,51	28,14	3,90	0,000	636
Satisf	Satis=3	25,38	19,29	30,06	3,36	0,000	336
Estam	PAS	20,35	25,43	18,28	-2,62	0,004	443
Satisf	Satis=5	24,37	31,92	17,45	-3,67	0,000	556
InterComp	Inter=1	53,02	61,77	19,61	-4,00	0,000	1076

Tabla 3.6: Caracterización de la clasificación sobre los 3 primeros factores del ACP mediante las variables ilustrativas.

3.4.2. ACM sobre las valoraciones de los productos y clasificación

En este apartado la tabla objeto de análisis es la que cruza los 1742 individuos con las categorías correspondientes a las valoraciones de los 26 artículos que se muestran en la encuesta on-line como susceptibles de compra. Las valoraciones están medidas en una escala 1-4, que son 1 = muy poco probable, 2 = poco probable, 3 = probable y 4 = muy probable y se refieren a la probabilidad de compra del artículo visualizado.

Las valoraciones se consideran como categorías mutuamente excluyentes de variables cualitativas, por lo que se considera apropiado el uso del Análisis de Correspondencias Múltiples (ACM) para extraer los principales factores existentes detrás de dichas valoraciones, así como una descripción de los individuos encuestados. En ACM la posición final de una categoría sobre un eje es un múltiplo del centro de gravedad corregido de los individuos que escogen dicha categoría. Los ejes se interpretan en base a las coordenadas de las categorías que más contribuyen a la formación de los mismos y que mejor están representadas en ellos, mediante los cosenos cuadrado.

Los factores principales se extraen a partir del examen de los valores propios obtenidos a partir del proceso de diagonalización de la matriz de inercia objeto del ACM. La Tabla 3.7 contiene los mayores valores propios y sus inercias proyectadas, incluyendo las inercias corregidas según el método de Benzécri.

	Valor propio	% Inercia	% Inercia acumulada	% Inercia acum. (Benzécri)
1	0,4347	14,13	14,13	80,18
2	0,2053	6,67	20,80	94,41
3	0,1045	3,40	24,20	96,63
4	0,0781	2,54	26,74	97,44
5	0,0719	2,34	29,07	98,01
6	0,0687	2,23	31,31	98,48
7	0,0635	2,07	33,37	98,80
8	0,0615	2,00	35,37	99,07

Traza de la matriz: 3.07692

Tabla 3.7: Tabla de los 8 primeros valores propios ACM.

Aún cuando la corrección de Benzécri puede tender a proporcionar estimaciones demasiado optimistas de la inercia proyectada, parece claro que existe un salto evidente entre el segundo y el tercer eje, por lo que puede ser suficiente considerar el plano principal conformado por los ejes 1 y 2 como un buen resumen de esta tabla. Una inercia sin corregir del 20 % cuando se tienen 26

variables × 4 categorias = 104 columnas en la matriz de variables indicadoras parece de igual forma un resultado aceptable.

La Figura 3.6 contiene el plano principal del ACM. Se observa una alineación curvilínea de las modalidades activas del ACM, donde las valoraciones de los productos están ordenadas de menor (1) a mayor (4) siguiendo el gráfico de izquierda a derecha. La nube de modalidades proyectada muestra el común efecto Guttman, de forma que el segundo factor opone categorías extremas (1 y 4) a las intermedias (2 y 3).

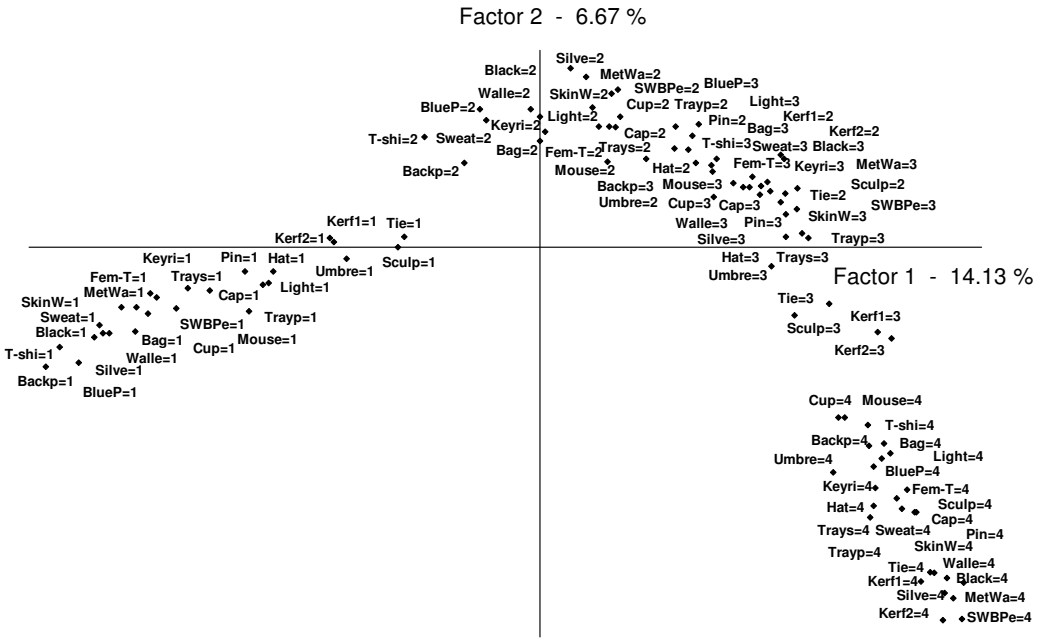


Figura 3.6: ACM: Categorías activas (valoraciones de productos) sobre el plano (1,2).

La Figura 3.6 muestra que, dado que las modalidades están en el centro de gravedad corregido de los individuos que las han elegido, los individuos que estén en el cuarto cuadrante serán los más interesados en comprar los productos, por lo que es interesante caracterizarlos de la forma más precisa posible. De la misma manera, los individuos del tercer cuadrante son los menos interesados y su identificación es igualmente interesante. Los que estén en los cuadrantes 1 y 2 son los individuos de interés medio. La caracterización de todos estos individuos se realiza mediante la proyección en suplementario de variables de caracterización, que son las mismas del apartado precedente en la que se realizaba el ACP de los términos semiométricos. Los valores test correspondientes a las proyecciones de las categorías suplementarias están en

la Tabla 3.8. Aquellas categorías con valor test en valor absoluto superior al cuantil 0,025 de la distribución normal (1,96) son las que se entiende como significativamente caracterizadoras para cada eje al nivel de significación del 5 %. Resulta significativo que las modalidades de las variables *interés previo de compra* (BuyLogo) y *Satisfacción con la institución* (Satis) son todas significativas, fuertemente en muchos casos. Además de éstas, son significativas las modalidades de alumnos, hombre, edad entre 23 y 29 años (edad = 2) en el lado negativo del primer eje y las de profesores, mujer y mayores de 45 en el lado positivo de ese eje.

La visualización de algunas de las proyecciones de las modalidades suplementarias sobre el plano principal pueden verse en la Figura 3.7. Se muestran sólo las de las modalidades de satisfacción y propensión previa de compra, uniendo las de satisfacción mediante segmentos. Así, queda reflejada una trayectoria evidente que sigue la línea marcada por la forma de la nube de modalidades proyectada sobre el plano principal.

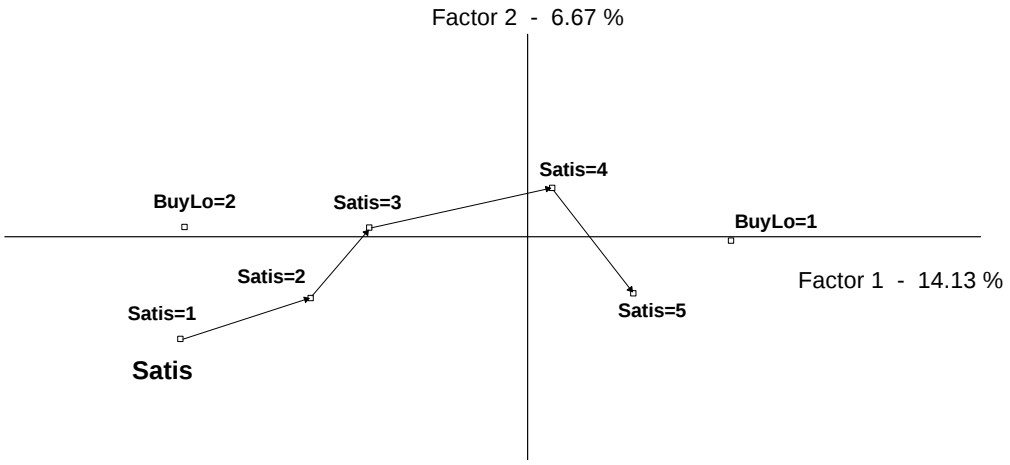


Figura 3.7: ACM: Categorías suplementarias sobre el plano (1,2).

Vistas las proyecciones de las modalidades más características sobre el plano principal, parece que las posibilidades de compra de los artículos mostrados están directamente relacionadas con la satisfacción de pertenencia a la institución, así como con la buena disposición previa a comprar este tipo de productos. Asimismo, hay una mayor disposición de compra entre profesores, mujeres y personas de edad más avanzada. La disposición es escasa entre alumnos, hombres e individuos de entre 23 y 29 años.

Variable o categoría	Efec.	Dist. al orig.	Eje 1	2	3	4	5
Link							
ALUMNOS	547,00	2,18	-3,39	2,57	2,39	7,92	3,52
PAS	443	2,93	1,05	-1,06	-1,85	-3,09	-2,60
PROFESORES	752,00	1,31	2,25	-1,48	-0,61	-4,70	-1,01
Gender							
HOMBRE	803	1,16	-4,26	2,50	-0,11	-1,54	0,58
MUJER	939	0,85	4,26	-2,50	0,11	1,54	-0,58
Campus							
ARABA	254	5,85	-0,42	0,09	-1,03	1,49	1,15
BIZKAIA	1046	0,66	2,93	-0,71	0,41	-0,80	-2,52
GIPUZKOA	442	2,94	-2,96	0,73	0,37	-0,31	1,90
Satis							
1	20	86,10	-3,83	-1,78	-0,19	1,75	-0,31
2	72	23,19	-4,62	-2,06	0,11	0,81	-1,24
3	336	4,18	-7,96	0,68	0,38	1,30	0,89
4	743	1,34	2,14	6,66	0,20	0,66	1,02
5	556	2,13	7,43	-6,41	-0,58	-2,63	-1,27
N.R.	15	115,13	-0,57	0,26	0,26	0,41	0,16
BuyLogo							
1 (si)	1076	0,61	26,54	-1,08	-2,99	1,65	-4,13
2 (no)	636	1,73	-26,64	1,10	2,71	-1,27	3,71
N.R.	30	57,06	-0,53	-0,04	1,11	-1,47	1,68
Age							
1	307	4,67	-1,18	2,05	2,67	6,80	2,09
2	228	6,64	-3,42	1,54	-0,32	3,86	1,63
3	649	1,68	-1,31	-1,14	-1,16	0,24	-0,54
4	527	2,30	5,11	-1,51	-1,00	-8,25	-3,00
N.R.	31	55,19	-0,81	-0,44	0,84	-1,66	2,25
Education							
1	0	0,00	0,00	0,00	0,00	0,00	0,00
2	13	133,00	1,73	-0,26	0,08	0,57	-0,99
3	574	2,03	0,33	-0,02	1,91	5,19	1,25
4	1134	0,53	-0,12	0,10	-2,19	-5,17	-1,42
N.R.	21	81,95	-2,24	-0,13	1,28	-0,24	1,60

Tabla 3.8: Valores test de las categorías suplementarias respecto a los ejes del ACM.

Clasificación sobre los factores El ACM deja entrever que hay diferentes tipos de individuos que quedan repartidos sobre el plano principal de forma que las posiciones de las categorías, que están en el centro de gravedad de los individuos que las han elegido, indican grupos distintos si están alejadas unas de otras. Una clasificación sobre los factores del ACM permite definir claramente y cuantificar grupos de individuos que son homogéneos internamente y diferentes entre sí, medido a través de las diferentes categorías existentes.

La clasificación realizada es de tipo mixto. En la Tabla 3.9 se detallan los últimos índices de nivel de los nodos de la jerarquía en la clasificación jerárquica. El histograma sugiere elegir una partición en 3, 4 o 6 clusters, que corresponden con los mayores saltos de los índices de nivel. Se decide proceder a la reubicación de centros e individuos a través del método de $K-$medias con una partición consolidada en 6 grupos.

Descripción de los nodos

Núm.	Prim.	Últ.	Efec.	Peso	Índice	Histograma de índices de nivel
75	70	58	8	278.00	0.00812	***
76	71	72	11	232.00	0.00886	***
77	67	68	7	225.00	0.01013	***
78	76	69	16	315.00	0.01520	*****
79	66	1	5	306.00	0.01615	*****
80	3	73	5	366.00	0.02374	*******
81	78	77	23	540.00	0.03466	**********
82	81	75	31	818.00	0.03848	**********
83	82	74	33	1070.00	0.06603	******************
84	83	79	38	1376.00	0.15956	**********...*****
85	80	84	43	1742.00	0.28665	**********...**********

Suma de índices de nivel = 0.71349

Tabla 3.9: Histograma de los índices de nivel de la clasificación jerárquica.

En la Figura 3.8 se muestran las posiciones de los centros de los clusters tras la reasignación iterativa de centros e individuos de la clasificación mixta. Los centros están representados por círculos de diámetros proporcionales a los tamaños de los clusters. Dada la interpretación de los ejes principales del ACM, el cluster etiquetado como 1, debería de ser el cluster de los individuos más proclives a comprar artículos corporativos y más satisfechos con la institución. En el lado contrario, en el cluster 6 deberían figurar los individuos menos proclives a comprar, menos satisfecho y jóvenes.

La caracterización completa de los clusters obtenidos puede realizarse mediante la comparación, para cada categoría, de los porcentajes de individuos que han elegido la categoría dentro del cluster y en el total de la muestra, así como examinando sus valores test, obtenidos de la comparación de dichas

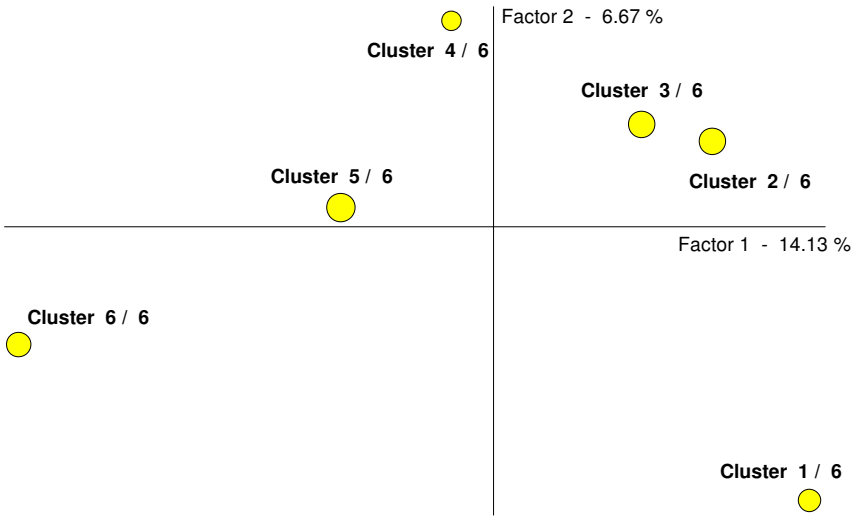


Figura 3.8: Clasificación sobre los factores del ACM. Centros y tamaños relativos de los clusters representados por círculos y sus diámetros.

categorías.

Las Tablas 3.10 a 3.15 contienen la descripción de las clases obtenidas a partir de la clasificación mixta, en base a las categorías de las variables ilustrativas categóricas. En la Tabla 3.16 se incluye la descripción en base a las variables ilustrativas continuas, que se realiza de forma similar, pero comparando las medias aritméticas de las variables continuas en el cluster y en el total de la muestra.

La descripción de los clusters se resume de la siguiente manera:

Cluster 1 Es un cluster de tamaño medio asociado a una mayor probabilidad de compra. Está integrado por individuos que comprarían a priori, muy satisfechos con la institución, de mayor edad, mujeres y profesores. Valoran principalmente los términos *Elegante, Sobrio, Tradicional* y *Artístico*.

Cluster 2 Su tamaño es medio y se asocia a personas que comprarían a priori y de edad más avanzada. Hay muy pocos que manifiestan su negativa a comprar. Predomina la valoración de los términos Tradicional y Sobrio. Es muy similar al cluster 1, salvo por el menor nivel de satisfacción con la institución.

Cluster 3 De tamaño también medio, se asocia a encuestados que comprarían a priori, bastante satisfechos, muy pocos se negarían a comprar y valoran los términos de Moderno, Original y Práctico. Es un grupo propenso a

la compra pero que se diferencia por los términos semiométricos, más valorados.

Cluster 4 De tamaño inferior a la media, es el primero de los grupos en los que se manifiesta una propensión de compra previa de los productos negativa. Solamente es característico, adicionalmente, que infravaloran el término *Elegante*; el resto aproximadamente igual que la media.

Cluster 5 De tamaño ligeramente superior a la media, es un grupo caracterizado por alumnos, más jóvenes (entre 18 y 22), que no comprarían, del campus de gipuzkoa y de satisfacción media. Valoran menos que la media términos como Tradicional, Sobrio, Elegante y Artístico.

Cluster 6 Este grupo de tamaño medio, sería el claramente menos interesado en comprar los artículos, de forma que hasta un 90,3 % manifiesta no estar interesado en comprar, incluso antes de ver los artículos, y son también en porcentaje decreciente, de satisfacción media y baja, de edad entre 30 y 44 y hombres. Los términos semiométricos reciben en general valoraciones muy bajas, en especial los de Elegante, Moderno, Tradicional y Sobrio.

En general, el análisis mediante ACM de las valoraciones de los artículos presentados muestra un perfil y una tipología de los individuos similar a la obtenida mediante el ACP de los términos generales semiométricos, que aquí tienen un papel meramente ilustrativo. Las personas de mayor edad y de mayor vinculación profesional y personal con la universidad son las más satisfechas con la institución y más proclives a comprar artículos con su logotipo. Además, como en el caso anterior, prefieren artículos de corte elegante y serio sobre otros de índole más moderna y atrevida. Por contra, los menos interesados suelen ser alumnos, más jóvenes, y que precisamente, valoran menos las características citadas en los artículos que pueden ser vendidos en la tienda universitaria.

Cluster 1 / 6 (Efectivo: 274 - Porcentaje: 15,73)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
BuyLogo	BuyLo=1	90,51	61,77	23,05	11,51	0,000	1076
Satis	Satis=5	50,00	31,92	24,64	6,74	0,000	556
Age	Age=4	38,32	30,25	19,92	3,05	0,001	527
Gender	MUJER	62,04	53,90	18,10	2,89	0,002	939
Link	PROFESORES	50,00	43,17	18,22	2,41	0,008	752
Satis	Satis=4	35,40	42,65	13,06	-2,59	0,005	743
Link	ALUMNOS	24,09	31,40	12,07	-2,82	0,002	547
Gender	HOMBRE	37,96	46,10	12,95	-2,89	0,002	803
Satis	Satis=3	10,22	19,29	8,33	-4,30	0,000	336
BuyLogo	BuyLo=2	7,30	36,51	3,14	-12,00	0,000	636

Tabla 3.10: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Primer cluster.

Cluster 2 / 6 (Efectivo: 312 - Porcentaje: 17,91)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
BuyLogo	BuyLo=1	84,29	61,77	24,44	9,45	0,000	1076
Age	Age=4	38,78	30,25	22,96	3,50	0,000	527
Education	Educa=4	71,15	65,10	19,58	2,44	0,007	1134
Education	Educa=3	26,92	32,95	14,63	-2,46	0,007	574
Satis	Satis=2	1,60	4,13	6,94	-2,52	0,006	72
Age	Age=2	8,33	13,09	11,40	-2,77	0,003	228
Age	Age=1	11,86	17,62	12,05	-2,97	0,002	307
Link	ALUMNOS	22,12	31,40	12,61	-3,92	0,000	547
BuyLogo	BuyLo=2	13,78	36,51	6,76	-9,70	0,000	636

Tabla 3.11: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Segundo cluster.

Cluster 3 / 6 (Efectivo: 309 - Porcentaje: 17,74)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
BuyLogo	BuyLo=1	88,35	61,77	25,37	11,30	0,000	1076
Satis	Satis=4	51,46	42,65	21,40	3,37	0,000	743
Satis	Satis=3	14,24	19,29	13,10	-2,46	0,007	336
BuyLogo	BuyLo=2	10,68	36,51	5,19	-11,14	0,000	636

Tabla 3.12: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Tercer cluster.

Cluster 4 / 6 (Efectivo: 223 - Porcentaje: 12,80)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
BuyLogo	BuyLo=2	52,91	36,51	18,55	5,28	0,000	636
BuyLogo	BuyLo=1	45,29	61,77	9,39	-5,27	0,000	1076

Tabla 3.13: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Cuarto cluster.

Cluster 5 / 6 (Efectivo: 325 - Porcentaje: 18,66)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
Link	ALUMNOS	42,77	31,40	25,41	4,74	0,000	547
BuyLogo	BuyLo=2	46,77	36,51	23,90	4,15	0,000	636
Age	Age=1	25,85	17,62	27,36	4,08	0,000	307
Campus	GIPUZKOA	31,69	25,37	23,30	2,79	0,003	442
Satis	Satis=3	24,31	19,29	23,51	2,42	0,008	336
Campus	BIZKAIA	53,23	60,05	16,54	-2,70	0,003	1046
Age	Age=4	23,69	30,25	14,61	-2,83	0,002	527
Link	PROFESORES	35,69	43,17	15,43	-2,97	0,001	752
BuyLogo	BuyLo=1	51,38	61,77	15,52	-4,17	0,000	1076

Tabla 3.14: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Quinto cluster.

Cluster 6 / 6 (Efectivo: 299 - Porcentaje: 17,16)

Variable	Categorías	% de cat. en grupo	% de cat. en total	% grupo en categ.	Valor test	Prob.	Peso
BuyLogo	BuyLo=2	90,30	36,51	42,45	21,41	0,000	636
Satis	Satis=3	32,11	19,29	28,57	5,80	0,000	336
Satis	Satis=2	9,03	4,13	37,50	4,11	0,000	72
Age	Age=3	45,15	37,26	20,80	3,01	0,001	649
Gender	HOMBRE	53,18	46,10	19,80	2,63	0,004	803
Gender	MUJER	46,82	53,90	14,91	-2,63	0,004	939
Age	Age=4	22,41	30,25	12,71	-3,24	0,001	527
Satis	Satis=4	33,44	42,65	13,46	-3,50	0,000	743
Satis	Satis=5	21,07	31,92	11,33	-4,48	0,000	556
BuyLogo	BuyLo=1	8,03	61,77	2,23	-21,47	0,000	1076

Tabla 3.15: Descripción de la clasificación en 6 clusters sobre los factores del ACM: Sexto cluster.

Variable	Media en cluster	Media total	Des. Típ. total	Des. Típ. cluster	Valor test	Prob.
Cluster 1 / 6 (Efectivo: 274)						
Stylish	5,630	4,850	1,359	1,625	8,58	0,000
Sober	4,295	3,738	1,566	1,644	6,08	0,000
Traditional	3,996	3,488	1,633	1,593	5,72	0,000
Artistic	5,255	4,823	1,531	1,631	4,75	0,000
Modern	5,561	5,299	1,226	1,441	3,25	0,001
Practical	6,232	6,002	1,166	1,326	3,11	0,001
Cluster 2 / 6 (Efectivo: 312)						
Traditional	3,823	3,488	1,484	1,593	4,10	0,000
Sober	4,067	3,738	1,529	1,644	3,91	0,000
Stylish	5,106	4,850	1,411	1,625	3,07	0,001
Cluster 3 / 6 (Efectivo: 309)						
Modern	5,497	5,299	1,276	1,441	2,64	0,004
Original	5,722	5,530	1,287	1,521	2,44	0,007
Practical	6,166	6,002	1,072	1,326	2,39	0,009
Cluster 4 / 6 (Efectivo: 223)						
Stylish	4,552	4,850	1,511	1,625	-2,92	0,002
Cluster 5 / 6 (Efectivo: 325)						
Artistic	4,608	4,823	1,775	1,631	-2,63	0,004
Stylish	4,635	4,850	1,723	1,625	-2,64	0,004
Sober	3,495	3,738	1,716	1,644	-2,94	0,002
Traditional	3,169	3,488	1,683	1,593	-3,97	0,000
Cluster 6 / 6 (Efectivo: 299)						
Original	5,238	5,530	1,885	1,521	-3,59	0,000
Artistic	4,493	4,823	1,874	1,631	-3,78	0,000
Practical	5,729	6,002	1,668	1,326	-3,86	0,000
Sober	3,310	3,738	1,730	1,644	-4,83	0,000
Traditional	2,955	3,488	1,598	1,593	-6,24	0,000
Modern	4,791	5,299	1,786	1,441	-6,55	0,000
Stylish	4,232	4,850	1,863	1,625	-7,09	0,000

Tabla 3.16: Descripción de la clasificación en 6 clusters sobre los factores del ACM: variables ilustrativas continuas.

3.5. Herramientas confirmatorias o predictivas

Las herramientas de corte confirmatorio o predictivo que se utilizan en esta sección difieren de la sección anterior en que hay al menos una variable que es el objeto de principal interés, y como tal, se toma como *dependiente* de otras de las variables disponibles. De hecho, estás técnicas están de alguna forma relacionadas con los modelos clásicos de regresión donde una variable, denominada dependiente, se intenta explicar a través de otras, denominadas explicativas, y cuya relación viene determinada por los correspondientes coeficientes de la relación. El foco se centra sobre estos coeficientes, cómo se estiman y que valores alcanzan dichas estimaciones.

Las dos técnicas que se van a emplear son de uso preferente en el ámbito del análisis de encuestas y del marketing: los modelos logit y los modelos de ruta o senda PLS. En ambos casos, la variable que se considera de interés, a explicar, será la disposición a comprar los productos corporativos que se muestran en la encuesta on-line.

3.5.1. PLS path modelling

En este apartado se busca la descripción de los datos de la encuesta EHUdenda a través de una técnica de modelización como es el PLS Path Modelling, o modelos de rutas PLS (Mínimos cuadrados parciales). En esencia se trata de modelizar una relación entre variables latentes que son construidas a partir de las variables disponibles, generalmente provenientes de una encuesta de opinión. Este tipo de análisis empieza a ser frecuente en estudios de marketing.

Inicialmente se trata de obtener el modelo de medida (de las variables latentes) o modelo interno. Las variables latentes se construyen a través de variables observadas, denominadas manifiestas. En ocasiones el significado o concepto de las variables latentes está predeterminado desde un principio, otras se deja a los datos sugerir cuál debe de ser aquél. Para que las variables latentes ξ_j tengan un significado definido de la manera más claramente posible, se elige una especificación reflexiva, según la cual es necesario que las variables manifiestas x_i que las conforman definan un grupo lo más unidimensional posible, en el sentido multivariante del término.

Las variables manifiestas elegidas son las valoraciones en escala de 1 a 4 de los 26 productos mostrados en la encuesta, que se reescalan. Las variables latentes se buscan de forma que se respete la condición de unidimensionalidad, mediante el examen de los ejes principales de un ACP de la tabla de dichas valoraciones. Es necesario para ello examinar bastantes ejes. Finalmente, se

llega a una estructura de 8 variables latentes, cuyas manifiestas aparecen detalladas en la Tabla 3.17. Las etiquetas y la numeración de las latentes han sido libremente escogidas.

Etiqueta	V. Latente	Productos
umbh	ξ_1	umbrella, hat
tie	ξ_2	tie, kerchief no.1, kerchief no.2
textiles	ξ_3	T-shirt, T-shirt-V, sweater, cap
bag	ξ_4	plastic tray, leather tray, backpack, bag, cup
wat	ξ_5	leather-strapped watch, metallic-strapped watch, wallet
mous	ξ_6	keyring, lighter, mousepad
scul	ξ_7	pin, sculpture
pens	ξ_8	blue pen, black pen, silver pen, silver pen in wooden case

Tabla 3.17: Variables latentes y sus variables manifiestas (valoraciones de los productos).

La Tabla 3.18, por su lado, contiene las pruebas de unidimensionalidad a partir de los estadísticos α de Cronbach, ρ de Dillon-Goldstein y los dos primeros valores propios de un ACP normado de las valoraciones de los productos que componen cada latente.

V. Latente	Dimen.	α Cronbach	ρ Dillon-Goldstein	λ_1	λ_2
Umbh	2	0,5693	0,8228	1,3979	0,6021
Tie	3	0,7875	0,8793	2,1409	0,7257
Textiles	4	0,8477	0,898	2,7535	0,5553
Bag	5	0,8132	0,874	2,8704	0,8481
Wat	3	0,645	0,8087	1,7549	0,6553
Mous	2	0,5873	0,8289	1,4157	0,5843
Scul	4	0,9209	0,9445	3,2396	0,4411
Pens	3	0,8524	0,9112	2,3235	0,4986

Tabla 3.18: Prueba de unidimensionalidad de las variables latentes parciales.

Los bloques de variables que definen las variables latentes mantienen la consistencia interna, dado que todos los segundos valores propios λ_2 de los ACP normados son menores que 1, todos los coeficientes ρ son elevados, claramente por encima de 0,7 y solamente 3 de los grupos poseen coeficientes α relativamente pequeños (umbh, wat y Mous), inferior a $0,7$ lo que lleva a considerar este modelo de medida como bastante aceptable.

Las variables ξ_i conforman propensión de compra de productos similares. Es decir, variables a las que los encuestados tienden a dar valoraciones similares,

quizás porque sean productos parecidos, como parece ser el caso.

Después se construye una variable latente que refleje una propensión global de compra, a partir de todas las variables manifiestas. Necesariamente las 26 variables manifiestas forman un grupo que no es o no tiene por qué ser unidimensional, por lo que se elige una especificación formativa para esta variable.

Finalmente se construye el modelo interno, que relaciona la variable latente global, la propensión global de compra, con las propensiones parciales de compra, de grupos similares:

$$\xi = \sum_j \beta_j \xi_j + \nu \quad \forall j = 1, \ldots, 8. \tag{3.1}$$

El modelo estimado figura en la ecuación (3.2), y muestra cómo la propensión global de compra depende principalmente de los grupos de bolígrafos *(Pens)*, prendas de vestir *(Textiles)* y el mix de artículos que hay en *bag* que son los que gozarían de una mayor aceptación. Los coeficientes, dado que las variables están tipificadas son, en realidad, coeficientes de correlación parciales, y dada la construcción de las variables latentes, todos aquellos valores mayores que $1/k = 1/8 = 0{,}125$ significan que el grupo correspondiente está particularmente relacionado con la propensión de compra o, dicho de otra manera, que ésta se debe principalmente a los productos que aparecen en el grupo.

$$\begin{aligned} \widehat{E(\xi)} \;=\; & 0{,}0865 * umbh + 0{,}1335 * tie + 0{,}2041 * textiles + 0{,}2114 * bag + \\ & + 0{,}1791 * wat + 0{,}1292 * mous + 0{,}0881 * scul + 0{,}2322 * pens \end{aligned} \tag{3.2}$$

La Figura 3.9 contiene información visual relevante, ya que muestra la especificación del diagrama de flechas que resume la especificación del modelo interno (relación entre variables latentes) y externo (relación entre latentes y manifiestas), así como la especificación, reflexiva o formativa, de las variables latentes dada por el sentido de las flechas que las unen. Se acompañan los coeficientes de correlación simples entre todas las variables, latentes y manifiestas, que permiten comprobar hasta qué punto las variables manifiestas están relacionadas con los constructos, y que no es inferior a 0,50 más que en unos pocos casos, a poca distancia y sólo para el caso de la propensión global de compra.

El modelo PLS permite construir una propensión global de compra y dilucidar cuáles de las variables son las principales determinantes del mismo, es decir, cuáles tienen mayor aceptación y con qué otros productos están asociados. Para conocer algo más, en particular, sobre las características de los individuos que han participado en la encuesta, podemos relacionar la propensión

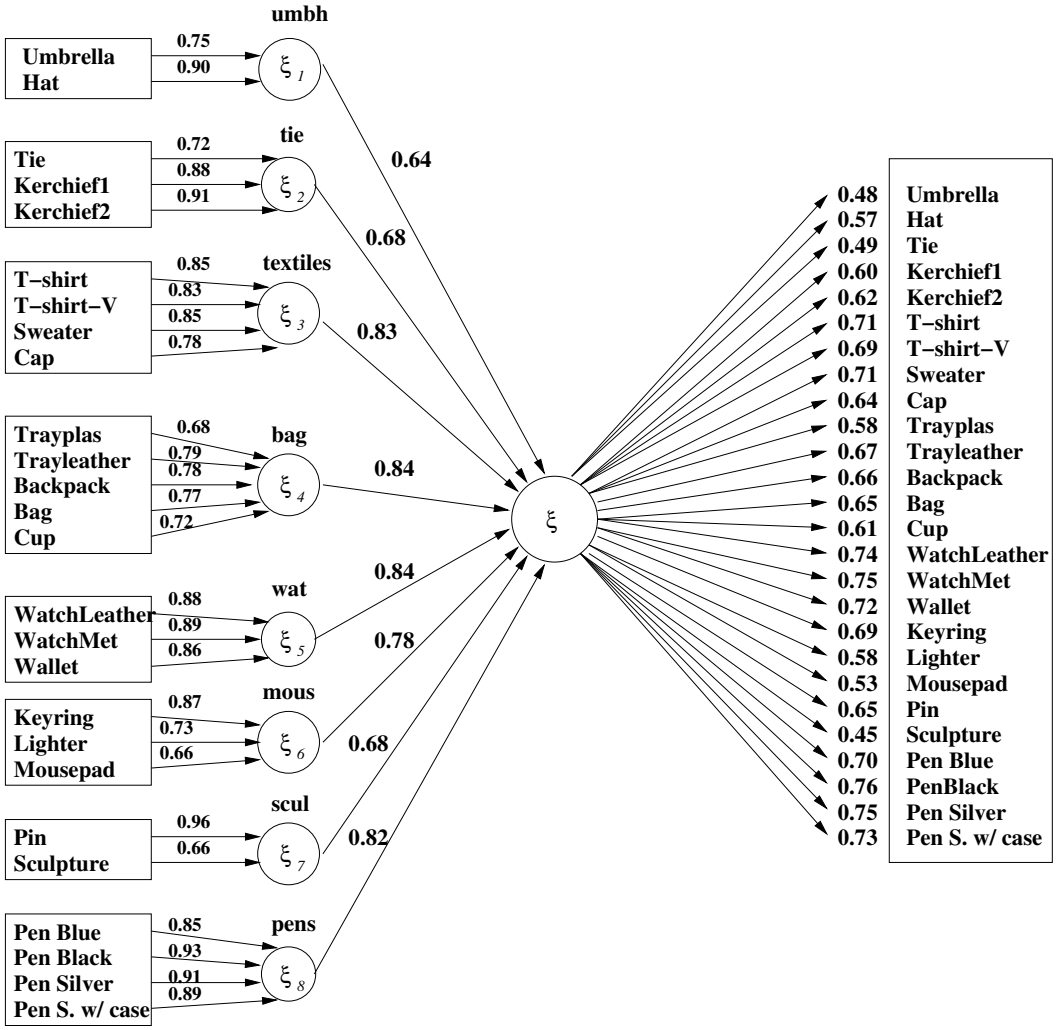


Figura 3.9: Diagrama de flechas del modelo externo PLS.

de compra construida, ξ con las variables caracterizadoras de los individuos, las variables que en ACM, por ejemplo, se proyectarían en suplementario.

Los modelos PLS son modelos de regresión, por lo que en este caso se opta por realizar una regresión de la variable latente general sobre todas las características conocidas de los individuos. Además, se añade la información sobre las valoraciones de los términos semiométricos que expresan características muy generales, como *original, atrevido* y otros. En este punto, los términos tienen en algunos casos valoraciones muy parecidas para la mayoría de individuos de la muestra, de manera que los 7 términos propuestos presentan altas correlaciones, provocando excesiva colinealidad. Es por esto, que se realiza una regresión en componentes principales parcial, en el sentido de que los compo-

nentes se extraen únicamente de las valoraciones de tales términos, dejando el resto de variables como están. Éstas, al ser cualitativas, son introducidas como variables ficticias. El resultado está en la ecuación (3.3).

$$\begin{aligned}\widehat{E(\xi)} \;=\; & -0,85 + 0,07 * \text{comp1 (original, atrevido, practico, artistico, moderno)} + \\ & +0,11 * \text{comp2 (tradicional, serio, elegante)} + 0,15 * \text{satisfecho} + \\ & +0,26 * \text{muy satisfecho} - 0,25 * \text{hombre} + 0,06 * \text{profesores} + \\ & +1,18 * \text{intencion previa de compra de un producto con el logo} \\ & +0,14 * \text{campus de Alava} + 0,12 * \text{campus de Vizcaya} \\ & -0,10 * \text{educacion superior} + 0,07 * \text{edad (+45)} \\ & R^2 = 0{,}4848\end{aligned} \tag{3.3}$$

El modelo estimado (3.3) muestra que la propensión de compra se debe más a las valoraciones de términos como Tradicional,Serio o Elegante que a la de términos como Original, Atrevido, Práctico, Artístico o Moderno, que tienen un coeficiente inferior. Asimismo, la satisfacción influye de manera positiva en la propensión de compra, así como la intención previa de compra de un producto con el logotipo, incluso antes de su visualización. Se advierte cómo los hombres tienen una propensión de compra inferior y, en menor medida, las personas de mayor edad y los profesores son más propensos a la compra de estos productos.

La significación de los coeficientes de la regresión en la ecuación (3.3) es difícil de realizar dada la falta de supuestos realizados sobre los términos de error de los modelos PLS, así como por la naturaleza de sus variables. En la Tabla 3.19 se muestran los coeficientes, junto con las desviaciones típicas y estadísticos t correctos bajo los supuestos de un modelo de regresión lineal general, más los intervalos de confianza bootstrap de los coeficientes. Éstos han sido generados a partir de 1000 muestras de tamaño igual al tamaño de muestra ($N = 1742$ individuos). Desde un punto de vista inferencial, habría que añadir, a lo dicho en el párrafo anterior, que los profesores no tienen una diferencia significativa con el resto, ni lo tiene la edad o el nivel educativo.

Los resultados de la modelización PLS ponen de manifiesto, al igual que las técnicas puramente descriptivas como el ACP o el ACM, que la intención de compra depende principalmente de la satisfacción con la institución y de la intención previa de compra. Otras características como el género o la edad emergen también como caracterizadoras de la propensión de compra en el mismo sentido pero, sin embargo, no siempre de forma significativa desde un punto de vista inferencial. Al mismo tiempo, se muestra de una manera visual la relación de la propensión de compra general con los productos y permite conformar grupos de productos similares. Todo esto desde un enfoque de

	Estim.	Error Est.	Estad. t	Cuantil 2.5 %	Cuantil 97.5 %
constante	-0.84788	0.06217	-13.638	-0.96821	-0.72230
comp1	0.06777	0.01286	5.271	0.04238	0.09387
comp2	0.11243	0.01463	7.683	0.08518	0.14148
satis4	0.15395	0.04995	3.082	0.05535	0.25248
satis5	0.26256	0.05491	4.782	0.13834	0.37877
male	-0.25432	0.03985	-6.382	-0.32960	-0.18057
professor	0.05829	0.05202	1.121	-0.04493	0.15292
buylogo	1.18159	0.04243	27.850	1.09303	1.26913
araba	0.13608	0.06253	2.176	0.01188	0.26209
bizkaia	0.12201	0.04601	2.652	0.03027	0.21372
edu.sup	-0.09743	0.05149	-1.892	-0.20196	0.00369
edad45	0.06956	0.04514	1.541	-0.01385	0.16045

Tabla 3.19: Regresión de la propensión global de compra sobre las variables caracterizadoras. Intervalos Bootstrap (1000 rep., tamaño = N) al 95 %.

modelización, diferente por tanto de las técnicas multivariantes y con algunas ventajas técnicas como la posibilidad de tratar un gran tamaño muestral, o la no necesidad de supuestos distribucionales.

3.5.2. Modelos Logit

En un modelo Logit la atención se centra en una variable que es discreta, sobre la que se pretende saber qué valores de otras variables llevan a tomar una decisión u otra, es decir, a elegir uno de los valores de la variable discreta. En el caso del conjunto de datos que estamos analizando, la variable de interés es si los individuos comprarían o no artículos de la tienda corporativa.

La variable mencionada, en realidad, se mide en la encuesta en dos momentos de tiempo claramente diferenciados:

1. De forma genérica, antes de visualizar ningún artículo con el logotipo de la UPV/EHU.

2. De forma indirecta y, sin embargo, concreta, en la medida de que se visualizan 26 artículos con el logotipo de la UPV/EHU y se pregunta si se comprarían separadamente en una escala de 1 a 4 cada uno de ellos.

En ambos casos el objetivo es estimar la probabilidad de compra de artículos con el logotipo de la UPV/EHU por parte del personal de la institución y expresarla en función de sus características personales y de sus preferencias

sobre las características de los productos, así como de su grado de satisfacción respecto a la institución.

La probabilidad de compra se define de la manera habitual en los modelos logit, tal y como aparece en la ecuación (3.4),

$$P(y_i = 1) = F\left(\sum_j x_{ij}\beta_j\right) = \frac{e^{\sum_j x_{ij}\beta_j}}{1 + e^{\sum_j x_{ij}\beta_j}} \tag{3.4}$$

donde y es una variable dicotómica que toma el valor $y_i = 1$ si el individuo i compra y el valor $y_i = 0$ si no lo hace.

En el conjunto de datos disponible, la probabilidad de compra viene dada por el deseo manifestado por cada individuo sobre cada uno de los 26 artículos concretos mostrados, valorado en una escala de 1 a 4. Para utilizar una probabilidad global de compra, basada en los resultados de tales valoraciones, se considera que un individuo es un comprador (potencial) de productos de la tienda si valora como probable o muy probable (puntuaciones 3 y 4) la compra de, al menos, el 25 % de los productos de la tienda. Esto equivale a valoraciones positivas de, al menos, 7 productos.

Las variables explicativas son las mismas que se usaron en la sección anterior, PLS path modelling, y que se usaron como ilustrativas en la sección de ACM. Esto es, satisfacción, características personales y valoración de características deseables de los productos utilizando términos semiométricos. Al igual que en el caso de PLS, algunos de los términos semiométricos son altamente colineales, por lo que los 7 términos se han reducido a tres componentes principales obtenidas a partir de un ACP previo, denominadas *innovador* (similar al componente 1 usada en PLS), *clásico* (igual al componente 2 del PLS) y *práctico* (componente asociada en exclusiva al término *práctico*).

La Tabla 3.20 contiene los coeficientes estimados del modelo. Como en análisis anteriores, la probabilidad de compra depende de forma muy significativa del interés previo de compra (antes de visualizar los productos) y de la satisfacción con la institución. También aparecen las componentes asociadas a los términos asociados con lo clásico y el término práctico, y el género, en el mismo o parecido sentido que en anteriores ocasiones. A diferencia de los análisis previos, el estamento resulta significativo, pero no de igual forma; los profesores resultan más proclives a comprar productos con el logotipo (respecto a la categoría omitida, personal de administración y servicios), pero más aún los alumnos.

No hay ninguna razón evidente que justifique el resultado anterior respecto del grupo de alumnos, más allá del hecho de que se trata de análisis diferentes y no directamente comparables. Se ha probado a estimar otros modelos logit similares a éste pero para cada uno de los 8 subconjuntos de artículos de

Variable	Categoría	Coeficiente	Sig.
Estamento	Alumno	0,805	**
	Profesor	0,409	*
Sexo	Hombre	-0,471	**
Campus	Alava	-0,027	
	Vizcaya	0,200	
Edad	Menos de 23	0,400	
	23-29	0,015	
	30-44	0,118	
Carácter	Innovador	0,133	
	Clásico	0,263	**
	Práctico	0,187	**
Satis	Bastante Satis	0,529	**
	Muy Satis	0,428	*
Int.ini		3,642	**
Constante		-0,830	
R^2	(Nagelkerke)	0,441	

Tabla 3.20: Coeficientes estimados del modelo logit para la probabilidad de compra de artículos con logotipo (** = significativo al 5 %, * = significativo al 10 %).

valoraciones similares utilizados en el PLS path modelling (Tabla 3.17) para la estimación de las variables latentes parciales. Se observa cómo es cierto que los alumnos valoran de forma positiva y significativa la compra de los artículos de los grupos 4 (bag), 6 (mous) y 8 (pens) que son artículos que incluyen mochilas, bolsos, llaveros, mecheros y bolígrafos, siempre mediante modelos logit como el descrito.

Finalmente, la medida de bondad de ajuste empleada en el modelo logit está muy en línea con la que se obtuvo en el modelo PLS (ver ecuación (3.3)), y manifiesta un ajuste medio-bajo, que no es despreciable sin embargo, dada la cantidad de variables indicadores que aparecen en la regresión.

3.6. Conclusiones

En este capítulo se han utilizado métodos estadísticos que sirven de apoyo a la Minería de Datos. Los métodos incluyen, además de técnicas exploratorias multivariantes, técnicas de carácter predictivo como el Partial Least Squares (PLS) path modelling y los modelos logit, a todas las cuales se ha hecho referencia en el capítulo 2. El criterio de selección de estas técnicas ha sido la

naturaleza del estudio empírico realizado y los objetivos perseguidos con el mismo: Viabilidad de una tienda corporativa de la Universidad del País Vasco (UPV/EHU).

La matriz de datos, obtenida a través de una encuesta on-line, está integrada por numerosas variables de distinta naturaleza, si bien la mayoría son variables cualitativas que reflejan, por un lado, las valoraciones de los encuestados sobre los productos corporativos y, por otro lado, las características personales de los individuos, algunas de ellas relacionadas con la institución objeto de estudio.

Las diferentes técnicas aplicadas sobre los datos así definidos han permitido obtener muchas y muy diversas conclusiones. Las divergencias entre las conclusiones están asociadas a los específicos objetivos de cada una de las metodologías. No obstante, se complementan entre sí, enriqueciendo notablemente los resultados del estudio. Los métodos exploratorios como el Análisis de Componentes, Correspondencias o Clasificación, ayudan a describir la información a veces eclipsada dentro de un conjunto relativamente grande de datos, incluso a nivel de grupos de individuos relativamente homogéneos, como son los clusters. Por otro lado, los métodos predictivos, como el PLS o los modelos logit, permiten modelizar el comportamiento de los individuos, utilizando herramientas inferenciales para buscar y seleccionar un modelo mejor o para establecer claramente las características de los individuos. Al mismo tiempo, es posible computar medidas de bondad de ajuste que permiten evaluar la adecuación del modelo seleccionado a los datos existentes.

Señalar, finalmente, que el estudio realizado y sus conclusiones han resultado de gran utilidad para la UPV/EHU. Los análisis efectuados permiten establecer las características de los individuos que son los más probables compradores de los productos corporativos y los que, en última instancia, serían los principales soportes de la viabilidad de una tienda que venda este tipo de productos. Al mismo tiempo, es posible identificar los tipos de individuos que no están interesados en los mismos, así como algunas características de los mismos, lo que abre la puerta a la consideración de otros productos diferentes, de características conocidas, o intuidas, para su posible incorporación a la tienda corporativa en el caso de que se quiera acceder a este tipo de clientes con una mayor probabilidad de éxito. Estas conclusiones son valiosas desde un punto de vista de objetivos de mercado y proporciona pautas útiles de marketing en este caso concreto.

CAPÍTULO 4

Tablas Múltiples de tablas de efectivo diferente

4.1. Introducción

El análisis multivariante de datos se centra en el tratamiento de tablas de datos donde habitualmente las filas representan individuos (personas, empresas, países) y las columnas variables que miden características observadas de los individuos. Cuando el número de columnas ó variables presentes es grande, de forma que la cantidad de información disponible excede con mucho la capacidad de interpretación visual de la tabla, ya sea de forma directa ó a través de gráficos bidimensionales, es cuando el análisis multivariante de datos es de máxima utilidad.

Una extensión del análisis se produce cuando tiene sentido estructurar a priori el conjunto de variables ó columnas disponibles en diversos grupos en los que las variables en el seno de cada grupo tienen alguna relación difícilmente extraíble de los propios datos. De esta forma cada grupo constituye una tabla, comúnmente denominada subtabla, que puede todavía contener un número elevado de variables. De esta forma, tendríamos un conjunto de tablas cada una con un número posiblemente elevado de filas y columnas pero que están relacionadas, bien sea por contener individuos similares, por medir características comunes, o los dos aspectos a la vez. Por ejemplo, puede pensarse en tablas que contengan variables económicas como saldos de los diversos elementos de la balanza de pagos (en columnas) medidas en un conjunto de países (en filas), y todo ello para varios años. Esto es, se tendría un conjunto de tablas donde

cada de ellas se asocia a cada uno de los respectivos años. El estudio de la información así estructurada permitiría poner de manifiesto posibles trayectorias temporales en las relaciones entre las variables. Otro ejemplo habitual se produce con información procedente de una misma encuesta realizada en diferentes países. En este caso se dispondrían en filas los individuos encuestados y en columnas las preguntas medidas en la encuesta, obteniendo una tabla para cada uno de los países analizados. En este otro caso, el objetivo se centraría en detectar las posibles diferencias y semejanzas entre las relaciones de las preguntas entre los países. Dentro de este marco de análisis de tablas múltiples, hay varias metodologías con posibilidades de aplicación, como son el Análisis Factorial Múltiple de Escofier & Pagés (1998), la metodología Statis desarrollada por el grupo del profesor Y. Escoufier cuyos primeros trabajos aparecieron publicados en L'Hermier des Plantes (1976) y posteriormente Lavit (1988).

Los métodos citados anteriormente presentan una limitación que puede ser en ocasiones determinante. En particular, es la que se refiere a la necesidad de que las tablas sean de igual efectivo, esto es, de que el número de individuos contenido en cada tabla sea el mismo. En este capítulo se trabaja dicho problema y se propone una extensión como solución que permite la incorporación de todos los individuos de las subtablas, aunque éstas sean de distinto tamaño.

4.2. Herramientas de análisis de Tablas Múltiples

En frecuentes ocasiones se realizan investigaciones en las que la información puede presentar diferentes estructuras de comportamiento. Estas diferentes estructuras pueden quedar eclipsadas en los análisis de la información en su conjunto. Por ello, esta información debe ser estudiada desde la óptica de tabla múltiple, esto es, teniendo en cuenta la existencia de diversos grupos.

En el extenso campo de investigación del Análisis de Datos existen numerosas y diversas técnicas desarrolladas para el estudio exploratorio de tablas susceptibles de ser estructuradas en subtablas (también denominadas tablas múltiples, tablas de tres dimensiones o tablas de tres entradas) Esta estructuración puede estar asociada tanto a la dimensión de las columnas (variables) como a la dimensión de las filas (individuos). Las investigaciones en las que se dispone de información de esta naturaleza múltiple tienen objetivos más ambiciosos, ya que no se limitan a la búsqueda de relaciones entre variables y tipologías de los individuos, sino que se amplían al análisis comparativo de las realidades presentes en el seno de cada una de las tablas. Esta riqueza interpretativa, junto con la gran casuística de datos, ha animado a muchos

investigadores a desarrollar metodologías apropiadas a estos objetivos.

Este trabajo se enmarca en una línea de investigación iniciada en la década de los 80 por los profesores Brigitte Escofier y Jérôme Pagés en el seno de la Escuela Francesa de Análisis de Datos, el Análisis Factorial Múltiple o AFM (Escofier & Pagés 1986, 1992, 1994, 1998). Desde que estos autores pusieron en conocimiento de la comunidad científica sus avances sobre el tratamiento de tablas múltiples, son muchos los investigadores que se han sumado a su filosofía y han contribuido a afianzar el AFM como una metodología con una gran versatilidad en el tratamiento de información de tres dimensiones. Así lo ponen de manifiesto los numerosos trabajos que en los últimos años han visto la luz y han consolidado esta escuela, poniendo de manifiesto su potencialidad, tanto desde el punto de vista teórico como empírico, en muy diversas áreas. De la vasta literatura científica relacionada con esta técnica merecen ser destacados los siguientes trabajos: Pagés (1996, 2005), Pagés & Tenenhaus (2001, 2002) y Husson & Pagès (2006*a*,*b*) en los que se compara, reflejando similitudes y diferencias, el AFM con otros métodos como el modelo INDSCAL, el análisis Procusto o la metodología STATIS; Bécue-Bertaut & Pagès (2004) con el método MFACT y Goitisolo (2002), Zárraga & Goitisolo (2002, 2006), Goitisolo & Zárraga (2008), Zárraga & Goitisolo (2009) con el Análisis Simultáneo para el análisis múltiple de tablas de contingencia; Pagès (2002, 2004), Pagés & Camiz (2008), Bécue-Bertaut & Pagès (2001, 2004, 2008), Abascal et al. (2006) que presentan una extensión del AFM para el tratamiento de tablas mixtas y de tablas de frecuencias; Le Dien & Pagés (2003) una adaptación del AFM para el tratamiento de encuestas en las que las preguntas están agrupadas en temas (estructuración jerárquica); Le Dien & Pagés (2010) una extensión del AFM para el tratamiento simultáneo de variables cuantitativas medidas en varios grupos de individuos; Morand & Pagès (2007) una extensión del AFM incorporando las rotaciones procustas; Abascal et al. (2001, 2004, 2008), Landaluce (1995), Landaluce et al. (1999), García Lautre (2001) y García Lautre & Abascal (2003), en los que se recogen diversas aplicaciones del AFM.

Todos estos trabajos ponen de manifiesto que el AFM con el tiempo y uso ha pasado de ser una técnica de análisis de tablas múltiples a toda una filosofía de análisis comparativo, tanto gráfico como a través de indicadores numéricos, de diferentes conjuntos de datos, sea cual sea su naturaleza y su estructura. Este capítulo se enmarca en este contexto de adaptaciones, extensiones y aplicaciones del AFM.

4.2.1. El Análisis Factorial Múltiple

4.2.2. Objetivos

El AFM es un método factorial estrechamente relacionado con el ACM y el ACP para el análisis de tablas de datos que contienen información sobre individuos descritos a partir tanto de variables continuas como categóricas. La diferencia con respecto a aquellos estriba en que las variables están dispuestas en torno a varios grupos, esto es, subtablas de variables del mismo tipo. Con respecto a los otros métodos de tablas múltiples, le diferencia especialmente el peso asignado a cada tabla en el análisis global con el objetivo de equilibrar la influencia de cada una de ellas.

La diferenciación entre subtablas tiene una justificación externa al análisis. Puede ser debida a las propias características de las variables (como, por ejemplo, puntuaciones de diferentes aspectos valorados en una cata de vinos, dando lugar a variables relacionadas con olfacción, gustación o presencia visual, siguiendo el ejemplo clásico de (Escofier & Pagés 1998, cap. 6)), a una diferenciación espacial (una misma encuesta llevada a cabo en diferentes países, por ejemplo, de la Unión Europea) o a una diferenciación temporal (una encuesta realizada a los mismos individuos en momentos de tiempo diferentes, como un panel de usuarios de transporte público).

En todos los casos citados el interés no se centra solamente en obtener una tipología de individuos en base a las variables disponibles, sino que se extiende a la búsqueda de las posibles relaciones entre las subestructuras correspondientes a cada grupo de individuos. El AFM ofrece para ello resultados específicos sobre la estructura en grupos de la tabla, lo que incluye a la relación entre los grupos, los factores comunes a los grupos y la representación simultánea de los individuos y las variables, tanto desde un punto de vista global como desde el de los grupos a los que pertenecen. En definitiva, se trata de analizar la estructura común de las subtablas especificadas, poniendo de manifiesto cuáles son los elementos comunes y los elementos discordantes en dicha estructura.

4.2.3. Metodología

Una vez se ha definido la tabla global como una yuxtaposición de tablas con diferentes grupos de variables, el AFM consiste en un Análisis de Componentes Principales para las subtablas que contienen variables continuas y en un Análisis de Correspondencias Múltiples para las subtablas compuestas por variables categóricas. En realidad, también en este caso, el ACM puede verse como un ACP sobre la subtabla de variables indicadoras de las categorías, utilizando una ponderación adecuada sobre las variables indicadoras (Landaluce

1995).

Los factores del AFM, denominados factores globales, se extraen de un ACP normado de la tabla global compuesta por las subtablas de variables continuas y de indicadoras de las modalidades ponderadas. Sin embargo, como se ha dicho antes, el AFM busca equilibrar la influencia de cada subtabla, para que ninguna determine por sí sola, en un caso extremo, o en parte excesiva los factores globales que se van a extraer.

Previamente a la extracción de los factores globales del AFM, se realizan análisis separados (ACP o ACM) de las subtablas, denominados análisis parciales. De ellos puede obtenerse información relativa a su estructura interna. Aún más importante, de los valores propios de dichos análisis parciales se extrae la ponderación que equilibra la influencia de cada una de las subtablas. Se selecciona, para cada una, el primer, el mayor valor propio del análisis parcial y se usa la raíz cuadrada del inverso como ponderación en el análisis global. Esto es, analiza la tabla múltiple $\boldsymbol{X}$ que surge al yuxtaponer de forma horizontal los diferentes $j = 1, ..., J$ grupos considerados. La ponderación seleccionada es el inverso del primer valor propio λ_1^j obtenido en los análisis separados de cada grupo $\boldsymbol{X}_j$. En definitiva, se realiza el ACP sobre la matriz global definida esquemáticamente a través de la ecuación (4.1).

$$X = \left[\frac{1}{\sqrt{\lambda_1^1}}\boldsymbol{X}_1 \,\middle|\, \frac{1}{\sqrt{\lambda_1^2}}\boldsymbol{X}_2 \,\middle|\, \cdots \,\middle|\, \frac{1}{\sqrt{\lambda_1^j}}\boldsymbol{X}_j \,\middle|\, \cdots \,\middle|\, \frac{1}{\sqrt{\lambda_1^J}}\boldsymbol{X}_J \right] \tag{4.1}$$

Esta manera de proceder permite alcanzar dos objetivos:

1. Ninguna subtabla puede determinar, al menos, el primer factor del análisis global. La inercia de la primera componente de cada grupo queda de esta manera normalizada al valor 1.

2. La estructura interna dentro de cada grupo queda inalterada, ya que el peso es común a todas las variables del grupo.

La distancia al cuadrado entre dos individuos queda definida, una vez efectuada la ponderación sobre los grupos, por:

$$d^2(i,l) = \sum_{j \in J} \alpha_j d^2(i^j, l^j) \qquad \text{donde} \quad \alpha_j = \frac{1}{\lambda_1^j} \tag{4.2}$$

En AFM, asimismo, las relaciones de transición entre los factores del espacio de los individuos y los de las variables se derivan de las relaciones de transición del ACP. Así, la relación de transición en el ACP es:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K} x_{ik} G_s(k) \tag{4.3}$$

donde $F_s(i)$ es la coordenada del individuo i sobre el eje s, λ_s es el valor propio asociado al eje s, x_{ik} el valor que toma la variable k para el individuo i y $G_s(k)$ la coordenada de la variable k sobre el eje s.

En el caso de que la variable k se vea afectada por el peso m_k,

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K} x_{ik} m_k G_s(k) \tag{4.4}$$

A partir de esta expresión y con la introducción del peso asignado por el AFM a la subtabla j, esto es, $1/\lambda_1^j$, se sostiene que la relación de transición entre los factores globales de un AFM resulta como sigue:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J} \frac{1}{\lambda_1^j} \sum_{k \in K_j} x_{ik} G_s(k) \tag{4.5}$$

mientras que la relación de transición entre factores parciales del AFM es, en el seno del grupo j,

$$F_s(i^j) = \frac{1}{\sqrt{\lambda_s}} \frac{J}{\lambda_1^j} \sum_{k \in K_j} x_{ik} G_s(k) \tag{4.6}$$

de forma que la relación de transición para la tabla total puede escribirse como la media aritmética de las relaciones parciales:

$$F_s(i) = \frac{1}{J} \sum_{j \in J} F_s(i^j) \tag{4.7}$$

4.2.4. Herramientas de interpretación de los resultados

Una vez se han efectuado los análisis parciales de las J subtablas y el análisis conjunto, se tienen $J+1$ conjuntos de resultados habituales, conteniendo entre otros, valores propios y factores. Constituyen las herramientas básicas para la determinación e interpretación de la estructura en grupos de la tabla.

Los resultados sobre la tabla global incluyen a los clásicos en ACP o ACM, como coordenadas, contribuciones, cosenos cuadrados y correlaciones entre variables continuas y factores, que son de utilidad solo para dicho análisis global. Los más interesantes son los relacionados con la estructura en grupos, como se expone a continuación.

Correlaciones entre factores parciales y globales

Los factores parciales (provenientes de los análisis parciales de las subtablas) son específicos de cada subtabla, mientras que los factores globales (que provienen de la tabla completa incluyendo la ponderación por grupos del AFM) representan el comportamiento global. Con dicha ponderación, el valor máximo de puede tomar el primer valor propio global es igual a J, el número de grupos y se obtiene en el caso en que la dirección global de mayor inercia es común a todos los grupos. Sobre el resto de valores propios, su valor refleja la estructura interna de cada grupo (siendo menores que los primeros).

La semejanza de los grupos con respecto a la estructura global se mide a través de correlaciones entre los factores parciales del mismo orden y los factores globales, $corr(F_i^j, F_i),\ \forall j = 1, ..., J$. Una correlación elevada implica que el comportamiento representado por el factor está presente en todos los grupos, mientras que una correlación baja es indicadora de que el comportamiento es específico de algún grupo en particular.

En esta fase, si los factores parciales de un grupo presentan correlaciones bajas con los globales implica que se trata de un grupo con escasa comunalidad en su estructura interna con el resto y, en consecuencia, debería desecharse en un análisis conjunto de este tipo, al menos como grupo activo. Esto no impide su consideración en suplementario.

Calidad de representación de los grupos: Coeficientes de ligazón

Independientemente de la relación entre factores globales y parciales, es necesario determinar cuán relacionados están los grupos con los factores, de la misma manera que en ACP se usan las correlaciones entre variables y factores.

Coeficiente de ligazón $\mathcal{L}_g$ Para calcular la correlación entre una variable y un grupo de variables, se puede usar el cuadrado del coeficiente de correlación múltiple (Carrol 1968). Es sabido, sin embargo, que este coeficiente es poco fiable, sobreestimando el grado de asociación, en el caso en que las variables del grupo estén muy correlacionadas.

En el contexto de tablas múltiples es más frecuente utilizar la inercia proyectada de la tabla compuesta por el grupo de variables sobre la variable cuya correlación es de interés. Para ello, se computa

$$W_j = X_j X_j'$$

la matriz de productos escalares entre los individuos considerando las variables del grupo j (posiblemente afectadas por una matriz de pesos M_j, en cuyo caso $W_j = X_j M_j X_j'$). Se calcula el coeficiente de ligazón entre una variable z

y una tabla de variables X_j (posiblemente afectadas por la matriz de pesos de individuos D) como el producto escalar entre zz' y W_j:

$$\mathcal{L}_g(z, X_j) = \langle zz'D, W_jD \rangle \tag{4.8}$$

Este operador coincide con el coeficiente de correlación múltiple al cuadrado cuando las variables no están correlacionadas. Además, es menos sensible ante pequeñas variaciones de los datos (Escofier & Pagés 1998, sec. 7.3.4.3)).

Ligazón entre dos subtablas Si, fuera del estricto ámbito del AFM, interesa una medida de asociación o ligazón entre dos subtablas, de manera similar a lo que representa un coeficiente de correlación entre dos variables, puede utilizarse el producto escalar entre dos grupos de variables

$$CovV(K_j, K_l) = \langle W_jD, W_lD \rangle$$

que no está acotado superiormente y está más relacionado con lo que sería una covarianza en el caso univariante. Como medida similar al coeficiente de correlación, puede utilizarse el coeficiente RV de Escoufier (1973) y Robert & Escoufier (1976).

$$RV(K_j, K_l) = \left\langle \frac{W_jD}{||W_jD||}, \frac{W_lD}{||W_lD||} \right\rangle \tag{4.9}$$

El coeficiente RV tiene la interesante propiedad de que $0 \leq RV \leq 1$. Sin embargo, no recoge la relación entre las subtablas del AFM en cuanto no incluye la ponderación de las tablas.

Desde el punto de vista del AFM, tiene más sentido utilizar como coeficiente de ligazón el producto escalar entre subtablas, incluyendo la ponderación:

$$\mathcal{L}_g(K_j, K_l) = \left\langle \frac{W_jD}{\lambda_1^j}, \frac{W_lD}{\lambda_1^l} \right\rangle \tag{4.10}$$

Es interesante notar que en AFM la norma del grupo K_j es:

$$\left|\left| \frac{W_jD}{\lambda_1^j} \right|\right| = \sum_s \left(\frac{\lambda_s^j}{\lambda_1^j} \right)^2$$

y se interpreta como una medida de dimensionalidad de la tabla K_j. Esta norma es mayor cuanto mayor es el número de factores parciales de importancia similar al primero de ellos. En el caso extremo de que el grupo solo contenga variables tipificadas incorrelacionadas entre sí, la norma es igual al número de variables.

Esto significa que el coeficiente $\mathcal{L}_g$ es mayor cuanto más multidimensionales son los grupos K_j y K_l y cuantas más direcciones comunes de inercia importante presenten. Comparativamente, el coeficiente RV no se ve afectado por la dimensionalidad de la estructura común. En este sentido, ambas medidas son complementarias, eligiendo el coeficiente $\mathcal{L}_g$ cuando sea importante tener en cuenta la dimensionalidad de los grupos.

Calidad de representación de una subtabla sobre un factor La calidad de representación de una subtabla sobre un eje global se puede medir a través del coeficiente de ligazón entre la tabla K_j y el factor global correspondiente:

$$\mathcal{L}_g(K_j, F_s) \tag{4.11}$$

En la práctica, dada la ponderación del AFM, los grupos suelen estar bastante bien representados sobre el primer factor global.

En la práctica puede visualizarse en un único plano la calidad de representación de todas las subtablas para el subespacio de $\mathbb{R}^2$ engendrado por los ejes s y s' si se dibujan todos los puntos $\left(\mathcal{L}_g(K_j, F_s), \mathcal{L}_g(K_j, F_{s'})\right),\ \forall j = 1, ..., J$

Relación inercia inter/total

El AFM permite, en una aplicación del principio de Huygens, descomponer la inercia de la tabla global, o de un factor global de la tabla en particular, como la suma de inercia inter grupos más la inercia intra grupos. El cociente

$$\frac{\text{inercia inter}}{\text{inercia total}}$$

es siempre inferior a 1 y, cuando se calcula para un factor en particular da una idea de cuán común a los diferentes grupos es dicho factor y del sentido que tiene el análisis múltiple para la tabla en cuestión. Es particularmente interesante que esta magnitud sea elevada para, al menos, los primeros factores globales.

Representaciones gráficas

Desde el punto de vista de la interpretación, el AFM dispone de diversas representaciones gráficas, análogas a otros análisis factoriales. A éstas hay que añadir las representaciones relacionadas con la relación intra grupos propia del AFM como análisis de tablas múltiples.

Representación de los individuos Es posible representar en un subespacio de $\mathbb{R}^2$ mediante un diagrama de dispersión la proyección de la nube media de individuos como en un ACP. Sin embargo, es posible también proyectar las J nubes parciales de forma separada sobre los mismos ejes globales. Para ello, se define la matriz

$$\tilde{X}_j = \left[\; 0 \; \middle| \; X_j \; \middle| \; 0 \; \right] \tag{4.12}$$

la cual se puede proyectar en suplementario sobre los ejes globales para obtener la proyección de la nube parcial j sobre el eje s:

$$F_s^j = \tilde{X}_j M u_s = \frac{1}{\sqrt{\lambda_s}} \tilde{X}_j M X' D F_s = \frac{1}{\sqrt{\lambda_s}} W_j D F_s \tag{4.13}$$

siendo u_s el vector unitario correspondiente al eje s y teniendo en cuenta las habituales relaciones de transición del ACP.

Representación de las variables En el caso de las variables, la representación puede hacerse, sobre los factores globales o parciales, siendo claramente de mayor interés la proyección sobre los globales, lo que no tiene ninguna diferencia reseñable respecto de una proyección correspondiente a un método factorial habitual como el ACP.

Representación de los grupos Como se ha citado en la página 99, la representación gráfica de los grupos se corresponde con el diagrama de dispersión de los puntos compuestos por los coeficientes de ligazón entre tablas y factores, $\mathcal{L}_g(K_j, F_s)$.

Elementos suplementarios

Como en otros análisis, es posible proyectar en suplementario subtablas adicionales. Sin embargo, debe realizarse una normalización similar a la de las tablas activas cuando se pretende obtener y representar sobre los ejes globales las proyecciones de las variables suplementarias. También es posible obtener coeficientes de ligazón para considerar su calidad de representación o la relación con las tablas activas mediante coeficientes RV.

Sin embargo, no es posible obtener una representación gráfica en suplementario de nubes de individuos asociadas a variables suplementarias, dados los espacios no coincidentes en que se proyectan la nube suplementaria y la global. No obstante, pueden aún calcularse otras medidas de similaridad, como las correlaciones entre componentes principales y factores globales o coeficientes de ligazón. En cualquier caso, desde un punto de vista interpretativo, es

mucho más relevante la relación entre las variables de las tablas que entre los individuos de las tablas.

4.3. Análisis de tablas de efectivo diferente

Como se ha expuesto anteriormente, el análisis de tablas múltiples se interesa por el análisis de diversas tablas rectangulares de forma que al interés existente por la relación entre individuos por un lado, y variables por otro, se añade otro consistente en la relación entre las tablas.

La metodología AFM expuesta en la sección 4.2.1 se dirige a este tipo de problemática, de una forma en que, además, equilibra la influencia de las tablas, limitando que una o unas pocas de las subtablas dominen el análisis global. El análisis AFM, como ya se ha comentado, tiene un mayor sentido cuando las subtablas están suficientemente relacionadas; lo contrario supondría que un análisis conjunto no tiene sentido.

Desde esta perspectiva, es habitual en el ámbito del análisis de encuestas la existencia de tablas de datos de variables categóricas provenientes de las preguntas formuladas en la encuesta. En este contexto, el análisis de tablas múltiples cobra sentido cuando se dispone de una encuesta que se realiza de forma repetida en un ámbito temporal o espacial. En este caso, las variables objeto de la encuesta son idénticas alcanzando, en el caso de las variables categóricas, a la escala (y composición de la misma) en que se miden (sean ordinales o no).

Cuando se considera un análisis de tablas múltiples donde las subtablas se caracterizan por contener las mismas variables, es frecuente que, sin embargo, contengan distintos, y diferente número de, individuos. Esto sucede simplemente por cualquier variación en el muestreo realizado. Podemos tener diferentes entidades (por ejemplo, países) con diferentes tamaños, diferentes presupuestos para la muestra y diferentes tasas de respuesta. Pero incluso en el caso en el que el número de individuos sea igual, es claro que son anónimos y son distintos de una tabla a otra (como lo son los entrevistados en un país y otro).

Las consideraciones anteriores hacen que el AFM sea una herramienta de nula o escasa utilidad cuando las subtablas contienen individuos que son distintos. En esta sección se propone una alternativa que permite una aplicación del método en este tipo de situaciones considerando información relevante sobre relaciones entre las variables y los grupos conformados por las subtablas, al tiempo que se mantiene una de las ventajas del AFM, como es la ponderación de las subtablas.

La Figura 4.1 contiene el esquema de la metodología propuesta, que se

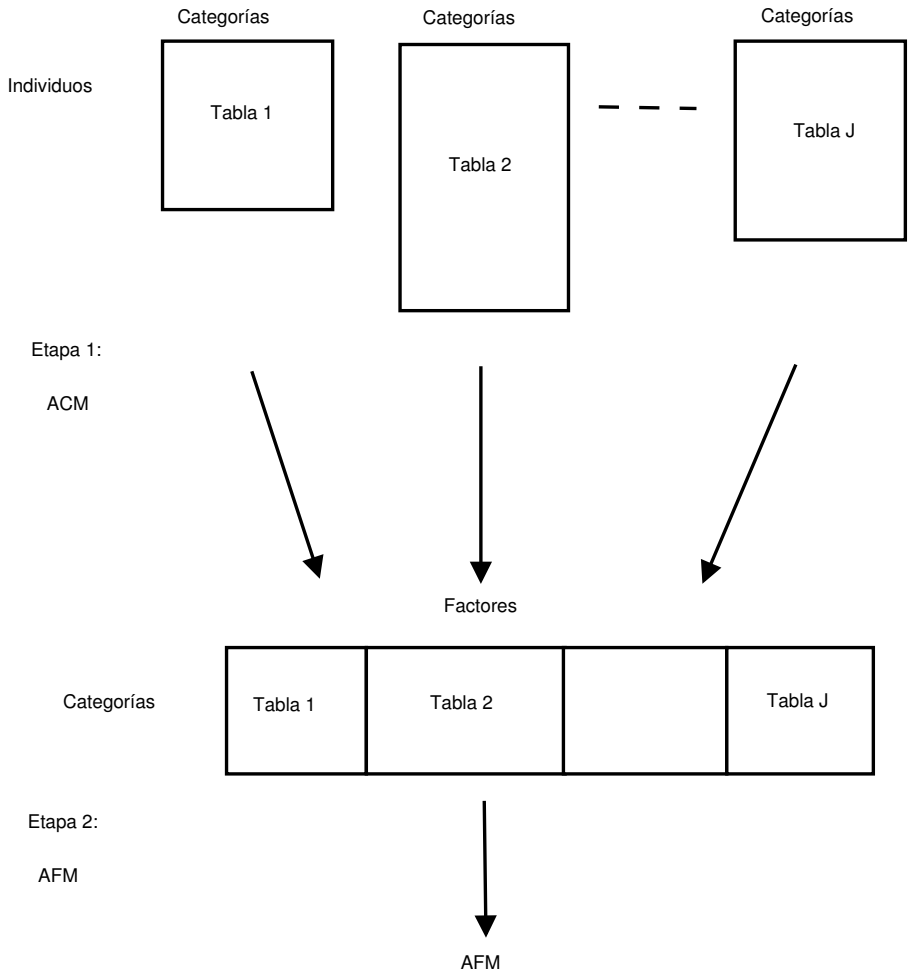


Figura 4.1: Esquema de la metodología propuesta para el análisis por AFM de tablas de efectivo diferente.

compone de dos etapas:

1. Se realiza un Análisis de Correspondencias Múltiples sobre cada una de las subtablas asociadas a los grupos. Las subtablas tienen por filas a los individuos encuestados en cada caso y por columnas las variables observadas, que han de ser las mismas y, en el caso habitual de que no sean continuas, categorizadas de idéntica manera. La dimensión en filas puede ser distinta, pero no así en columnas. A partir de cada ACM realizado, se extrae un número reducido de factores, en base a los objetivos habituales de los análisis factoriales, en un proceso de reducción de la dimensionalidad y de eliminación de los factores de escaso contenido común. No es formalmente necesario que el número de factores extraídos de cada tabla

sea el mismo, pero es conveniente que no sea muy dispar, para favorecer el uso de los elementos de análisis de la interestructura del análisis global posterior.

2. Se yuxtaponen las tablas formadas a partir de los factores correspondientes a las proyecciones de los puntos variable obtenidos en el paso anterior. Concretamente, se sustituyen las subtablas originales por dichos factores principales obtenidos de esas mismas subtablas. Con esta forma de proceder, se crea una gran tabla que contiene, por filas, las modalidades de todas las variables de las subtablas anteriores y que son comunes a todas ellas. Las columnas de esta gran tabla se conforman, en base a los mismos grupos definidos por las subtablas originales, mediante la yuxtaposición de los factores elegidos en el paso anterior para todas y cada una de las tablas. Esta tabla múltiple es de naturaleza cuantitativa y es analizada por medio de un AFM que permita obtener una visualización de la interestructura de estas tablas a partir de una representación razonablemente fiel de la estructura existente en las tablas originales.

El método descrito permite el análisis comparado mediante AFM de tablas de variables categóricas. Esta metodología plantea dos limitaciones:

1. Es necesario que las tablas recojan las mismas variables, y además deben de estar codificadas de manera homogénea.

2. En el segundo paso, al usar las coordenadas de las variables como puntos fila, se pierde la información específica sobre los individuos originales, de forma que desaparecen del análisis de la interestructura.

Con respecto a la primera limitación, más que una restricción en mi opinión supone una definición del ámbito de actuación, que no es despreciable en términos prácticos. La desaparición de la información sobre los individuos es, en principio, relevante. Sin embargo, es natural pensar que el interés real recae sobre las clases definidas a través de las categorías de las variables de las tablas originales, las cuales permanecen intactas en el análisis. De hecho, dado que los individuos son generalmente anónimos en el campo de las encuestas, no disponer de información directa sobre su representación en el análisis no parece que suponga un excesivo precio a pagar.

En la sección 4.4 se muestra una aplicación de la propuesta metodológica descrita para el análisis de una encuesta llevada a cabo en varios países de forma aproximadamente simultánea.

En la sección 4.3.1 se expone un pequeño ejercicio de simulación en el que se muestra el comportamiento esperable de algunos elementos del análisis sobre un conjunto de datos artificialmente construido.

4.3.1. Simulación

En esta sección se realiza una pequeña simulación que permite visualizar el comportamiento de algunos elementos del AFM aplicado a tablas de efectivo diferente presentado anteriormente.

Aspectos generales

Se generan $J = 2$ tablas de tamaño diferente, $N_1 = 50, N_2 = 36$ (de forma que $N_2 \approx 2N_1/3$) que contienen el mismo número de variables categóricas $K_1 = K_2 = 3$, todas ellas medidas en una escala de 5 puntos. Las variables a simular se generan de acuerdo con el esquema de la Figura 4.2 de la siguiente manera:

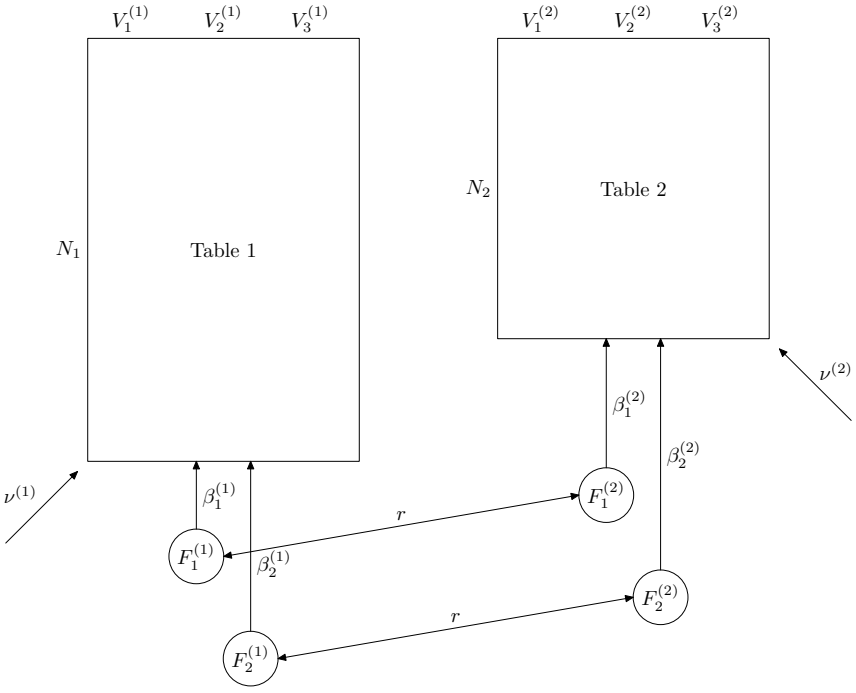


Figura 4.2: Esquema de formación de las tablas simuladas.

1. Se supone que existe una estructura subyacente común dentro de cada tabla de dimensionalidad $D = 2$. A partir de aquí, se generan $K_1 + K_2 + D \times J = 10$ muestras de tamaño $N_1 = \text{máx}\{N_1, N_2\} = 50$ de forma independiente como realizaciones de una variable aleatoria normal reducida. Se les denota $\nu_i^{(t)}, f_1^{(t)}, f_2^{(t)} t = 1, 2,\ i = 1, 2, 3$ y f_1, f_2.

2. Se impone una correlación r entre $f_l^{(1)}$ y $f_l^{(2)},\ l = 1, 2$.

3. Se construyen las variables continuas

$$z_i^{(t)} = \alpha^{(t)}\nu_i^{(t)} + \beta_{1i}^{(t)} f_1^{(t)} + \beta_{2i}^{(t)} f_2^{(t)} \quad t = 1, 2 \quad i = 1, 2, 3 \tag{4.14}$$

donde t indica el número de la tabla e i el de variable dentro de la tabla. Las variables $z_i^{(t)}$ son las variables continuas que están detrás de la formación de las variables categóricas que realmente aparecen en las tablas; se asemejan a una variable latente. f_1 y f_2 representan dos factores comunes, mientras que $\nu_i^{(t)}$ es un factor específico interpretado como *ruido*. Siguiendo la idea contenida en Markus (1994), se normalizan los coeficientes de la ecuación 4.14 con arreglo a la restricción

$$\left(\alpha^{(t)}\right)^2 + \sqrt{\left(\beta_{1i}^{(t)}\right)^2 + \left(\beta_{2i}^{(t)}\right)^2} = 1. \tag{4.15}$$

Los coeficientes $\beta_{ki}^{(t)}$ se eligen de forma que $\beta_{ki}^{(1)} \approx \beta_{ki}^{(2)}$ con la intención de que las tablas presenten un comportamiento similar.

4. Finalmente, las variables construidas $z_i^{(t)}$ se discretizan en 5 clases, tratando de preservar aproximadamente la igualdad entre los tamaños de las clases. Las variables se agrupan entonces en 2 tablas, borrándose aleatoriamente 14 líneas de la segunda de ellas para obtener el tamaño muestral deseado de $N_2 = 36$.

Una vez se han obtenido las variables generadas y las correspondientes tablas, se puede realizar el AFM para tablas de efectivo diferente y observar su comportamiento.

Aspectos particulares y resultados

La generación de las variables objeto de simulación necesita de la elección de los parámetros especificados en la ecuación (4.14) así como el coeficiente r que implica una determinada interestructura.

La tabla 4.1 contiene los parámetros elegidos para la simulación. La elección no es completamente libre, dado que tienen que satisfacer la restricción de la ecuación (4.15). Esta restricción es importante, puesto que implica que una variable claramente asociada con un factor subyacente $f_i^{(t)}$ no puede estarlo de forma significativa con otro factor, teóricamente ortogonal al primero. Dicho de otra forma, garantiza una fidelidad de representación elevada si se utilizan las variables y las tablas que se generan para un análisis de reducción de dimensionalidad como, por ejemplo, el ACP.

Table 1	Table 2
$\alpha^{(1)} = 0{,}01$	$\alpha^{(2)} = 0{,}03$
$\beta_{11}^{(1)} = 0{,}1$	$\beta_{11}^{(2)} = 0{,}2$
$\beta_{21}^{(1)} = 0{,}9946$	$\beta_{21}^{(2)} = 0{,}9789$
$\beta_{12}^{(1)} = 0{,}97$	$\beta_{12}^{(2)} = 0{,}95$
$\beta_{22}^{(1)} = 0{,}24$	$\beta_{22}^{(2)} = 0{,}3$
$\beta_{13}^{(1)} = 0{,}2$	$\beta_{13}^{(2)} = 0{,}1$
$\beta_{23}^{(1)} = 0{,}9797$	$\beta_{23}^{(2)} = 0{,}9941$

Tabla 4.1: Parámetros de la simulación.

Se han generado 100 replicaciones de todas las variables anteriormente descritas con los parámetros de la Tabla 4.1 y un coeficiente de correlación entre los factores subyacentes del mismo orden de las dos tablas igual a $r = 0{,}85$. Con esas variables se han generado 100 pares de tablas como las mencionadas anteriormente con las variables discretizadas como ya se ha comentado. La Figura 4.3 muestra el histograma de las 100 replicaciones de los coeficientes RV computados a partir de los 100 pares de tablas simuladas. Los valores centrales de la distribución se asocian claramente a la correlación impuesta de partida entre los ejes del mismo orden.

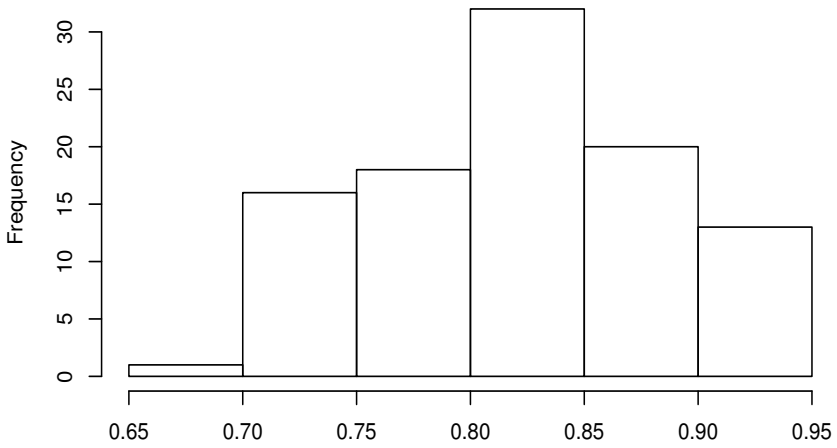
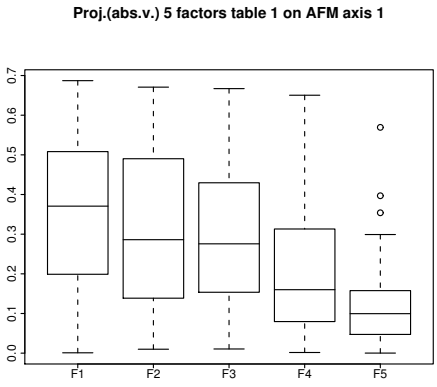


Figura 4.3: Histograma de los coeficientes RV entre las 2 tablas simuladas con $r = 0{,}85$. 100 replicaciones.

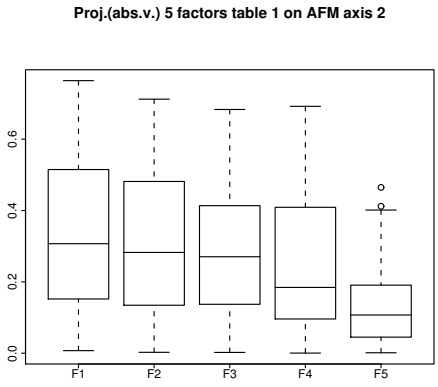
Se han realizado los AFM para estas tablas simuladas de efectivo diferente utilizando las 100 replicaciones. Es difícil realizar una representación gráfica

superpuesta de todas ellas, y más aún representar los factores parciales sobre los ejes del AFM, particularmente teniendo en cuenta que los signos de los ejes son arbitrarios, al igual que sucede, por ejemplo, en el bootstrap total (ver, p. ej., Lebart (2006) para una descripción del bootstrap en el contexto del ACP-ACM). En su lugar, se han obtenido las coordenadas de los factores parciales principales sobre los ejes principales del AFM y se han calculado diagramas de caja de factores parciales sobre cada eje global. Los resultados están en las Figuras 4.4(a)-4.4(d). En ellas puede verse cómo los factores parciales (F1-F5, eje horizontal) de mayor inercia están mejor representados en los primeros ejes globales (cada uno es una figura distinta). Parece que los ejes parciales están mejor representados sobre ejes globales del mismo orden, pero la relación es débil y no se cumple siempre, como es el caso de la Figura 4.4(b). Sí es cierto que a partir del factor parcial 5, la representación sobre los ejes globales es ya de muy escasa calidad. El hecho de que la dimensionalidad subyacente de las dos tablas sea de 2 y que aparezcan más de dos factores con buena calidad de representación en los dos primeros ejes globales puede tener que ver con que esa dimensionalidad tiende a *repartirse* entre las variables indicadoras objeto del ACM una vez que las tablas simuladas han sido discretizadas.

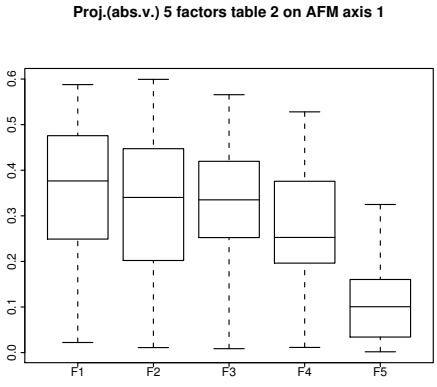
Estos resultados permiten corroborar la estabilidad de los resultados proporcionados por la extensión del AFM propuesto en este capítulo para el análisis simultáneo de tablas cualitativas de efectivo diferente y, por ende, su idoneidad como método exploratorio para este tipo de análisis de tablas múltiples.



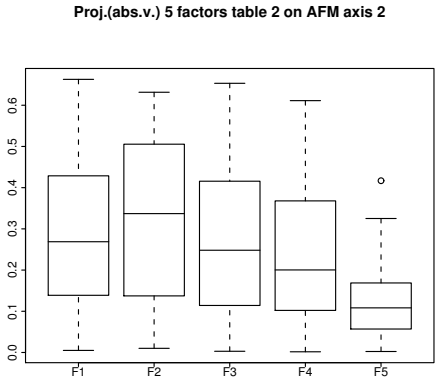
(a) Factores parciales de la Tabla 1 sobre el eje 1 del AFM



(b) Factores parciales de la Tabla 1 sobre el eje 2 del AFM



(c) Factores parciales de la Tabla 2 sobre el eje 1 del AFM



(d) Factores parciales de la Tabla 2 sobre el eje 2 del AFM

Figura 4.4: Diagramas de caja de las proyecciones de los 5 primeros factores parciales sobre los dos primeros ejes del AFM.

4.4. Aplicación: Encuesta de desigualdad social

Se ha llevado a cabo un análisis exploratorio con la Encuesta de Desigualdad Social proporcionada por el International Social Survey Programme (ISSP – Zentralarchiv für Empirische Sozialforschung, Universidad de Colonia) para el año 1992, originalmente realizada en 19 países. Las variables proporcionadas son las mismas para todos los países, siendo la mayoría cualitativas, medidas en una escala de Likert de 5 puntos. Sin embargo, el número de individuos entrevistados disponibles para cada país es bastante diferente, oscilando desde 749 hasta 2502.

La muestra seleccionada para el análisis es un subconjunto de 22 variables medidas para todos los individuos entrevistados en 10 países de los disponibles. Todas las variables contienen 5 categorías más una categoría de respuesta ausente que recogen opiniones del tipo:

- ¿Cuánto cree que un ... debería ganar?' o
- ¿Cuanto cree que un ... realmente gana?'

donde los puntos suspensivos son reemplazados en cada pregunta por una profesión de las expuestas en la Tabla 4.2, incluyendo profesiones correspondientes a distintos estatus sociales. Así, se incluyen profesiones de alto nivel (a veces denominadas de cuello blanco) como gerentes o jueces y otras relacionadas con oficios manuales o de menor cualificación (respectivamente denominadas de cuello azul) como granjeros, obreros o dependientes de tienda. Estas variables han sido recodificadas a categóricas preservando, en la medida de lo posible, la igualdad de tamaño de las clases en las que la codificación divide la muestra.

Profesiones de cuello azul	**Profesiones de cuello blanco**
Obrero cualificado	Médico
Dependiente de gandes almacenes	Directivo en gran empresa
Propietario de tienda pequeña	Abogado
Granjero	Gerente de fábrica
Obrero no cualificado	Juez
	Ministro del gobierno

Tabla 4.2: Profesiones correspondientes a las variables seleccionadas de la Encuesta de Desigualdad Social.

En la encuesta hay un gran número de variables socioeconómicas susceptibles de ser usadas como variables ilustrativas (esto es, no usables en la determinación de los factores, pero sí en la interpretación de los mismos al proyectarse sobre ellos). Se han seleccionado Religión, Sexo, Años de Educación y Tendencia Política (otras resultaron poco ó nada informativas). Estas variables ilustrativas sólo pueden ser de utilidad en la primera fase del análisis.

Los 10 países analizados finalmente son: la antigua República Federal de Alemania (WG), Gran Bretaña (GB), Estados Unidos (USA), Hungría (HUN), Noruega (NW), la antigua Checoslovaquia (CZ), la antigua URSS (RU), Nueva Zelanda (NZ), Filipinas (FIL) y España (SP).

Los ACM llevados a cabo sobre cada una de las 10 tablas correspondientes a los 10 países, incluyendo cada una las 22 variables seleccionadas, proporcionan resultados que concluyen que los factores tienen una interpretación similar en todas ellas. El primer factor refleja la no respuesta, típico de este tipo de encuestas, asociado a las personas de edad avanzada. El segundo factor es, básicamente, un factor tamaño que ordena las respuestas a las preguntas en orden ascendente y que tiene alguna relación con Edad, en el sentido de que los jóvenes tienden a presentar los valores más bajos. El tercer factor muestra un efecto Guttman cuando se dibuja frente al segundo: las categorías extremas se oponen a las medias. El cuarto factor es menos interpretable: muestra relaciones entre opiniones que asocian que ciertas profesiones de cuello blanco deberían y son bien pagadas junto a opiniones que consideran que otras de cuello azul son, y deberían de ser, mal remuneradas. Las restantes variables socioeconómicas (aparte de Sexo y Edad), que entran en el análisis como ilustrativas, no son de mucha ayuda y en algunos países los factores 3 y 4 aparecen intercambiados.

Decidimos entonces mantener 5 factores de cada ACM como representativos de cada una de las 10 tablas, aunque sólo hemos encontrado relevancia e interpretabilidad en los 4 primeros. De esta forma, la tabla formada por las coordenadas de las categorías en los factores es una tabla de $10\times5=50$ columnas y 132 filas ((5 categorías + no respuesta)$\times$22 variables =132 categorías totales). Esta es la tabla que se analiza mediante AFM en la segunda etapa del procedimiento propuesto en la sección 4.3.

La matriz de correlación entre los factores retenidos de los 10 ACMs (antes de efectuar la ponderación citada) se muestra en las Tablas 4.3(a)-4.3(c). Puede observarse que los factores del mismo orden están altamente correlacionados, además de tener la misma interpretación para sus respectivas tablas. También se vislumbra que la correlación entre los factores del mismo orden para los distintos países decae cuando se consideran factores de orden superior. Interpretamos esto como una consecuencia del hecho de que los factores

de orden superior tienen mayor influencia de fluctuaciones muestrales y cubren cada vez menos características comunes de las variables originales.

Para obtener una medida de similaridad entre tablas, calculamos los coeficientes RV de Escofier (ver ecuación 4.9) entre las tablas, que pueden interpretarse como coeficientes de correlación entre pares de tablas. Dados los resultados de la sección 4.3.1, estos coeficientes son un indicador de la estructura subyacente común entre las tablas objeto de análisis. La tabla que contiene dichos coeficientes, similar a una matriz de correlación, aparece en la Tabla 4.4.

Así, por ejemplo, Estados Unidos y Alemania presentan una alta correlación (0,75), mientras la antigua URSS y Nueva Zelanda por el contrario no (0,42). Estos coeficientes tienen dos dificultades prácticas. La primera es la relativa a la interpretación, que se complica cuando aumenta el número de tablas a considerar. La segunda concierne a los coeficientes RV, que sólo tienen en cuenta correlaciones dos a dos, sin tener en consideración relaciones netas de influencia de otras variables o tablas. Esto sucede de la misma manera que un coeficiente de correlación entre dos variables en un contexto multivariante puede revelar información sesgada, en comparación con la que puede ofrecer un coeficiente de correlación parcial, por ejemplo. En este sentido cobra importancia la utilización de un AFM, que además de tener en cuenta todas las variables a la vez, equilibra la influencia de los grupos que se forman con ellas.

La Figura 4.5 refleja el diagrama de dispersión que contiene los coeficientes de ligazón correspondientes a las 10 tablas consideradas en el AFM. En ella se ve cómo la ligazón con los ejes globales es razonablemente elevada, siendo siempre superior a 0,50 para todos los países. Supera el valor 0,75 para 5 países (USA, FIL, SP, NW y NZ) con respecto al primer eje global y superior a 0.70 para 6 de ellos (USA, FIL, HUN, RU, GB, WG) con respecto al segundo. En este sentido, USA y FIL son los países mejor representados mediante el análisis múltiple, al menos en el eje principal.

Por su parte, la Figura 4.6 refleja las proyecciones de los puntos categoría de las tablas de la primera etapa del análisis sobre los ejes globales del AFM realizado en la segunda etapa. Formalmente es la proyección de la nube de individuos sobre los ejes globales.

Las etiquetas se interpretan de la siguiente forma: el texto es el indicativo del país y el dígito es el orden del factor que corresponde a los factores extraídos de la primera etapa, utilizados como individuos activos en la segunda fase. Así, por ejemplo, `hun5` representa a la coordenada del factor inicial de orden 5 para Hungría.

La Figura 4.6 muestra que las proyecciones de los factores del mismo orden están muy próximas y en una dirección muy similar, lo que refuerza las simila-

	WG1	WG2	WG3	WG4	WG5	GB1	GB2	GB3	GB4	GB5	USA1	USA2	USA3	USA4	USA5	HUN1	HUN2
WG1	1,00																
WG2	-0,07	1,00															
WG3	0,10	0,20	1,00														
WG4	0,14	0,06	0,15	1,00													
WG5	0,13	0,27	0,02	0,43	1,00												
GB1	-1,00	0,07	-0,09	-0,13	-0,12	1,00											
GB2	-0,05	0,92	0,22	0,00	0,25	0,05	1,00										
GB3	0,17	0,20	0,85	0,37	0,26	-0,16	0,26	1,00									
GB4	-0,06	-0,02	-0,04	-0,83	-0,23	0,05	0,01	-0,28	1,00								
GB5	-0,09	-0,48	-0,24	-0,26	-0,61	0,08	-0,57	-0,43	0,18	1,00							
USA1	0,99	-0,09	0,08	0,12	0,11	-1,00	-0,07	0,15	-0,05	-0,06	1,00						
USA2	0,00	0,90	0,20	0,19	0,19	0,00	0,84	0,17	-0,13	-0,30	-0,03	1,00					
USA3	0,18	-0,03	0,89	0,08	-0,04	-0,18	0,02	0,75	0,07	-0,10	0,16	0,04	1,00				
USA4	0,05	-0,12	0,23	0,76	0,07	-0,04	-0,22	0,27	-0,69	0,12	0,03	0,08	0,21	1,00			
USA5	-0,03	-0,40	-0,18	-0,16	-0,74	0,01	-0,34	-0,29	0,02	0,59	0,02	-0,29	-0,14	0,05	1,00		
HUN1	0,97	-0,15	0,08	0,13	0,10	-0,96	-0,15	0,14	-0,06	-0,03	0,95	-0,09	0,17	0,09	-0,01	1,00	
HUN2	0,08	0,79	0,03	0,01	0,24	-0,08	0,85	0,06	0,04	-0,37	0,06	0,81	-0,07	-0,23	-0,29	-0,04	1,00
HUN3	0,19	0,03	0,69	0,26	-0,05	-0,19	0,10	0,70	-0,25	-0,07	0,17	0,15	0,70	0,32	-0,05	0,15	0,08
HUN4	0,00	-0,26	-0,15	-0,40	-0,50	-0,01	-0,23	-0,27	0,36	0,40	0,03	-0,31	-0,12	-0,28	0,49	0,03	-0,21
HUN5	-0,09	-0,36	-0,30	0,12	-0,45	0,09	-0,44	-0,37	-0,25	0,53	-0,07	-0,29	-0,30	0,32	0,58	-0,01	-0,47
NW1	-1,00	0,07	-0,10	-0,12	-0,11	1,00	0,04	-0,17	0,04	0,08	-1,00	0,00	-0,18	-0,03	0,01	-0,95	-0,09
NW2	-0,15	0,90	0,19	-0,08	0,17	0,14	0,92	0,17	0,11	-0,41	-0,16	0,85	0,01	-0,24	-0,29	-0,24	0,83
NW3	0,14	0,04	0,75	0,30	0,16	-0,13	0,03	0,78	-0,27	-0,26	0,12	0,07	0,72	0,33	-0,25	0,14	-0,10
NW4	0,01	-0,32	0,30	-0,50	-0,61	-0,01	-0,26	0,08	0,40	0,40	0,01	-0,32	0,37	-0,25	0,42	0,04	-0,28
NW5	0,15	-0,16	0,04	0,08	-0,17	-0,19	-0,16	0,08	-0,16	0,18	0,22	-0,10	0,06	0,08	0,32	0,05	-0,23
CZ1	-0,96	0,03	-0,09	-0,11	-0,12	0,95	0,01	-0,15	0,04	0,09	-0,94	-0,03	-0,17	-0,04	0,06	-0,96	-0,12
CZ2	-0,07	0,78	0,03	-0,16	0,11	0,07	0,83	0,00	0,15	-0,29	-0,08	0,77	-0,10	-0,29	-0,18	-0,18	0,88
CZ3	0,09	0,17	0,59	0,26	0,01	-0,09	0,25	0,63	-0,26	-0,09	0,09	0,27	0,50	0,32	0,02	0,05	0,17
CZ4	0,03	0,00	-0,17	0,43	0,14	-0,05	-0,03	-0,09	-0,47	0,05	0,07	0,12	-0,21	0,36	0,17	-0,05	-0,01
CZ5	-0,07	-0,18	-0,17	-0,03	-0,33	0,10	-0,21	-0,25	-0,01	0,29	-0,12	-0,16	-0,15	0,13	0,21	0,07	-0,16
RU1	0,99	-0,10	0,08	0,12	0,10	-0,99	-0,09	0,14	-0,05	-0,04	0,98	-0,03	0,16	0,06	0,00	0,98	0,04
RU2	0,02	0,72	0,13	-0,15	0,16	-0,02	0,80	0,07	0,19	-0,34	0,01	0,73	0,04	-0,27	-0,25	-0,09	0,88
RU3	0,11	0,17	0,46	0,33	0,14	-0,10	0,23	0,57	-0,42	-0,19	0,09	0,26	0,39	0,35	-0,22	0,07	0,18
RU4	0,09	0,38	0,28	0,04	0,40	-0,08	0,43	0,39	-0,04	-0,49	0,06	0,34	0,25	-0,06	-0,54	0,03	0,39
RU5	0,05	-0,08	-0,13	0,30	0,32	-0,05	-0,11	-0,07	-0,21	-0,09	0,05	-0,01	-0,10	0,23	-0,12	0,05	0,05
NZ1	0,99	-0,07	0,10	0,12	0,11	-1,00	-0,04	0,17	-0,05	-0,08	1,00	0,00	0,19	0,03	-0,01	0,95	0,08
NZ2	-0,12	0,88	-0,06	-0,10	0,20	0,11	0,91	-0,04	0,09	-0,37	-0,13	0,87	-0,21	-0,26	-0,29	-0,21	0,85
NZ3	0,08	0,05	0,89	0,06	-0,02	-0,07	0,11	0,82	0,01	-0,23	0,06	0,07	0,89	0,17	-0,18	0,07	-0,02
NZ4	0,08	-0,11	0,20	0,64	-0,05	-0,09	-0,14	0,29	-0,70	0,16	0,09	0,11	0,20	0,72	0,24	0,04	-0,17
NZ5	-0,14	-0,09	0,08	0,15	-0,18	0,17	-0,06	0,04	-0,18	0,16	-0,20	0,00	0,10	0,27	0,07	-0,01	-0,01
FIL1	-0,97	0,15	-0,07	-0,10	-0,05	0,97	0,15	-0,11	0,05	-0,01	-0,97	0,08	-0,17	-0,10	-0,02	-0,97	0,03
FIL2	0,06	0,86	-0,01	-0,16	0,18	-0,07	0,84	-0,03	0,17	-0,33	0,05	0,81	-0,14	-0,28	-0,31	-0,03	0,82
FIL3	-0,03	-0,24	-0,61	-0,46	0,11	0,02	-0,27	-0,60	0,46	0,02	-0,01	-0,40	-0,49	-0,54	-0,07	0,01	-0,21
FIL4	-0,13	-0,05	-0,50	0,21	-0,27	0,13	-0,14	-0,45	-0,36	0,37	-0,12	0,05	-0,52	0,32	0,39	-0,11	-0,05
FIL5	0,11	-0,12	-0,07	-0,38	-0,52	-0,12	-0,02	-0,16	0,36	0,32	0,13	-0,08	-0,01	-0,34	0,48	0,10	0,05
SP1	0,92	-0,19	0,06	0,07	0,05	-0,92	-0,23	0,07	-0,02	0,06	0,91	-0,14	0,14	0,12	0,02	0,95	-0,13
SP2	0,00	-0,84	-0,05	0,06	-0,17	0,00	-0,89	-0,08	-0,08	0,35	0,02	-0,81	0,10	0,30	0,24	0,11	-0,89
SP3	0,23	-0,10	0,76	0,16	0,01	-0,22	-0,05	0,78	-0,05	-0,07	0,21	-0,05	0,82	0,23	-0,09	0,20	-0,09
SP4	0,07	0,25	0,21	0,28	0,08	-0,06	0,14	0,30	-0,36	-0,01	0,05	0,28	0,12	0,30	-0,17	0,05	0,17
SP5	0,00	0,33	-0,18	-0,27	0,40	0,00	0,35	-0,15	0,39	-0,44	-0,01	0,20	-0,20	-0,51	-0,50	-0,05	0,39
	WG1	WG2	WG3	WG4	WG5	GB1	GB2	GB3	GB4	GB5	USA1	USA2	USA3	USA4	USA5	HUN1	HUN2

Tabla 4.3(a): Matriz de correlaciones entre los 5 primeros factores ACM de 10 países: factores 1-17.

	HUN3	HUN4	HUN5	NW1	NW2	NW3	NW4	NW5	CZ1	CZ2	CZ3	CZ4	CZ5	RU1	RU2	RU3	RU4
HUN3	1,00																
HUN4	-0,25	1,00															
HUN5	-0,25	0,45	1,00														
NW1	-0,19	-0,01	0,09	1,00													
NW2	0,11	-0,21	-0,42	0,14	1,00												
NW3	0,67	-0,35	-0,24	-0,13	-0,04	1,00											
NW4	0,26	0,51	0,12	-0,01	-0,17	0,09	1,00										
NW5	0,04	0,13	0,21	-0,21	-0,17	0,04	0,05	1,00									
CZ1	-0,18	0,00	0,09	0,94	0,11	-0,13	-0,02	-0,03	1,00								
CZ2	-0,02	-0,08	-0,36	0,06	0,80	-0,11	-0,19	-0,20	0,04	1,00							
CZ3	0,72	-0,18	-0,14	-0,10	0,23	0,51	0,07	0,06	-0,06	0,19	1,00						
CZ4	-0,08	-0,26	0,26	-0,06	-0,03	-0,11	-0,44	0,56	0,05	-0,08	-0,01	1,00					
CZ5	-0,10	0,36	0,45	0,13	-0,21	-0,20	0,25	-0,57	-0,02	-0,16	-0,15	-0,39	1,00				
RU1	0,16	0,04	-0,04	-0,98	-0,18	0,12	0,03	0,12	-0,97	-0,11	0,06	0,00	-0,02	1,00			
RU2	0,02	-0,13	-0,41	-0,03	0,76	-0,07	-0,11	-0,20	-0,05	0,86	0,11	-0,08	-0,17	-0,02	1,00		
RU3	0,72	-0,44	-0,22	-0,10	0,18	0,46	-0,09	0,01	-0,11	0,11	0,71	0,14	-0,10	0,07	0,14	1,00	
RU4	0,28	-0,55	-0,71	-0,08	0,39	0,31	-0,19	-0,12	-0,10	0,32	0,26	-0,10	-0,41	0,04	0,32	0,44	1,00
RU5	-0,08	-0,46	-0,13	-0,05	-0,09	-0,07	-0,43	-0,05	-0,04	-0,11	-0,10	0,33	-0,23	0,05	-0,05	0,06	0,15
NZ1	0,19	0,02	-0,09	-1,00	-0,14	0,14	0,01	0,22	-0,94	-0,06	0,10	0,07	-0,13	0,98	0,03	0,10	0,08
NZ2	-0,09	-0,22	-0,37	0,11	0,92	-0,20	-0,31	-0,18	0,08	0,84	0,08	0,02	-0,17	-0,15	0,78	0,09	0,36
NZ3	0,71	-0,21	-0,36	-0,08	0,10	0,78	0,36	-0,01	-0,08	-0,04	0,54	-0,24	-0,15	0,06	0,09	0,48	0,32
NZ4	0,32	-0,25	0,34	-0,09	-0,18	0,33	-0,22	0,52	-0,01	-0,20	0,37	0,61	-0,21	0,06	-0,23	0,35	-0,07
NZ5	0,21	-0,02	0,14	0,20	-0,06	0,10	0,20	-0,74	0,03	-0,02	0,13	-0,44	0,74	-0,09	-0,05	0,14	-0,06
FIL1	-0,15	-0,04	0,00	0,96	0,24	-0,14	-0,05	-0,15	0,95	0,16	-0,05	-0,03	0,01	-0,98	0,07	-0,07	-0,04
FIL2	-0,11	-0,16	-0,38	-0,07	0,82	-0,13	-0,23	-0,17	-0,10	0,84	-0,01	-0,05	-0,15	0,03	0,82	0,04	0,39
FIL3	-0,64	0,22	-0,02	0,03	-0,25	-0,46	0,01	-0,09	0,02	-0,15	-0,69	-0,19	0,02	0,00	-0,16	-0,63	-0,13
FIL4	-0,35	0,11	0,54	0,13	-0,14	-0,40	-0,16	0,09	0,15	-0,02	-0,18	0,36	0,29	-0,11	-0,10	-0,14	-0,32
FIL5	0,01	0,56	0,15	-0,12	-0,01	-0,28	0,47	0,10	-0,17	0,07	-0,07	-0,14	0,22	0,13	0,04	-0,20	-0,33
SP1	0,11	0,01	0,05	-0,91	-0,29	0,12	0,06	0,11	-0,90	-0,25	0,00	0,02	0,02	0,94	-0,16	0,03	0,00
SP2	-0,06	0,10	0,40	0,01	-0,88	0,11	0,20	0,13	0,03	-0,85	-0,18	0,05	0,18	0,04	-0,76	-0,13	-0,36
SP3	0,74	-0,08	-0,34	-0,22	-0,05	0,73	0,42	0,10	-0,21	-0,16	0,56	-0,22	-0,16	0,20	-0,07	0,41	0,26
SP4	0,25	-0,37	-0,08	-0,06	0,16	0,32	-0,17	0,08	-0,05	0,13	0,33	0,06	-0,16	0,06	0,07	0,33	0,26
SP5	-0,24	-0,18	-0,53	0,00	0,38	-0,23	-0,31	-0,23	-0,02	0,29	-0,25	-0,16	-0,24	-0,02	0,34	-0,13	0,29
	HUN3	HUN4	HUN5	NW1	NW2	NW3	NW4	NW5	CZ1	CZ2	CZ3	CZ4	CZ5	RU1	RU2	RU3	RU4

Tabla 4.3(b): Matriz de correlaciones entre los 5 primeros factores ACM de 10 países: factores 18-34.

	RU5	NZ1	NZ2	NZ3	NZ4	NZ5	FIL1	FIL2	FIL3	FIL4	FIL5	SP1	SP2	SP3	SP4	SP5
RU5	1,00															
NZ1	0,04	1,00														
NZ2	-0,03	-0,11	1,00													
NZ3	-0,11	0,08	-0,14	1,00												
NZ4	0,12	0,10	-0,19	0,14	1,00											
NZ5	-0,05	-0,20	-0,07	0,14	-0,05	1,00										
FIL1	-0,04	-0,96	0,20	-0,06	-0,09	0,10	1,00									
FIL2	-0,08	0,07	0,89	-0,10	-0,25	-0,11	0,01	1,00								
FIL3	-0,03	-0,03	-0,11	-0,54	-0,56	-0,25	0,01	-0,03	1,00							
FIL4	0,05	-0,13	0,02	-0,52	0,27	0,15	0,08	0,00	-0,04	1,00						
FIL5	-0,31	0,12	-0,04	-0,07	-0,19	0,05	-0,10	-0,04	0,01	0,05	1,00					
SP1	0,08	0,91	-0,25	0,04	0,07	-0,09	-0,96	-0,06	0,04	-0,07	0,02	1,00				
SP2	0,08	0,00	-0,89	0,05	0,15	0,05	-0,12	-0,81	0,16	0,07	-0,11	0,22	1,00			
SP3	-0,15	0,23	-0,28	0,83	0,22	0,11	-0,20	-0,19	-0,45	-0,46	0,02	0,16	0,11	1,00		
SP4	0,03	0,06	0,12	0,22	0,35	0,05	-0,08	0,11	-0,37	0,12	-0,28	0,11	-0,14	0,26	1,00	
SP5	0,22	0,00	0,42	-0,15	-0,57	-0,29	0,06	0,36	0,33	-0,26	-0,03	-0,06	-0,34	-0,25	0,00	1,00
	RU5	NZ1	NZ2	NZ3	NZ4	NZ5	FIL1	FIL2	FIL3	FIL4	FIL5	SP1	SP2	SP3	SP4	SP5

Tabla 4.3(c): Matriz de correlaciones entre los 5 primeros factores ACM de 10 países: factores 35-50.

	WG	GB	USA	HUN	NW	CZ	RU	NZ	FIL	SP
WG	1,00									
GB	0,73	1,00								
USA	0,75	0,61	1,00							
HUN	0,54	0,55	0,59	1,00						
NW	0,64	0,58	0,61	0,54	1,00					
CZ	0,44	0,45	0,45	0,52	0,55	1,00				
RU	0,45	0,50	0,48	0,68	0,44	0,51	1,00			
NZ	0,58	0,59	0,68	0,51	0,75	0,63	0,42	1,00		
FIL	0,61	0,57	0,66	0,55	0,51	0,49	0,49	0,58	1,00	
SP	0,50	0,53	0,59	0,57	0,56	0,45	0,42	0,61	0,50	1,00

Tabla 4.4: Coeficientes RV entre pares de tablas para 10 países seleccionados.

ridades entre los países. Los factores de orden 1 de los países están fuertemente relacionados entre sí y bien representados en el primer eje global. El hecho de que aparezcan en lados opuestos del eje carece de relevancia, puesto que los signos de los factores son arbitrarios. La situación en el segundo eje global es muy similar con respecto a los factores iniciales de orden 2. Únicamente se aprecia una ligera rotación de los ejes hacia la izquierda, de lo que sería una representación, digamos, *perfecta* en cuanto a la asociación de factores iniciales con ejes globales.

Se han proyectado también algunos de los puntos correspondientes a los ejes iniciales de los países de orden igual o superior a 3 (y hasta orden 5). Se aprecia cómo, en líneas generales, muestran una alta relación entre ejes del mismo orden, al igual que los primeros factores iniciales, y una apreciable representación sobre los ejes principales, aunque claramente inferior a la de los primeros ejes. En este sentido, la representación es sensiblemente peor que en caso de la sección 4.3.1, donde un mayor número de factores (hasta el 4) estaban bastante bien representados en el plano principal, posiblemente por tratarse de conjuntos de datos de dimensionalidad menor.

Dada la elevada relación entre los factores iniciales y su representación en el AFM global, esto significa que la interpretación de los ejes del AFM global se debería de corresponder con la de los factores iniciales. Esto quiere decir que el primer eje global se asocia a la modalidad de no respuesta y el segundo a un efecto talla. Los ejes siguientes se interpretan de la misma forma en tanto en cuanto las proyecciones de los factores iniciales estén bien representadas sobre tales ejes. El AFM pone de manifiesto con claridad cuán comunes son los factores subyacentes en las tablas y la proximidad, medida a través de

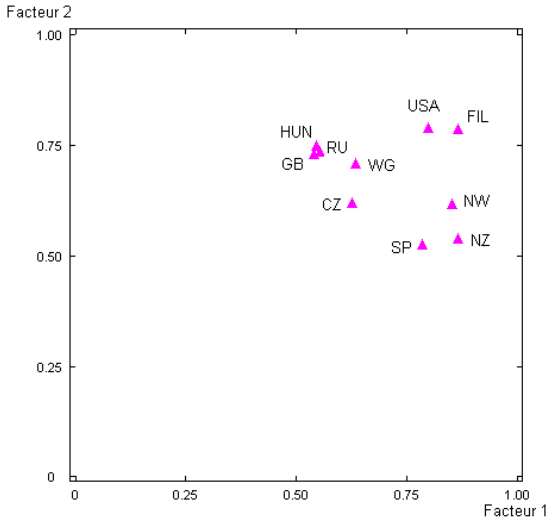


Figura 4.5: Coeficientes de ligazón $\mathcal{L}_g$ entre las 10 tablas y los factores globales.

coeficientes de ligazón, entre ellas, que en el caso de los conjuntos de datos empleados es muy elevada.

4.5. Conclusiones

En este capítulo se ha presentado una de las técnicas de análisis exploratorio de tablas múltiples más versátiles, el Análisis Factorial Múltiple (AFM). El AFM con el tiempo y uso se ha consagrado como una filosofía de análisis comparativo, tanto gráfico como a través de indicadores numéricos, de diferentes conjuntos de datos, sea cual sea su naturaleza y su estructura. Así lo ponen de manifiesto los numerosos trabajos que han visto la luz desde que en 1986 Escofier B. y Pagés J., en el seno de la Escuela Francesa de Análisis de Datos, pusieran en conocimiento de la comunidad científica una nueva técnica de análisis de datos que permitía la estructura en distintos grupos de los datos, equilibrando la influencia de cada uno de ellos.

El núcleo de este capítulo, sin embargo, lo constituye una nueva aportación que se enmarca en el contexto de adaptaciones, extensiones y aplicaciones del AFM. Un nuevo procedimiento, combinación de técnicas factoriales de tabla única (ACM) y de tabla múltiple (AFM), que permite el tratamiento simultáneo, desde un punto de vista descriptivo y comparativo, de grupos de individuos en los que se ha medido la misma información mediante variables nominales. Por tanto, es un método que hereda las importantes bondades que,

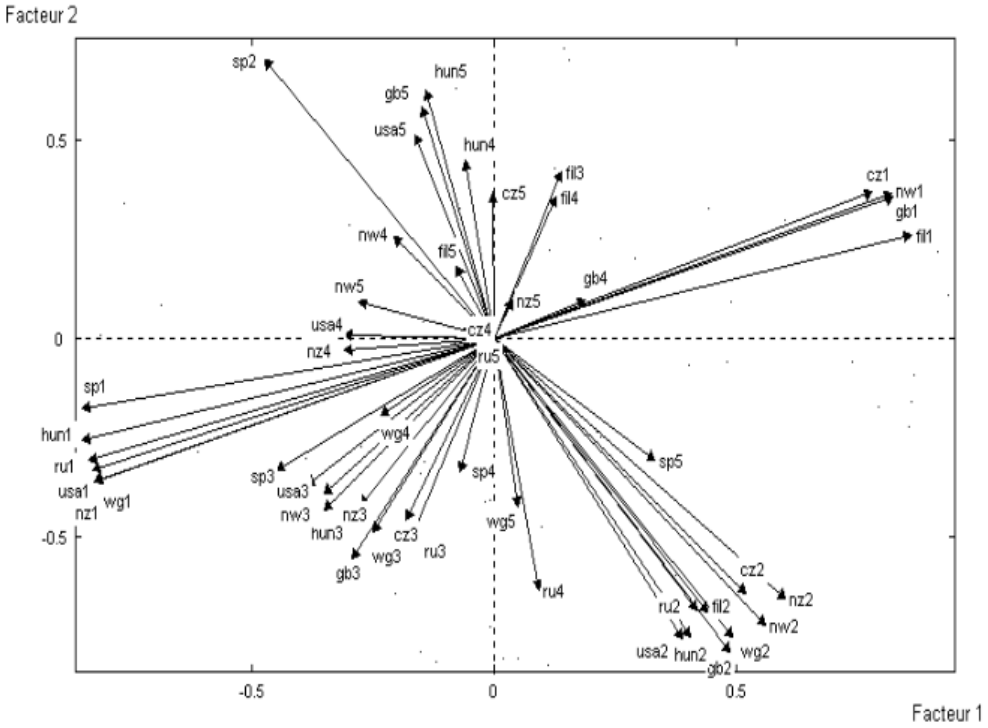


Figura 4.6: Proyecciones de los 5 primeros ejes parciales sobre el plano principal del AFM para las 10 tablas.

en el campo del análisis de datos, las dos metodologías en las que se basa poseen. Merecen una mención especial las siguientes:

1. La naturaleza exploratoria de la técnica tiene la ventaja de dejar que sean los propios datos los que sugieran cuales son los factores más relevantes para cada grupo de individuos, cuál es su importancia relativa de cara a un análisis global y cómo de próximos están unos de otros (en definitiva, unas tablas de otras).

2. La metodología proporciona una visión global de los datos, sin que ninguno de los grupos considerados tenga protagonismo ni quede eclipsado por el resto. Además, proporciona unos resultados con gran riqueza interpretativa, entre los que destacan las distintas medidas de relación entre los grupos analizados. Estas medidas ayudan a tomar decisiones sobre los grupos a mantener en un análisis global, por su similitud, o a analizar por separado dado su comportamiento específico y distinto al resto.

3. Los individuos originales pierden su identidad. Sin embargo, es bien sabido que, habitualmente, en las encuestas los individuos de la muestra no tienen interés en sí mismos, siendo verdaderamente importantes las clases definidas a través de ellos en su diferente caracterización. De hecho, las categorías, que constituyen los nuevos individuos en la tabla analizada en la segunda etapa de la metodología propuesta, definen estas clases.

La aplicación incluida muestra como, con un conjunto de datos real, es posible extraer la información más relevante de la interestructura de un análisis de tablas múltiples como el AFM. El ejercicio de simulación, por su parte, muestra cómo los elementos básicos de interpretación de la interestructura utilizados se corresponden con lo que tendrían que medir, como es la correlación subyacente entre dos tablas de datos. En este sentido, este ejercicio nos proporciona confianza sobre los resultados que se obtienen en la aplicación presentada, y en otras que puedan presentarse.

En este capítulo también se ha querido dejar constancia de la existencia de otras metodologías que, con un enfoque distinto, más ambicioso que las recogidas en el primero, resultan igualmente idóneas para el proceso de descubrimiento del conocimiento implícito en grandes volúmenes de datos. Y, al igual que las allí presentadas, estas técnicas de análisis de tablas múltiples también pueden hacer grandes aportaciones en las etapas que conllevan la transformación en conocimiento explícito. No en vano, estas metodologías completan los patrones de comportamiento al permitir detectar trayectorias asociadas a las diferentes tablas en las que se estructuran los datos analizados, información eclipsada cuando el estudio se realiza desde la óptica de tabla única.

CAPÍTULO 5

Tablas múltiples en el análisis textual

5.1. Introducción

En este capítulo se aborda el problema de las tablas múltiples mixtas incluyendo tablas lexicales provenientes de textos relacionados pero escritos en dos o más lenguas diferentes. Las tablas múltiples consideradas se denominan mixtas puesto que pueden contener tres tipos de elementos de diferente naturaleza: valores que se obtienen de medir variables cuantitativas sobre un conjunto de individuos, categorías definidas a partir de variables categóricas medidas o seleccionadas por el mismo conjunto de individuos y, finalmente, frecuencias correspondientes a unidades léxicas (p. ej., palabras) empleadas en un corpus textual como pueden ser las respuestas a una pregunta abierta. La particularidad, en este caso, es la posibilidad existente para los individuos de responder a dicha pregunta abierta en idiomas diferentes, dando lugar a corpus textuales distintos.

5.1.1. El análisis textual

El análisis textual es una parte de la Estadística Exploratoria asociada a la Lingüística y cuyo objeto básico es el análisis de textos en cuanto al estudio de las frecuencias obtenidas a partir de las palabras empleadas en la redacción de los mismos, generalmente utilizando herramientas de la Estadística Multivariante. Tiene aspectos importantes en común con las técnicas de *data mining* en el sentido de que tratan de extraer patrones (información) de grandes masas de datos, que en este caso lo conforman las palabras contenidas en textos, que

pueden ser grandes como libros o colecciones de los mismos.

El análisis de frecuencias inherente al análisis textual implica el uso de variables discretas y/o cualitativas. El vocabulario que aparece en cualquier texto, incluso si éste no es muy extenso, generalmente resulta en una tabla de frecuencias de alta dimensionalidad donde aparecen muchas palabras diferentes y donde las frecuencias son, sin embargo, generalmente pequeñas. Este tipo de tablas se denominan *dispersas*. La elevada dimensionalidad favorece el uso de técnicas de análisis multidimensionales adecuadas a este tipo de datos (Benzécri 1973, 1977, Lebart 1982, Lebart et al. 1998), como el análisis de correspondencias y el análisis de clasificación.

El origen del análisis textual está en la Estadística Léxica (Yule 1944, Guiraud 1954, 1960), cuyo principal objetivo era la descripción del vocabulario léxico de autores clásicos. La motivación primaria de la descripción era el de la asignación y/ó verificación de la autoría de textos en los casos en la que ésta fuese dudosa debido a la falta de información sobre la misma dado el transcurrir del tiempo.

Etapas del análisis textual

La obtención de las frecuencias, base del análisis textual, requiere un proceso complejo de conversión de los elementos textuales pertenecientes a un texto a una tabla resumen de los mismos. Dicho proceso se realiza en una serie de etapas que se enumeran a continuación.

Normalización La existencia de sinónimos provoca que, en cualquier texto, dos palabras con el mismo significado puedan escribirse de diferente manera. Esto puede considerarse como una característica de riqueza lingüística, pero disminuye de manera ficticia los valores de las frecuencias de los términos empleados. En este caso, tales palabras pueden agruparse como una sola, aspecto que es necesariamente subjetivo por parte del analista y que debe realizarse a la luz del objeto del estudio y de las características del texto.

Segmentación Esta fase consiste en la partición del texto en sus unidades elementales o unidades léxicas, es decir, aquéllas que ya no se dividirán más y sobre las cuales se contarán frecuencias. Dichas unidades elementales no son necesariamente sólo palabras, sino que también aparecen algunas combinaciones de palabras. Estas unidades pueden ser de varios tipos:

- **Forma gráfica o palabra** Una secuencia de letras delimitada a izquierda y derecha por un espacio en blanco o un signo de puntuación, denominados delimitadores.

- **Lema** Consiste en la agrupación en torno a una única forma gráfica de varias palabras con la misma raíz o, mejor, que corresponden a una misma entrada del diccionario. Principalmente esto supone la conversión de todos los tiempos verbales a infinitivo, de todos los sustantivos a singular y de todos los adjetivos a un mismo género (masculino o femenino) singular.
- **Unidades léxicas complejas** Son unidades compuestas de dos o más palabras y que como tal tienen significado propio, diferente del que tienen las palabras que las componen por separado. El ejemplo clásico es *seguridad social*. A su vez, estas unidades pueden dividirse en dos:
  - **Segmentos repetidos** Se refiere a unidades complejas compuestas por formas gráficas contiguas.
  - **Cuasisegmentos repetidos** Unidades complejas que contienen formas no necesariamente contiguas (como *ir al cine*, que puede aparecer como *ir nada/un poco/mucho/siempre al cine*). Pueden existir unidades complejas compuestas por formas que se solapen, y, en particular, generalmente lo hacen con las formas gráficas de las que están compuestas.

Tanto en el caso de usar como unidades las formas gráficas como los lemas, es necesario resolver posibles problemas de ambigüedad. Una serie de caracteres alfanuméricos terminados a ambos lados por un delimitador se denomina *ocurrencia*. Dos series así definidas que sean idénticas forman dos ocurrencias de la misma forma gráfica. Entonces, es posible que dos ocurrencias de la misma forma gráfica o, más frecuentemente, lema, tengan significados diferentes (p. ej., *cola* puede entenderse como adhesivo o como parte de la anatomía de un animal). Generalmente, dichas ambigüedades se resuelven de forma manual examinando el contexto inmediato, aunque a veces es necesario hacer lo propio con el párrafo entero o incluso el texto completo. Esto es aún más necesario en los casos en los que la ambigüedad es intencionada. Existen escasas posibilidades de automatización de esta cuestión.

La elección como unidad elemental de formas gráficas o lemas depende del campo al que pertenece el texto. Por ejemplo, en el caso de textos científicos donde el lenguaje trata de ser claro y preciso generalmente es preferible una lematización por la reducción de número de unidades o *vocabulario* que supone (p. ej., es idéntico hablar de *molécula* que de *moléculas*). Sin embargo, en las Ciencias Sociales, como en Política, pequeños matices pueden hacer cambiar el significado y es probable que tenga más sentido utilizar palabras.

Como se deja entrever en párrafos anteriores, tanto el número total de ocurrencias como el de vocablos (o extensión del vocabulario) tiende a ser

menor en el caso de usar lemas. Esto es común a pesar de que es práctica habitual la descomposición de contracciones en sus elementos originales (*del* → *de el*) y de que la eliminación de términos ambiguos aumenta el tamaño del vocabulario.

Por su parte, la consideración de combinaciones de palabras como unidades complejas requiere de criterios como la inseparabilidad y la conmutación. La inseparabilidad requiere (en el caso de los segmentos repetidos) que no tenga sentido intercalar otra u otras palabras entre las palabras que conforman el segmento. La conmutación requiere que no sea posible sustituir una palabra del segmento por otra distinta sin que el significado del segmento sea alterado.

Documentos lexicométricos Son los documentos que se obtienen a partir del texto inicial mediante la elección de sus unidades elementales de análisis y del conteo de las frecuencias correspondientes a las ocurrencias de las mismas. Generalmente, la relación entre el documento lexicométrico y el texto inicial se establece a partir de un índice compuesto por coordenadas numéricas que enlazan las ocurrencias de las unidades elementales con su posición en el texto (página, párrafo, línea, ...).

El estudio de los contextos más próximos de las ocurrencias de las unidades puede realizarse mediante *concordancias*. Éstas corresponden a la presentación de una forma gráfica (o lema) junto con las formas que aparecen alrededor de ella para varias o todas sus ocurrencias.

Finalmente, en la elaboración del documento lexicométrico, suele exigirse un umbral mínimo de frecuencia. Asimismo, en el caso de formas gráficas con escaso significado per se, como artículos (*un, el, la, ...*) suele tomarse la decisión de excluirlas como unidades elementales a pesar de tener una frecuencia elevada. No obstante, pueden aparecer como parte de las unidades complejas.

Tablas léxicas La tabla léxica consiste en una agrupación de los elementos del documento lexicométrico en una matriz de palabras o lemas $\times$ número de partes en que se divide el corpus. En el caso de las respuestas a una pregunta abierta de una encuesta, la tabla puede ser una matriz de frecuencias que tiene por filas los individuos que responden a la encuesta y por columnas el número total de términos que han sido empleados en las respuestas.

Textos correspondientes a respuestas de preguntas abiertas

En el diseño y análisis de encuestas la mayoría de las preguntas se plantean de forma cerrada, ofreciendo una lista de opciones a elegir como respuesta. Tales preguntas cerradas tienen el aspecto positivo de refrescar la memoria al incluir quizás opciones que el entrevistado podría no recordar. Este aspecto,

sin embargo, puede ser suplido mediante una buena guía por parte del entrevistador, que puede conseguirse si en el proceso de diseño de la encuesta se ha realizado correctamente la encuesta piloto y el encuestador es adecuadamente formado, recordando opciones posibles y dejando en todo caso la respuesta al entrevistado. En general, puede decirse que el hecho de usar preguntas cerradas responde más bien a la facilidad de transmisión de la información generada por las mismas a un ordenador para un análisis cuantitativo relativamente simple de la misma.

Por contra, existe una creencia generalizada de que en el diseño de encuestas es preferible en principio plantear preguntas abiertas puesto que no restringen en ningún modo la opinión de los encuestados. De forma añadida, tienden a crear un clima de confianza y de mayor complicidad con el entrevistador, además de acortar el tiempo de la entrevista en sí. Con mucha frecuencia se incluye la pregunta *¿Por qué?* a continuación de una pregunta cerrada para explicar la respuesta a la misma.

La estadística textual o análisis textual ofrece herramientas útiles de análisis para las respuestas a las preguntas abiertas. En este capítulo se van a exponer las principales herramientas del análisis de tales preguntas así como una interesante extensión, que se concreta en el análisis simultáneo de respuestas a las mismas preguntas abiertas pero realizadas en idiomas diferentes.

Textos en idiomas diferentes

Los diferentes idiomas existentes susceptibles de ser utilizados en un análisis textual de textos poseen diferencias morfosintácticas y de tradiciones ortográficas.

Algunos idiomas, como el Inglés, utilizan una única forma gráfica en muchos casos, en particular en los verbos cuando afecta a diferentes personas o en adjetivos y algunos artículos atendiendo a género diferente. Esto simplifica la fase de segmentación y hace que sea factible la consideración de palabras como unidades elementales. Sin embargo, en otros idiomas, esto no es así y se hace necesaria la lematización. Entre las razones existentes están:

- La existencia de pronombres personales enclíticos (adosados al verbo) que es frecuente en Castellano o en Euskera, por ejemplo.
- Abundancia de verbos sintéticos.
- Declinaciones, como el Ruso, el Latín o el Euskera.
- Existencia de palabras compuestas, como en el Alemán, donde pueden estar compuestas de varios sustantivos o un nombre y un sustantivo.

En la práctica, el uso de lemas es generalmente necesario y, desde luego, imprescindible si el objetivo es algún tipo de comparación de resultados entre dos o más idiomas diferentes.

Análisis de Correspondencias de tablas léxicas

En el caso del análisis textual, la tabla de frecuencias objeto del análisis de correspondencias (ver sección 2.4) es una tabla que cruza las unidades elementales y los segmentos y cuasisegmentos repetidos escogidos de un texto con partes del mismo, dependiendo de la finalidad del análisis. En el caso de texto correspondiente a una pregunta abierta procedente de una encuesta, las partes pueden bien corresponder a cada uno de los individuos que la responden y que usan, cada uno, una parte diferente del vocabulario total; o bien, a clases de individuos agrupados por categorías.

La proyección en suplementario de modalidades de otras variables es también posible. Estas modalidades suelen corresponder a datos socioeconómicos de los individuos, como edad o género, y constituyen una ayuda a la interpretación de los ejes.

Clasificación sobre los factores principales

Al igual que en otros métodos factoriales, es habitual complementar un AC textual con un Análisis de Clasificación (ver, p.ej., Lebart (1994)), generalmente combinando un método jerárquico (como el criterio de Ward) con otro de agregación en torno a centros móviles.

La fase más útil, desde el punto de vista interpretativo, de este procedimiento es la de descripción automática de las clases. Ésta se basa en examinar la desviación entre los valores promedio o proporciones (según el tipo de variable considerada) de las variables dentro de una clase dada con respecto a los mismos valores o proporciones del total de la muestra (Lebart et al. 2006). Las variables o categorías más características de cada clase pueden ser seleccionadas mediante el examen de dichas desviaciones y/o mediante valores test (Morineau 1984), que aportan una caracterización *inferencial* de las clases. Esta metodología descriptiva de las clases de una clasificación es común a otros métodos combinados de análisis factorial y clasificación, como el ACP o el ACM, por ejemplo. En el caso del AC de tablas léxicas, las clases quedan caracterizadas por las palabras y segmentos lexicales de frecuencia (proporción) mayor que en el total de la muestra.

Otras ayudas para la caracterización de la partición: Respuestas modales

La selección de palabras y segmentos *aislados* de mayor frecuencia como elementos de caracterización de una partición hace que en algunos casos sea difícil su interpretación, ya que aisladamente su significado puede ser ambiguo o poco claro al estar fuera de contexto. En este caso sería útil examinar el contexto asociado a dichos términos característicos.

Una manera de realizar lo anterior es examinar las respuestas reales, sin modificarlas ni reducirlas en absoluto, y seleccionar aquéllas que sean más representativas del cluster al que pertenecen. A este tipo de respuestas características se les denomina *respuestas modales* (Lebart et al. 1998, capítulo 6, pp. 129-145).

Existen dos criterios para la selección de respuestas modales (ver Lebart, Salem & Bécue-Bertaut (2000, cap. 8, pp. 179-184)):

1. **Selección mediante elementos característicos** Una vez que se disponen de los valores test correspondientes a los términos disponibles para la caracterización de la clase, se computan los valores test medios considerando todos los términos de cada respuesta en el cluster. Entonces se seleccionan como respuestas modales aquéllas respuestas con mayores valores test medios. Este criterio favorece la selección de respuestas cortas.

2. **Criterio $\boldsymbol{\chi^2}$** Este criterio hace uso de la tabla léxica de individuos $\times$ términos. Una vez se ha elegido la partición de la clasificación se computan distancias χ^2 entre el perfil de cada respuesta (fila de la tabla léxica) con el perfil medio de las respuestas correspondientes a la clase a la que pertenece (media de las filas de la tabla léxica que corresponden a su clase). Entonces, las respuestas más representativas son las de menor distancia χ^2 al perfil medio de la clase a la que pertenecen. Este criterio favorece la selección de respuestas más largas, a diferencia del anterior.

Las respuestas modales, en definitiva, tratan de situar en el contexto verbal, el uso de los términos característicos de las clases, para clarificar el sentido exacto de los mismos y ayudar en la interpretación correcta del análisis textual. Ninguno de los dos criterios es intrínsecamente preferible al otro y, de hecho, pueden utilizarse ambos para una mejor caracterización de los clusters obtenidos de la tabla léxica.

5.2. Análisis Factorial Múltiple de tablas mixtas (frecuencias, variables cuantitativas y variables categóricas)

En ocasiones la información disponible es mayor que la puramente contenida en un texto. En particular, cuando el texto susceptible de analizar mediante análisis textual proviene de respuestas a una encuesta, puede decirse que siempre existe información sobre preguntas cerradas, generalmente de elección múltiple, y que guardan algún tipo de relación con la pregunta abierta, ya sea de forma directa (p. ej., tras pedir una opinión ó valoración sobre algo, a continuación se pregunta *¿Por qué?*) o ya sea porque el contexto y objeto de la encuesta es compartido.

El caso descrito en el párrafo anterior da lugar a respuestas que se codifican como variables de naturaleza diferente. En el caso de la pregunta abierta, la tabla léxica que se puede construir a partir del vocabulario de respuestas conforma una tabla de contingencia que cruza términos empleados (cualquiera que sea la unidad léxica elemental escogida) con individuos. En el caso de las preguntas cerradas, éstas dan lugar a unas variables categóricas que contienen los valores numéricos asignados a las respuestas. Dichos valores numéricos pueden ser ordinales o simplemente nominales pero esto no implica ninguna diferencia en lo que sigue. La encuesta puede asimismo contener preguntas con respuestas codificadas como variables cuantitativas o métricas.

Las tablas correspondientes a distintos tipos de variables requieren análisis específicos (véase sección 2.5). Sin embargo, el Análisis de Tablas Múltiples Mixtas plantea el análisis conjunto de la información de dichas tablas.

El problema de análisis de tablas mixtas ha sido estudiado previamente. Gower (1971) propuso una manera de equilibrar variables cuantitativas y categóricas utilizando distancias específicas para cada tipo y estandarizando su recorrido antes de agregar en torno a una distancia global. Además, las variables podían tener pesos diferentes definidos por el usuario. Este procedimiento fue generalizado en Podani (1999) y Grabmeier & Rudolph (2002) para considerar variables ordinales y distancias diferentes, respectivamente. En el caso de tablas de frecuencias o de contingencia, en Goitisolo (2002) y otros artículos posteriores se plantea el Análisis Simultáneo como un método muy versátil para analizar varias tablas de un modo conjunto y equilibrado.

Una extensión del Análisis Factorial Múltiple para incluir tablas múltiples de frecuencias se trata en Bécue-Bertaut & Pagès (2004). El caso que nos ocupa, en el que las tablas múltiples están compuestas por tablas mixtas de frecuencias, variables categóricas y variables cuantitativas se encuentra en

Bécue-Bertaut & Pagès (2008). El procedimiento para realizar un AFM de una gran tabla mixta consiste en realizar un ACP sobre la subtabla o subtablas de variables cuantitativas y ACP ponderados sobre las tablas de variables cualitativas (Pagès 2002) y de frecuencias (Bécue-Bertaut & Pagès 2004). En el primer caso se obtienen los resultados de un ACM de la tabla de variables cualitativas, mientras que en el segundo se obtiene un pseudo-AC, que sólo difiere del AC si se comparan varias tablas de frecuencias y tienen muy diferentes frecuencias marginales en las filas.

Un problema subyacente en un AFM sobre una tabla mixta es la diferencia entre los pesos de los individuos, que suelen ser uniformes en los casos de las tablas de variables cuantitativas y categóricas. En el caso de las tablas de frecuencias, éstos no son uniformes, sino iguales a la marginal de las filas. Si la tabla de frecuencias es una tabla léxica, este término recuenta el número total de términos (de las formas elegidas) que emplea cada individuo y la aplicación de ese peso supone favorecer a los individuos que dan respuestas más largas. Generalmente se suele considerar como preferible esta segunda opción, tal y como se ha realizado en el marco del AFM (Abdessemed & Escofier (1996), Bécue-Bertaut & Pagès (2008)).

En este capítulo se va a presentar una extensión de este último caso. En particular, se trata de un conjunto de datos compuesto por frecuencias y variables categóricas provenientes de las respuestas a preguntas abiertas y cerradas de una encuesta pero que se realiza en idiomas diferentes. Es decir, es una misma encuesta, presentada en dos idiomas y que se lleva a cabo en el mismo periodo de tiempo y sobre colectivos similares. Los resultados de la encuesta tienen una parte común a los dos idiomas, las respuestas a las preguntas cerradas que son interpretadas y codificadas de igual manera. Además, existe una parte diferente formada por las tablas léxicas de frecuencias correspondientes a los términos empleados en las respuestas a las preguntas abiertas, necesariamente distintos al tratarse de dos idiomas. La diferencia es debida tanto a las dimensiones de la tabla de frecuencias como a las frecuencias contenidas en ella. El objetivo es múltiple y se concreta en obtener:

1. Una tipología de las respuestas en base a ambos tipos de preguntas y una caracterización de las mismas.

2. Una agrupación o clasificación de grupos homogéneos representativa de la tipología anterior.

3. Una comparación que nos permita estimar si hay diferencias apreciables entre los dos colectivos en cuanto a las respuestas en los dos idiomas.

Las tablas mixtas están compuestas por tablas yuxtapuestas conteniendo

dos o más tipos diferentes de *variables*: una tabla de frecuencias, una tabla de variables categóricas y, finalmente, una tabla de variables cuantitativas.

Adicionalmente, se va a extender el análisis a una situación en la que existen dos (o más) grupos de individuos sobre los que se tiene información para las mismas variables categóricas (y, quizás, cuantitativas) pero diferente en lo que respecta a la tabla de frecuencias.

5.2.1. Tablas apiladas de variables categóricas

En los métodos de análisis de encuestas es habitual disponer de tablas de variables categóricas provenientes de respuestas a preguntas de elección múltiple y, por tanto, cerradas. En este caso, es posible tener una noción de la diferencia entre grupos de individuos (por ejemplo, porque visualizan la encuesta en idiomas distintos) mediante una variable categórica adicional indicadora de la pertenencia a cada grupo.

Dado que el número de individuos no tiene por qué ser idéntico ($I = I_1 + I_2 + \ldots + I_n$) no es factible, en principio, la aplicación estándar de un método de tablas múltiples como el AFM para un análisis global.

Es habitual, en este caso y dado que las variables categóricas y sus categorías son idénticas para los grupos de individuos, apilar las n tablas de forma que se obtiene una única tabla de variables categóricas de dimensión $I \times J$. Esta tabla así definida permite examinar las posibles diferencias entre los n grupos añadiendo una variable categórica adicional que tenga en cuenta el grupo al que pertenece cada individuo. Dicha variable no interviene en la obtención de ejes y se proyecta en suplementario *a posteriori*.

El tratamiento de la gran tabla compuesta de las tablas apiladas puede realizarse mediante un Análisis de Correspondencias Múltiples, que ha sido expuesto en la sección 2.5. No obstante, este tratamiento ni equilibra la influencia de los grupos sobre el análisis, ni proporciona información sobre la estructura interna en grupos de los datos.

5.2.2. Análisis de Tablas Mixtas

El marco especificado en la sección 5.2.1 está enfocado en el tratamiento de dos o más tablas de datos de idéntica naturaleza, cuyo análisis por medio de un procedimiento estándar de tablas múltiples, como el AFM de Escofier & Pagés (1992), no es adecuado dado el diferente número de individuos con el que cuentan o pueden contar las tablas objeto de análisis.

No obstante, lo anterior no es del todo cierto. Es verdad que tablas con las mismas variables no pueden yuxtaponerse, en un sentido clásico, si no se dispone del mismo número de individuos para todas las tablas. Pero dos tablas

de diferente efectivo pueden combinarse y analizarse conjuntamente mediante AFM si pueden combinarse de forma que mantengan al menos un subespacio de referencia común, esto es, manteniendo en común un cierto número de variables. Esto puede visualizarse esquemáticamente en la Figura 5.1.

	Variables comunes	**Variables particulares de la tabla 1**	**Variables particulares de la tabla 2**
Individuos $(I_1 + I_2)$	X_A	X_{1B}	0
		0	X_{2B}

Figura 5.1: Tablas de diferente número de individuos relacionadas a través de un subespacio común de representación.

Esta modelización para la representación conjunta de tablas de datos ha sido usada anteriormente en varias ocasiones, por ejemplo, en Bárcena (2001), aunque en este caso las variables particulares eran tratadas como suplementarias. Además, es posible que las subtablas correspondientes a las variables comunes y a las variables particulares de las tablas iniciales sean de diferente naturaleza. Por ejemplo, que las variables comunes sean variables categóricas susceptibles de ser analizadas parcialmente mediante un ACM, pero que las variables particulares de cada tabla sean continuas o, incluso, que sean tablas de frecuencias como las tablas léxicas que se obtienen en el análisis textual. En tal caso, estaríamos ante un análisis conjunto de tablas de diferente naturaleza o mixtas, que ha sido tratado en Bécue-Bertaut & Pagès (2008). Estos autores, sin embargo, no analizan situaciones como las representadas en la figura 5.1, sino que consideran el análisis de tablas múltiples a través de un AFM de matrices con el mismo número de individuos para todas las variables analizadas.

Vamos a realizar una exposición del AFM de Escofier & Pagés (1992), de la extensión a tablas mixtas de Bécue-Bertaut & Pagès (2008) para finalizar con la extensión al caso descrito en la Figura 5.1.

5.2.3. El Análisis Factorial Múltiple de tablas de idéntica naturaleza

El Análisis Factorial Múltiple permite el análisis conjunto de una serie de tablas relacionadas entre sí por la naturaleza de los datos que contienen. La

relación viene dada porque para cada tabla, por un lado, los individuos son los mismos y, por otro, las variables pueden constituir grupos temáticos relacionados entre sí. En el ejemplo clásico de Escofier & Pagés (1992) perteneciente a la Sensometría o Análisis Sensorial, diversos jueces (o grupos de ellos) realizan valoraciones y asignan puntuaciones a las diferentes características de un número determinado de vinos, características que se agrupan en: vista, olfacción en reposo, olfacción tras agitado, origen, etc. Cada grupo de características, elaborado sobre todos los vinos, conforma una tabla. También se puede pensar en los datos de una encuesta realizada sobre un mismo panel de individuos y repetida en varios momentos de tiempo y/o circunstancias diferentes.

El esquema inicial supone una yuxtaposición de J tablas o *subtablas* denominadas $X_j,\ j = 1, \ldots, J$ en horizontal hasta formar una tabla completa denominada X, como en la Figura 5.2. El número de individuos I es único para las subtablas y coincide con el de la tabla completa y cada tabla tiene un número K_j de variables, siendo $K = \bigcup K_j$ el conjunto de variables total de la tabla completa. J, K y K_j son, a la vez, el conjunto de tablas de variables y su cardinal.

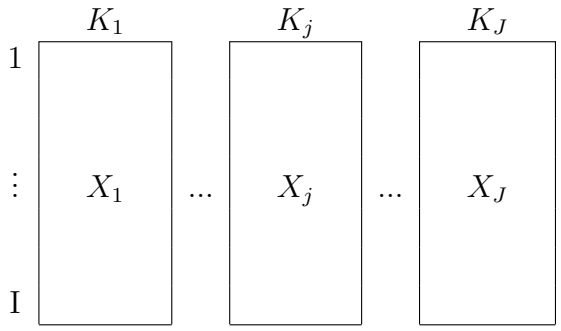


Figura 5.2: Tabla completa de datos y las J subtablas.

Al igual que en el ACM, los individuos están afectados por un peso $p_i,\ i = 1, \ldots, I$ tal que $\sum_i p_i = 1$ y las variables por un peso $m_k,\ k = 1, ..., K$ que pueden agruparse en torno a las matrices diagonales de pesos D (individuos), M_j (para las variables de la tabla j) y M (para todas las variables de la tabla completa).

El AFM consiste en un ACP ponderado de la tabla completa de la figura 5.2 pero con algunas particularidades:

1. En el caso de que las variables sean **cuantitativas**, el AFM consiste en realizar un ACP de la tabla completa, ponderando las variables de cada grupo por el inverso del primer valor propio de un ACP separado o parcial de las variables pertenecientes al grupo j. Es decir, el peso

correspondiente a la variable k perteneciente a la subtabla j es $m_k = 1/\lambda_1^j,\ \forall k \in K_j$, elemento correspondiente de la matriz diagonal M. En el ACP separado previo, del que se obtiene λ_1^j, la variable k tiene un peso igual a 1. Normalmente, en este caso, todos los individuos tienen peso uniforme e igual a $p_i = 1/I$.

Efecto de la ponderación del AFM El peso asignado a cada variable de la tabla j-ésima, $1/\lambda_1^j$ significa que la inercia proyectada máxima de las variables de un grupo j es siempre 1, al estar ponderadas por el valor $1/\lambda_1^j$ que es el inverso del primer valor propio o inercia proyectada sobre el primer factor, precisamente el de mayor inercia. Este peso provoca que las subtablas con una estructura interna más fuerte tengan un menor peso y sean menos determinantes a la hora de obtener los ejes globales del AFM.

2. En el caso de que las variables sean **cualitativas**, el AFM se realiza utilizando equivalencias entre el ACM y un ACP sobre la tabla disyuntiva completa ponderada de una determinada manera (ver Escofier & Pagés (1992) y Pagès (2002)). En concreto, el AFM puede obtenerse de dos maneras (suponiendo, al igual que en ACP, que todos los individuos tienen el mismo peso $p_i = 1/I$):

   - Si el número de variables categóricas dentro de la tabla j es Q_j, dado que en ACM la distancia χ^2 entre dos individuos de una misma tabla j y considerando sólo esa tabla es (ecuación (2.21)),

   $$d^2(i,i') = \frac{1}{Q_j} \sum_{k \in K_j} \frac{I}{z_{.k}} \left(z_{ik} - z_{i'k}\right)^2 \tag{5.1}$$

   ya que $f_{.k} = \frac{z_{.k}}{IQ_j}$, $f_{ik} = \frac{z_{ik}}{IQ_j}$ y $f_{i.} = 1/I$. Para obtener la misma distancia a partir de un ACP no normado sobre la TDC de variables indicadoras (centradas), basta con poner una ponderación

   $$m_k = \frac{I}{z_{.k} Q_j}$$

   a cada variable indicadora centrada. Es decir, se realiza un ACP no normado sobre la tabla de término general

   $$\frac{z_{ik} - w_k}{w_k} \tag{5.2}$$

donde

$$w_k = \sum_{i=1}^{I} p_i z_{ik} = \frac{z_{.k}}{I} \quad \text{con pesos } p_i = 1/I \text{ uniformes para } i = 1, ..., I,$$

$$\sum_{k \in K_j} w_k = Q_j$$

que da lugar a la distancia cuadrática de (5.1) añadiendo el peso $1/Q_j$. Entonces se realiza el ACP no normado de la tabla completa formada por las variables indicadoras correspondientes a las variables categóricas de las subtablas o tabla disyuntiva completa total, previamente ponderada. Adicionalmente, como en el caso de las variables cuantitativas, las variables correspondientes a la subtabla j llevan el mismo peso $1/\lambda_1^j$, donde λ_1^j es el primer valor propio correspondiente a un análisis separado mediante ACM de la subtabla j correspondiente. Una vez impuesta la ponderación de cada tabla, la distancia entre dos individuos de una misma tabla resulta:

$$d^2(i, i') = \frac{1}{Q_j \lambda_1^j} \sum_{k \in K_j} \frac{I}{z_{.k}} (z_{ik} - z_{i'k})^2 \tag{5.3}$$

Si consideramos la distancia entre individuos de tablas distintas, y como en el AFM la distancia ha de medirse sobre la tabla completa conteniendo las J tablas yuxtapuestas, esta distancia es:

$$d^2(i, i') = \sum_j \frac{1}{Q_j \lambda_1^j} \sum_{k \in K_j} \frac{I}{z_{.k}} (z_{ik} - z_{i'k})^2 \tag{5.4}$$

- También es posible realizar un ACP normado para obtener el ACM parcial de una tabla. Para que éste sea equivalente al ACM de la sección 2.5, es necesario que las variables indicadoras lleven el peso $m_k = (I - z_{.k})/IQ_j$, donde $z_{.k}$ es el número de individuos que han escogido la categoría correspondiente a la variable indicadora k (ver Pagès (2002)):

$$d^2(i, i') = \frac{1}{Q_j} \sum_{k \in K_j} \frac{I - z_{.k}}{I} \frac{I^2 (z_{ik} - z_{i'k})^2}{(I - z_{.k}) z_{.k}} \tag{5.5}$$

ya que la varianza de una variable indicadora es $(I - z_{.k})z_{.k}/I^2$. La distancia entre individuos resultante es la expresión (5.1). Al igual que en el caso anterior, las variables correspondientes a la

subtabla j llevan adicionalmente el mismo peso $1/\lambda_1^j$, donde λ_1^j es el primer valor propio correspondiente a un análisis separado mediante ACM de la subtabla j, por lo que las expresiones (5.3) y (5.4) son igualmente válidas.

Efecto de las ponderaciones del AFM El peso asignado a cada variable de la tabla j-ésima, $1/\lambda_1^j Q_j$ tiene dos componentes:

- Las variables de cada subtabla tienen una ponderación diferente, $1/Q_j$, que penaliza a las variables que pertenecen a subtablas conteniendo mayor número de variables.
- La inercia proyectada máxima de las variables de un grupo j es siempre 1, al estar ponderadas por el valor $1/\lambda_1^j$.

El efecto de estas ponderaciones es equilibrar el peso de las diferentes tablas de forma que se limita la posibilidad de que los ejes globales se vean dominados por las variables de una tabla conteniendo muchas variables o variables muy relacionadas entre sí.

3. Un último caso es aquél en el que las subtablas que figuran en la tabla completa de la figura 5.2 sean tablas de **frecuencias**. En este caso, los elementos de las matrices que ahí aparecen son de la forma f_{ikj}, que contienen las frecuencias relativas correspondientes al número de veces que aparece el elemento k para el individuo (o unidad) i sobre el efectivo total correspondiente a la tabla completa, de forma que

$$\sum_j \sum_k \sum_i f_{ikj} = 1$$

y las frecuencias marginales para las subtablas resultan:

$$f_{i.j} = \sum_{k \in K_j} f_{ikj} \quad \text{frecuencias marginales de las filas de una subtabla } j.$$

$$f_{.kj} = \sum_{i \in I} f_{ikj} \quad \text{frecuencias marginales de las columnas de una subtabla } j.$$

$$f_{i..} = \sum_{j \in J} \sum_{k \in K_j} f_{ikj} \quad \text{frecuencias marginales de las filas de la tabla total.}$$

El problema fundamental de yuxtaponer tablas de frecuencias (o tablas de frecuencias con otros tipos de tablas) se centra en las diferencias entre

las marginales de las filas de las subtablas. Si las marginales de las filas son iguales o al menos proporcionales entre subtablas, es posible extender directamente el AFM a este tipo de datos (Abdessemed & Escofier 1996). En caso contrario, es necesario modificar el método del AFM o utilizar otro diferente. Como extensión del AFM, es posible utilizar el Análisis Factorial Múltiple para Tablas de Contingencia (MFACT, en inglés) que puede verse en Bécue-Bertaut & Pagès (2004).

El método MFACT se basa en una generalización del AC obtenido como ACP ponderado de una matriz de desviaciones de las frecuencias observadas, respecto al cual el AC no es más que un caso particular. El *AC generalizado* (Escofier 1984) se obtiene como un ACP no normado sobre la tabla de término general:

$$\frac{f_{ikj} - m_{ikj}}{f_{i..}\, f_{.kj}} \tag{5.6}$$

donde se usan las frecuencias $f_{i..},\ i = 1, ..., I$ como pesos de las filas y $f_{.kj},\ k = 1, ..., K_j;\ j = 1, ..., J$ como pesos[1] de las columnas. Los elementos m_{ikj} corresponden a las frecuencias esperadas bajo un determinado modelo de independencia con las mismas marginales que la tabla completa, es decir, $m_{i..} = f_{i..}$ y $m_{.kj} = f_{.kj}$. Así, por ejemplo, en el caso del AC estándar, que corresponde a un modelo de independencia de la tabla completa, $m_{ikj} = f_{i..}f_{.kj}$ y los pesos de filas y columnas son, respectivamente, $f_{i..}$ y $f_{.kj}$.

MFACT utiliza el modelo de independencia correspondiente al Análisis de Correspondencias Interno (ICA) o AC intra tablas (Benzécri 1983, Escofier & Drouet 1983, Cazes & Moreau 1991, 2000) que corresponde al modelo de independencia intra tablas y de término general

$$m_{ikj} = \left(\frac{f_{i.j}}{f_{..j}}\right) f_{.kj} \tag{5.7}$$

de forma que, sustituyendo en (5.6), se realiza un ACP no normado sobre la tabla de término general:

$$\frac{f_{ikj} - \left(\frac{f_{i.j}}{f_{..j}}\right) f_{.kj}}{f_{i..}f_{.kj}} = \frac{1}{f_{i..}}\left[\frac{f_{ikj}}{f_{.kj}} - \frac{f_{i.j}}{f_{..j}}\right] \tag{5.8}$$

utilizando las marginales de la tabla total, $f_{i..}$ y $f_{.kj}$ como pesos de filas y columnas, respectivamente.

[1] Ambos pesos se convierten en métricas en sus respectivos espacios duales.

Finalmente, MFACT combina los pesos del análisis ICA con los del AFM tradicional. Los primeros solucionan el problema de las diferencias en las marginales de las filas, y el segundo equilibra la inercia de las diferentes subtablas. Cuando las subtablas son todas de frecuencias, se realiza un ACP no normado en la tabla total de término general (5.8) usando como pesos:

$$f_{i..} \qquad i = 1, ..., I \quad \text{para las filas, y} \tag{5.9}$$

$$\frac{f_{.kj}}{\lambda_1^j} \qquad k = 1, ..., K_j;\ j = 1, ..., J \quad \text{para las columnas} \tag{5.10}$$

donde λ_1^j es el primer valor propio asociado a la tabla j correspondiente a un ACP no normado separado sobre la tabla de término general y pesos correspondientes al ICA en (5.8). A esto Bécue-Bertaut & Pagès (2008) lo llaman AC *pseudo-separado*, puesto que no utiliza los pesos de las filas $f_{i.j}$ del AC clásico sino los pesos compromiso $f_{i..}$. Según dichos autores, la desviación que se produce respecto del AC clásico no es relevante a menos que las diferencias entre las marginales de las filas entre las subtablas sí lo sean.

Efecto de las ponderaciones del AFM En el caso de considerar un AFM sobre varias tablas de frecuencias se producen dos tipos de ponderaciones:

- La inercia proyectada máxima de las variables de un grupo j es siempre 1, al estar ponderadas por el valor $1/\lambda_1^J$, al igual que en los casos anteriores.
- Las filas de la tabla total de frecuencias llevan la ponderación $f_{i..}$ cuyo interés reside en que es el mismo para cada fila de todas las subtablas. Tiene el efecto de sobreponderar, dentro de una misma fila de la tabla global, a las filas correspondientes a subtablas de marginal débil relativo al de la tabla total y viceversa, equilibrando las subtablas en ese sentido.

5.2.4. El Análisis Factorial Múltiple de tablas mixtas

La consideración de un Análisis Factorial Multiple como un ACP ponderado sobre cualquier tipo de tablas, sean de variables cuantitativas, variables cualitativas/categóricas o de frecuencias (véase la sección 5.2.3) abre la puerta al análisis de una tabla completa con un espacio común, el de los individuos,

en la que se pueden yuxtaponer tablas de los tres tipos considerados. Así se pueden extraer factores comunes a todas las tablas consideradas después de equilibrar las tablas en la forma propuesta por el AFM (Abascal et al. 2006, Bécue-Bertaut & Pagès 2008). La única particularidad es que las subtablas llevan unas ponderaciones específicas a su propia naturaleza, como se ha especificado en la sección 5.2.3, además de la ponderación específica del AFM.

En lo que sigue se usará la notación de Bécue-Bertaut & Pagès (2008). Así, tenemos I individuos o unidades estadísticas y una tabla completa con J conjuntos o tablas de variables, de las cuales hay J_q cuantitativas, J_c categóricas y J_f tablas conteniendo frecuencias, de forma que $J = J_q + J_c + J_f$.

Para cualquier tipo de tabla, j se refiere a una tabla, k a una columna y K_j al número de columnas de la tabla j. El número de columnas de la tabla total es $K = \sum_{j \in J} K_j$

Todos los símbolos de los párrafos anteriores se refieren tanto al conjunto o tabla como a su cardinal.

Si la tabla j es cuantitativa, K_j es a la vez el número de columnas y de variables. Si es cualitativa, K_j es el número total de columnas, variables indicadoras y categorías correspondientes a las Q_j variables existentes en la tabla j. En el caso de ser una tabla de frecuencias, K_j es el número de columnas que coincide con el número de categorías (o variables) de la tabla de frecuencias. Denotamos x_{ijk} a los elementos de la tabla j-ésima de variables cuantitativas, z_{ijk} a los valores de las indicadoras de la tabla j-ésima de variables cualitativas y f_{ijk} a las frecuencias relativas sobre el efectivo total correspondiente a todas las J_f subtablas de frecuencias, como en la sección 5.2.3.

Elección de los pesos de los individuos

La existencia de subtablas de variables cuantitativas y/o categóricas, por un lado, y de subtablas de frecuencias, por otro, origina un problema de ponderación de los individuos o unidades estadísticas o filas de la tabla total. En efecto, en los primeros tipos de tablas, la ponderación es habitualmente uniforme ($p_i = 1/I$), mientras que en las tablas de frecuencias, para que el análisis parcial de las mismas sea equivalente a un AC (en realidad a un pseudo AC, como se ha indicado en la sección anterior), el peso de los individuos es igual a la marginal de las filas $f_{i..}$, que claramente no es uniforme ni similar al anterior.

1. Si se escoge $f_{i..}$ como peso de los individuos entonces los análisis parciales de las subtablas de variables cuantitativas y categóricas son ACPs o ACMs ponderados con dichos pesos. Esto se traslada a su contribución a la distancia global, que no equivaldría a la de esos métodos. Por contra, en el caso de las tablas de frecuencias, tenemos exactamente que

los resultados parciales coinciden con los del (pseudo) AC de la sección anterior. En el caso de las tablas de las variables cuantitativas, el peso $f_{i..}$ afecta al análisis al modificar medias y varianzas (antes de la extracción de los valores propios). En el caso de las tablas de variables cualitativas, se modifican los coeficientes[2] w_{kj} de la ecuación (5.2) de forma que tengan en cuenta estos pesos no uniformes, igualmente antes de extraer los ejes parciales.

2. Si se escoge $1/I$ como peso de los individuos, las consecuencias son exactamente las contrarias. Los análisis parciales de las subtablas de variables cuantitativas y categóricas coincidirían con sus respectivos ACPs y ACMs y sus contribuciones a la distancia global serían iguales a las distancias correspondientes a esos métodos. Sin embargo, en el caso de las subtablas de frecuencias, los resultados no corresponden ni al AC, ni al pseudo AC del método MFACT; por supuesto, la contribución a la distancia global tampoco es la misma a la de ninguno de ellos.

   La elección de uno u otro peso es libre, puesto que no existe ninguna razón general que haga preferible uno u otro. Bécue-Bertaut & Pagès (2008) recomiendan decidir en base a la aplicación concreta para la que se quiera utilizar el método. En particular, en el caso de que las tablas procedan de una encuesta y que las subtablas de frecuencias sean tablas léxicas producidas a partir de las respuestas a una pregunta abierta, los pesos $f_{i..}$ favorecen aquellas respuestas más largas, muchas veces asociadas al uso de un mejor vocabulario, por lo que pueden considerarse en ese sentido como preferibles a los pesos uniformes.

Propiedades del AFM de tablas mixtas

El AFM de tablas mixtas es, finalmente, un método factorial sobre el que se pueden extraer todos los resultados típicos de estos métodos, tanto coordenadas factoriales como ayudas a la interpretación tales como contribuciones y cosenos al cuadrado. El espacio global de referencia de la tabla total es el resultado de adoptar una distancia entre los individuos que tiene en cuenta las distancias con respecto a las variables de los tres tipos de tablas que se han considerado en el análisis.

Distancias en el AFM sobre tablas mixtas para la nube de individuos La distancia es una suma de distancias sobre los tres tipos de tablas

[2] A los que se añade un subíndice j para indicar que es posible que exista más de una subtabla de variables categóricas, de forma que $j = 1, ..., J_c$.

Indiv.	Subtabla variables cuantitativas	Subtabla variables categóricas	Subtabla de frecuencias	Pesos p_i
1				
i	$\frac{x_{ikj} - \overline{x}_{kj}}{s_{kj}}$	$\frac{z_{ikj} - w_{kj}}{w_{kj}}$	$\frac{f_{ikj} - \frac{f_{i.j}}{f_{..j}} f_{.kj}}{p_i f_{.kj}}$	$\frac{1}{I}$ ó $f_{i..}$
I				
Pesos de columnas	$\frac{1}{\lambda_1^j}$	$\frac{w_{kj}}{Q_j \lambda_1^j}$	$\frac{f_{.kj}}{\lambda_1^j}$	

Tabla 5.1: Términos generales y pesos asociados para la columna k-ésima de cada uno de los tres tipos de tablas consideradas en un AFM de tablas mixtas.

consideradas, ponderadas por la ponderación extra del AFM de $1/\lambda_1^j$:

$$\begin{aligned} d^2(i,i') &= \sum_{j\in J_q} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \left[\frac{x_{ikj} - x_{i'kj}}{s_{kj}}\right]^2 + \sum_{j\in J_c} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{1}{Q_j w_{kj}} \left[z_{ikj} - z_{i'kj}\right]^2 + \\ &+ \sum_{j\in J_f} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{1}{f_{.kj}} \left[\left(\frac{f_{ikj}}{f_{i..}} - \frac{f_{i'kj}}{f_{i'..}}\right) - \frac{f_{.kj}}{f_{..j}} \left(\frac{f_{i.j}}{f_{i..}} - \frac{f_{i'.j}}{f_{i'..}}\right)\right]^2 \end{aligned} \quad (5.11)$$

Cada tipo de tabla contribuye a la distancia global entre individuos en la misma medida que lo hacen el ACP, el ACM o el MFACT, según corresponda, con la ponderación adicional propia del AFM, esto es, el inverso del primer valor propio parcial $1/\lambda_1^j$ para cada tabla.

Las nubes de columnas para las tablas de variables categóricas y de frecuencias están centradas de acuerdo a los pesos empleados. En la práctica, se suelen representar en un gráfico las variables cuantitativas y en otro, simultáneamente las categóricas y las frecuencias. La relación entre todas ellas puede visualizarse en las relaciones de transición.

Relaciones de transición Como en otros métodos factoriales, pueden obtenerse las relaciones de transición entre la coordenada sobre el eje global s del individuo i, denotada $F_s(i)$ y las coordenadas de las columnas de la tabla

completa mediante (Bécue-Bertaut & Pagès 2008):

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J_q} \frac{1}{\lambda_1^j} \left[\sum_{k \in K_j} x_{ikj} G_s(kj) \right] + \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J_c} \frac{1}{\lambda_1^j Q_j} \left[\sum_{k \in K_j} z_{ikj} G_s(kj) \right] + \\ + \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J_f} \frac{1}{\lambda_1^j} \frac{f_{i.j}}{f_{i..}} \left[\sum_{k \in K_j} \frac{f_{ikj}}{f_{i.j}} G_s(kj) \right] \quad (5.12)$$

donde la relación de transición tiene tres componentes bien diferenciados, cada uno de ellos correspondiente a un tipo de subtablas. La parte de la relación correspondiente a cada tipo de subtabla es equivalente a la de ACP, ACM y MFACT, respectivamente, siempre corregida por la ponderación extra del AFM.

La relación inversa entre coordenadas de las columnas y los individuos para la tabla completa permite la interpretación clásica, salvo por un coeficiente debido al AFM, según el tipo de subtabla considerado:

- Para subtablas cuantitativas, la coordenada de una variable cuantitativa en el eje s es el coeficiente de correlación de esa variable con el factor de orden s.

- Para subtablas cualitativas, una categoría se sitúa en el centroide de los individuos que presentan dicha categoría.

- Para subtablas de frecuencias, una columna (generalmente una categoría, como un término en análisis lexical) se ve atraída por las filas (individuos en el caso de tablas léxicas) con un grado de asociación mayor que el de independencia entre filas y columnas dentro de la tabla j-ésima. La columna es repelida cuando el grado de asociación es menor que el que se produce bajo independencia de esas mismas filas y columnas.

Representación superpuesta de las nubes parciales Es posible también realizar una representación superpuesta de las J nubes de individuos correspondientes a las J subtablas, denominadas nubes parciales. Estas nubes se obtienen de los J análisis separados o parciales de las subtablas. Esto puede realizarse a partir de las relaciones de transición (5.12) si en ellas utilizamos para cada nube j las j columnas correspondientes. Ver Bécue-Bertaut & Pagès (2001, 2008).

5.2.5. El Análisis Factorial Múltiple de tablas mixtas de efectivo diferente con un subespacio común de representación

En la sección 5.2.4 se ha dibujado una situación en la que se consideran tres tipos de tablas diferentes (con variables cuantitativas, categóricas y con frecuencias), susceptibles de ser analizadas conjuntamente mediante Análisis Factorial Múltiple o alguna variante de él, relacionada con las características de los datos, como en el caso de las tablas de frecuencias.

El análisis anterior supone que todas las subtablas objeto de estudio comparten la dimensión correspondiente al número de filas, de forma que son los mismos individuos (u otras unidades estadísticas) sobre los que se dispone de información de varios tipos y grupos de variables formando las subtablas y, finalmente, la tabla global.

En esta sección el interés se centra en una variante del caso anterior, en la que se dispone de información común sobre los individuos disponibles a través de uno o varios grupos de variables medidas sobre todos los individuos. Pero además, existen otras variables (o columnas de una tabla de variables cuantitativas, categóricas o de frecuencias) que sólo están medidas sobre una parte de los individuos.

Suponemos que las variables que no están medidas sobre todos los individuos forman P subtablas que particionan la muestra en P partes, no necesariamente iguales. De esta forma, la primera subtabla contiene variables medidas sobre I_1 individuos, mientras que la segunda subtabla contiene variables medidas sobre los siguientes I_2 individuos, y así sucesivamente, de forma que $I_1 + I_2 + ... + I_P = I$. Si ordenamos los individuos de manera acorde con esta partición definida por las variables no comunes a todos ellos, se puede formar una tabla total, incluyendo variables comunes y no comunes rellenando los valores ausentes con ceros. Esto responde al esquema de la Figura 5.1 de la página 129, solo que en aquel caso la muestra está dividida en 2 partes.

Suponemos también que la tabla total (completada con ceros en los huecos en los que no hay información sobre los individuos) se ha construido de manera que las variables que no están presentes para todos los individuos corresponden en realidad a las J_f tablas de frecuencias, de forma que hay tantas partes en la muestra como subtablas de frecuencias, $P = J_f$. Así, la tabla total se presenta de la forma contenida en la Figura 5.3.

En la configuración de la Figura 5.3, el análisis de la tabla total mediante un AFM de tablas mixtas como el expuesto en la sección 5.2.4 no supone ninguna variante respecto de las subtablas de variables cuantitativas y categóricas.

	Variables Cuantitativas	**Variables Categóricas**	**Tablas de frecuencias**			
	(íd.)	V. Indicadoras	Tabla 1	Tabla 2	$\cdots$	Tabla J_f
I_1			f_{ik1}	0	$\cdots$	0
I_2			0	f_{ik2}	$\cdots$	0
$\vdots$	x_{ikj}	z_{ikj}	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_{J_f}			0	0	$\cdots$	f_{ikJ_f}
	$k = 1, ..., K_j$	$k = 1, ..., K_j$	$k = 1, ..., K_1$	$k = 1, ..., K_2$	$\cdots$	$k = 1, ..., K_{J_f}$
	$j = 1, ..., J_q$	$j = 1, ..., J_c$	$j = 1, ..., J_f$			

Figura 5.3: Tabla mixta total con variables cuantitativas y categóricas comunes y tablas de frecuencias distintas para los I individuos.

Con respecto a las subtablas correspondientes a tablas de frecuencias cambian las frecuencias marginales de las filas de la tabla total. Cuando las tablas de frecuencias se yuxtaponen, como en el caso de la sección anterior, resulta que $f_{i..} = \sum_{k \in K_j} \sum_{j=1}^{J_f} f_{ikj} = \sum_{j=1}^{J_f} f_{i.j}$. En el caso de la Figura 5.3 las tablas de frecuencias no se yuxtaponen, sino que se sitúan a lo largo de la diagonal de una matriz diagonal por bloques. En este caso, las marginales de las filas son $f_{i..} = \sum_{k \in K_j} f_{ikj} = f_{i.j}$ siendo j la subtabla correspondiente al elemento de la partición para el que tenemos frecuencias no nulas para el individuo i-ésimo. La frecuencia marginal de las filas de una subtabla j, $f_{i.j}$ vale 0 cuando corresponda a un bloque de ceros de la Figura 5.3. Las marginales de las columnas no se ven modificadas.

Por lo demás, es posible aplicar el AFM sobre tablas mixtas de la sección anterior. Para cada tipo de variables, se trata de realizar ACP no normado ponderado según el tipo de tabla por los pesos de las columnas de la Tabla 5.1, página 138. Los resultados parciales, salvo por la ponderación extra del AFM, corresponden a un ACP para las subtablas de variables cuantitativas, a un ACM para las subtablas de variables cualitativas y a un pseudo-AC (metodología MFACT) para el caso de las tablas de frecuencias.

Elección de los pesos de los individuos

Es preciso determinar los pesos de los individuos a utilizar en el AFM de la tabla total. La alternativa está entre utilizar un peso uniforme $p_i = 1/I$ como en ACP o ACM o un peso relacionado con las marginales de las filas de las tablas de frecuencias.

Sin embargo, en este caso, la elección del peso de los individuos en base a las marginales de la fila de la tabla total no es necesariamente la más conveniente.

1. Si elegimos $p_i = f_{i..}$, entonces este peso supone que la fila i lleva el peso $f_{i.j}$ para los individuos con frecuencias positivas en la tabla j. Por ejemplo, en el caso de que las tablas de frecuencias correspondan a tablas léxicas de frecuencias, los individuos que usan un léxico mayor y más variado tienen un peso mayor. Esto es lógico en el análisis textual.

2. Pueden existir alternativas diferentes, como elegir un peso

$$p_i = \frac{f_{i.j}}{J_f f_{..j}}$$

de forma que

$$\sum_{i \in I_j} p_i = \sum_{i \in I_j} \frac{f_{i.j}}{J_f f_{..j}} = \frac{1}{J_f}$$

$$\sum_{j=1}^{J_f} \sum_{i \in I_j} p_i = 1$$

De esta manera, los bloques correspondientes a las subtablas de frecuencias tienen el mismo peso para el conjunto de los individuos, pero dentro de cada bloque, cada individuo tiene un peso diferente según sea su frecuencia marginal. En el caso mencionado de diferentes tablas léxicas, esta ponderación podría ser interesante si no quisiéramos dar más peso a unas tablas que a otras simplemente por usar más o menos términos. Así, dentro de cada tabla, se preservaría el que individuos con más léxico tengan mayor peso.

Propiedades del AFM

Las propiedades, así como las distancias entre individuos y las relaciones de transición, se trasladan directamente de las ecuaciones (5.11) y (5.12).

Distancia entre dos individuos en la nube de individuos La distancia cuadrática entre dos individuos, i e i' es la expresada en la ecuación (5.11) donde, dada la disposición diagonal por bloques de las tablas de frecuencias de la Figura 5.3, la distancia entre dos individuos presenta la particularidad de que los cocientes de la forma $f_{i.j}/f_{i..}$ sólo pueden tomar el valor 0 o 1. Al igual que en la sección anterior, esta distancia tiene en cuenta, mediante una suma ponderada, las distancias debidas a variables cuantitativas, categóricas y de frecuencias de la misma manera que sus respectivos ACP, ACM o MFACT.

Relaciones de transición La relación de transición entre la coordenada sobre el eje global s del individuo i y las coordenadas de la tabla completa expuesta en la Figura 5.3 es la expresada en la ecuación (5.12) donde, nuevamente, el cociente $f_{i.j}/f_{i..}$ que aparece en el tercer término de la ecuación (5.12) solo toma los valores 0 o 1, según si estamos considerando un bloque de ceros o no.

5.2.6. Clasificación sobre los factores principales

Al igual que en otros métodos factoriales, es habitual complementar los análisis factoriales, en cualquiera de sus variedades, con un Análisis de Clasificación (ver, p.ej., Lebart (1994), en el ámbito del análisis de correspondencias), como se recuerda en la sección 2.7. Normalmente éste es de tipo mixto, combinando un método jerárquico (como el criterio de Ward) con otro de agregación en torno a centros móviles.

En el caso del Análisis Factorial Múltiple de tablas mixtas, se puede realizar también un análisis de clasificación sobre los factores globales, como la clasificación mixta descrita en la sección 2.7. Conviene notar que, en el caso del AFM de tablas mixtas, los pesos de los individuos pueden venir dados por alguna de las expresiones de la página 142 y no tienen por qué ser, por tanto, uniformes. Esto debe tenerse en cuenta en el cómputo de los índices de nivel de la jerarquía en base a la ecuación (2.72).

5.3. Aplicación: Encuesta on-line sobre Satisfacción sobre la Universidad y productos corporativos en dos idiomas

5.3.1. Descripción de las variables utilizadas

Los datos objeto de análisis corresponden a la encuesta on-line realizada en el año 2005 a los miembros de los diferentes colectivos pertenecientes a la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) sobre satisfacción con la institución y la valoración de productos corporativos con la imagen de la Universidad. La encuesta aparece en el Apéndice A y la descripción de las características de la misma así como de sus principales variables en la sección 3.3.

En este capítulo, seleccionamos las respuestas correspondientes a las preguntas cerradas correspondientes a la satisfacción con la institución y a la binaria que preguntaba por la disponibilidad genérica a comprar un artículo con el logotipo de la UPV/EHU (ver sección 3.3). La intención es que estas variables sean consideradas como activas en los análisis, mientras que las variables socioeconómicas de la encuesta puedan ser consideradas, de nuevo, como ilustrativas.

Adicionalmente a las preguntas cerradas anteriores, la encuesta contenía una **pregunta abierta**. Las respuestas a esta pregunta proporcionan un texto por encuestado, es decir, van a dar lugar a una tabla léxica de individuos $\times$ términos empleados en las respuestas. Esta pregunta se realizaba a continuación de la pregunta cerrada

> *¿Estarías interesado en comprar un producto con el logotipo de la UPV/EHU (tal como pañuelos, vacíabolsillos, camisetas, relojes, tazas, ...) para uso personal o para regalo?*

presentada en la página 59. La pregunta abierta era, simplemente,

> *¿Podrías escribir aquí por qué?*

Los encuestados podían responder libremente escribiendo un texto en un recuadro de diálogo clásico de una página web.

Conviene recordar, además, que la encuesta tiene la particularidad de que se planteó en dos idiomas diferentes (Euskera y Castellano). La elección del idioma era libre para el encuestado, resultando 304 respuestas en Euskera y 1243 en Castellano. El objetivo de esta aplicación es realizar un análisis de tablas múltiple, teniendo en cuenta que:

1. tenemos un análisis de tablas múltiple donde las subtablas tienen distintos tipos de variables, categóricas y frecuencias provenientes de un tabla léxica, y

2. los individuos que responden a las preguntas lo hacen en dos idiomas diferentes, por lo que los términos lexicales empleados no pueden coincidir por ese simple hecho.

El objetivo de esta aplicación es determinar, con ayuda de técnicas exploratorias multivariantes adecuadas, qué factores hay detrás de las respuestas a las preguntas seleccionadas y qué diferencias puedan encontrarse entre los colectivos que emplean idiomas diferentes, y utilizando adicionalmente como ilustrativas las variables socioeconómicas habituales en este tipo de encuestas.

5.3.2. Análisis de Correspondencias Múltiples (ACM) y clasificación sobre los factores principales. Tablas de idiomas apiladas

El primer objetivo del análisis es una descripción de la muestra procedente de la encuesta on-line, a partir de las variables categóricas mediante un Análisis de Correspondencias Múltiples (ACM) más una clasificación sobre sus ejes principales con una caracterización de la misma. En el ACM se utilizan como activas las variables correspondientes a las preguntas cerradas relacionadas con la satisfacción declarada por los encuestados (`Satis2`) y el interés mostrado por productos con logotipo (`BuyLogo`). Como variables suplementarias se usan las variables de caracterización enumeradas en la página 60 (*género, edad, vinculación, campus*) más una variable categórica adicional correspondiente al idioma empleado en la encuesta (Euskera o Castellano).

Sobre la variable `Satis` hay que decir que las categorías que representan los valores 1 y 2, correspondientes a las respuestas *En absoluto satisfecho/a* y *Poco satisfecho/a* tenían un efectivo extremadamente débil, por lo que se han unido a la categoría media 3 (*Medianamente satisfecho*). De esta manera, se define la variable `Satis2` que queda codificada de la siguiente manera:

1. *Nada, poco o medianamente satisfecho/a:* `Satis2=1`.

2. *Bastante satisfecho/a:* `Satis2=2`.

3. *Muy satisfecho/a:* `Satis2=3`.

Por su parte, la variable `BuyLogo` (*¿Estarías interesado en comprar... ?*) queda codificada como sigue:

1. *Sí:* `BuyLo=1`.

2. *No:* `BuyLo=2`.

Existen algunos valores ausentes en los datos sobre las variables activas que son reasignados aleatoriamente sobre el resto. El detalle puede consultarse en la tabla B.1 del apéndice B, página 221. En esa tabla puede apreciarse que, en este caso, a los individuos no se les va a asignar el peso uniforme $p_i = 1/I$ tradicional; la razón será evidente más adelante. Estos pesos van a ser tales que $\sum_i p_i = 200\,\%$, lo cual va a afectar al valor de la inercia total, que queda dividida por 2 (ver ecuaciones (2.60)-(2.64)).

La yuxtaposición de las tablas de los dos idiomas, supone, salvo por los pesos empleados, que a todos los individuos se les da la misma importancia, independientemente del idioma que hayan utilizado para responder.

Los resultados del ACM están en las Tablas 5.2 y 5.3. El plano principal conteniendo las proyecciones de las variables activas están en la Figura 5.4. Dado el escaso número de categorías totales, 2 ejes reproducen bastante bien el comportamiento de los individuos encuestados, con una tasa de inercia proyectada del 76,58 %. La interpretación de los ejes es muy clara. El primer eje opone individuos muy satisfechos con la universidad (`Satis2`=3) y al mismo tiempo interesados en comprar artículos con el logotipo de la universidad (`BuyLo=1`) a individuos con escasa o nula satisfacción (`Satis2`=1) y nada interesados en comprar (`BuyLo=2`). El eje 2, por su parte, opone individuos bastante satisfechos a individuos muy satisfechos.

Traza de la matriz 1.50000			
Número	Valor propio	%	% acumulado
1	0,6488	43,25	43,25
2	0,5000	33,33	76,58
3	0,3512	23,42	100,00

Tabla 5.2: Valores propios y tasas de inercia del ACM sobre las variables categóricas activas. Ambos idiomas apilados.

Las contribuciones a los dos primeros ejes se reparten entre las 5 categorías activas, siendo el segundo eje debido casi en exclusiva a la oposición entre las dos categorías de satisfacción mencionadas. Los cosenos cuadrado indican que todas las categorías, salvo `Satis2`=2 en el primer eje y `Satis2`=1 en el segundo, están bastante bien representadas en uno de los dos primeros ejes, con valores que oscilan entre 0,60 y 0,95, con lo que el plano principal parece un buen resumen de estos datos.

Categoría	Peso relativo	Dist. al origen	Eje 1	Eje 2	Eje 3
Coordenadas					
BuyLo=1	30,053	0,66375	-0,66	0,00	0,48
BuyLo=2	19,947	1,50661	0,99	0,00	-0,73
Satis2 = 1	11,835	3,22482	1,39	0,48	1,02
Satis2 = 2	21,348	1,34216	-0,21	-1,13	-0,16
Satis2 = 3	16,817	1,97314	-0,71	1,09	-0,52
Contribuciones					
BuyLo=1	30,053	0,66375	19,95	0,00	19,95
BuyLo=2	19,947	1,50661	30,05	0,00	30,05
Satis2 = 1	11,835	3,22482	35,39	5,55	35,39
Satis2 = 2	21,348	1,34216	1,47	54,37	1,47
Satis2 = 3	16,817	1,97314	13,14	40,08	13,14
Cosenos cuadrado					
BuyLo=1	30,053	0,66375	0,65	0,00	0,35
BuyLo=2	19,947	1,50661	0,65	0,00	0,35
Satis2 = 1	11,835	3,22482	0,60	0,07	0,33
Satis2 = 2	21,348	1,34216	0,03	0,95	0,02
Satis2 = 3	16,817	1,97314	0,26	0,60	0,14

Tabla 5.3: Coordenadas, contribuciones y cosenos cuadrado para las categorías activas del ACM.

Es muy interesante, en este caso, comprobar la relación con los ejes de las categorías suplementarias. Las proyecciones de las mismas sobre el plano principal están en la Figura 5.5 y un resumen aparece en la Tabla B.2 del apéndice B, página 222. Como ayudas a la interpretación tenemos la tabla de valores-test, que incluyen los valores correspondientes a las categorías activas y a las suplementarias. Están recogidas en la Tabla 5.4.

El gráfico de la Figura 5.5 indica cómo el primer eje, que en su lado izquierdo contiene las proyecciones correspondientes a los individuos más satisfechos e interesados en los productos corporativos, está asociado también a la edad y a la vinculación de los individuos con la universidad. En el lado izquierdo se proyectan individuos que son PDI o PAS y de mayor edad (los estratos 3 y 4 corresponden a individuos con edades comprendidas entre 30 y 44 años y a mayores de 45 años, respectivamente). En el lado derecho, jóvenes menores de 29 años y estudiantes, asociados a una menor satisfacción y a un menor interés por los productos. Ni el idioma ni el género parecen proyectarse de forma clara sobre este eje, ni sobre el segundo.

Los valores test calculados respecto de las categorías suplementarias reflejan

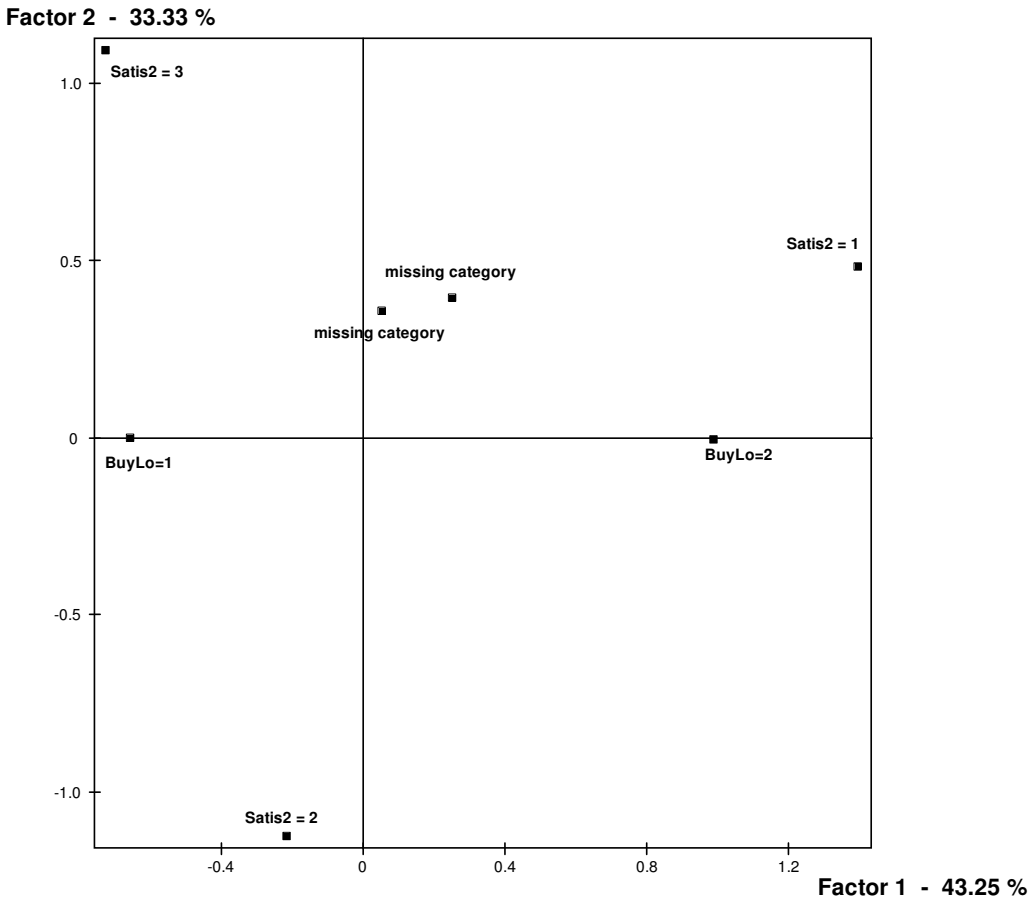


Figura 5.4: Plano principal del ACM de las variables categóricas activas `BuyLogo` y `Satis2`. 2 idiomas apilados.

cómo las categorías con valor test superior al valor crítico 1,96 para un nivel de significación del 5 % sobre el eje 1 son las de estudiantes, PDI, y edades entre 23-29 y 45 años o más. Tanto las dos de vinculación como las de edad tienen signos opuestos apareciendo en lados opuestos del eje (y del plano) con lo que refuerza la interpretación anterior. El segundo eje no contiene ninguna modalidad suplementaria con un valor test significativo.

En definitiva, el primer eje es claramente un eje de visión general sobre la universidad, que se asocia a una predisposición del mismo signo ante la apertura de la tienda corporativa. Asimismo, se caracterizan los grupos de individuos causantes de esta distinción en base a su edad y al tipo de vinculación con la universidad. No se aprecian diferencias significativas respecto del idioma empleado en la encuesta, por lo que ambos colectivos tienen un comportamiento

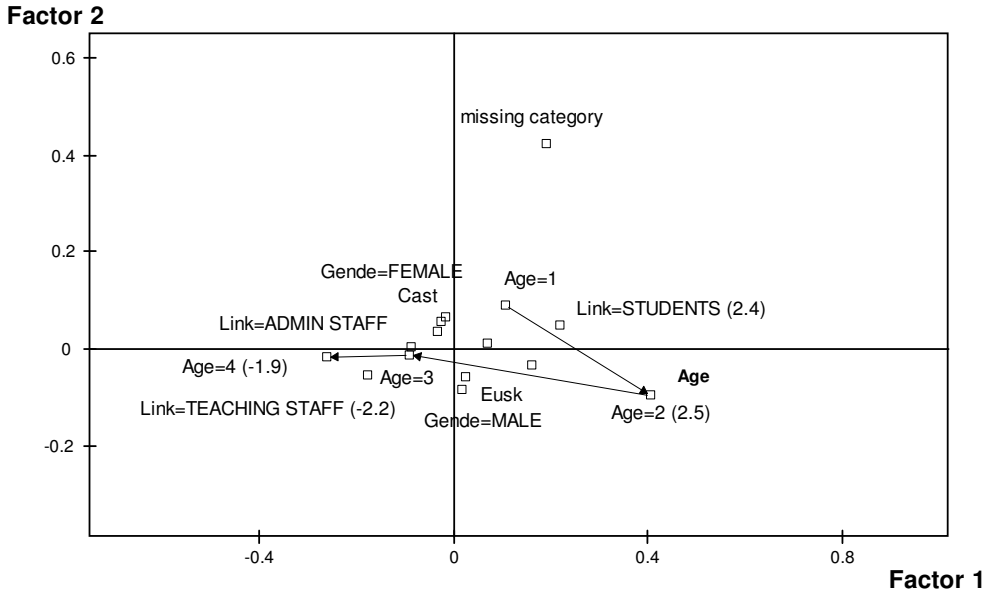


Figura 5.5: Proyecciones sobre el plano principal del ACM de las variables categóricas suplementarias.

similar.

Categoría	Efectivo	Peso absoluto	Dist. al orig.	Eje 1	Eje 2	Eje 3
BuyLo=1	963	118,78	0,67705	-11,19	-0,04	8,31
BuyLo=2	567	78,39	1,54114	11,22	-0,06	-8,23
missing cat.	17	2,03	97,12820	0,08	0,51	-0,55
Satis2 = 1	373	46,51	3,28296	10,86	3,76	7,91
Satis2 = 2	660	84,38	1,36075	-2,59	-13,62	-1,82
Satis2 = 3	499	65,62	2,03566	-7,15	10,77	-5,00
missing cat.	15	2,69	73,05210	0,42	0,65	-0,86
Students	509	75,55	1,63667	2,43	0,53	0,31
Admin Staff	371	35,43	4,62236	-0,22	0,23	0,39
Teaching Staff	667	88,22	1,25800	-2,20	-0,69	-0,60
Masc.	705	88,54	1,24984	0,22	-1,04	1,18
Femen.	842	110,66	0,80011	-0,22	1,04	-1,18
Araba	232	34,57	4,76223	1,04	-0,21	-0,17
Bizkaia	928	108,44	0,83696	-1,34	0,06	0,93
Gipuzkoa	387	56,19	2,54512	0,61	0,11	-0,89
Age=1	287	44,09	3,51804	0,80	0,66	0,15
Age=2	208	32,54	5,12170	2,53	-0,59	0,80
Age=3	571	76,54	1,60256	-1,01	-0,14	-1,20
Age=4	457	43,85	3,54277	-1,94	-0,13	0,57
missing cat.	24	2,18	90,37630	0,28	0,63	-0,11
Eusk	304	100,06	0,99081	0,36	-0,79	0,07
Cast	1243	99,14	1,00928	-0,36	0,79	-0,07

Tabla 5.4: Valores test de las proyecciones de las categorías activas y suplementarias.

Validación: estabilidad de las formas

El comportamiento de las categorías suplementarias correspondientes al idioma utilizado en la realización de la encuesta tiene un particular interés, como ya se ha comentado anteriormente.

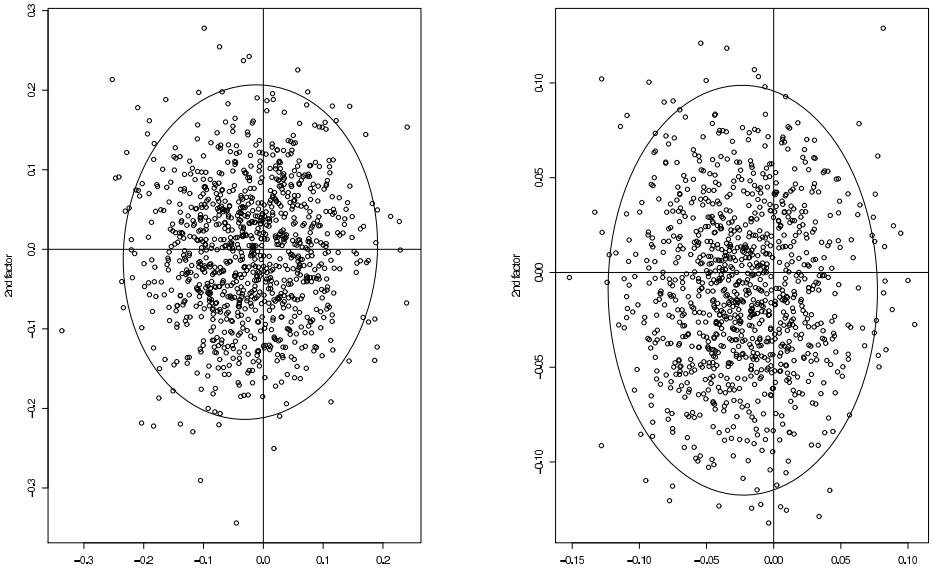
En esta sección se completa el análisis mediante un bootstrap parcial sobre dichas categorías, lo que permitirá valorar su estabilidad sobre el plano principal. Se han realizado 1000 réplicas[3] de muestras bootstrap de tamaño 1547, igual al de la muestra original. Dichas réplicas se han proyectado como suplementarias en los ejes originales. Se han calculado también medias y desviaciones típicas de las mismas sobre los ejes (véase Tabla 5.5).

[3]Lebart (2006) observa que en ACM no es necesario un número tan elevado de réplicas; de hecho probando otros valores más pequeños como 100 réplicas se obtuvieron resultados muy similares.

	Euskera			Castellano		
	Eje 1	Eje 2	Eje 3	Eje 1	Eje 2	Eje 3
Original	0.0254	-0.0560	0.0049	-0.0257	0.0565	-0.0050
Media	-0.0216	-0.0036	0.0601	-0.0230	-0.0094	0.0629
Desviación típica	0.0869	0.0859	0.0877	0.0410	0.0442	0.0436

Tabla 5.5: Medias y desviaciones típicas de las réplicas bootstrap (parcial) de las proyecciones del idioma (Euskera o Castellano) empleado por los encuestados.

Adicionalmente, se han dibujado las proyecciones en suplementario de las muestras bootstrap de las categorías correspondientes a los dos idiomas (véase Figura 5.6). Esta gráfica permite vislumbrar una mayor estabilidad en las repuestas de Castellano que en las de Euskera. En cualquier caso, la elipse contiene el origen de coordenadas, lo que concuerda con los valores test examinados anteriormente. Esta resultado corrobora el obtenido anteriormente: no parece haber diferencias significativas en las respuestas consideradas entre los dos idiomas.



(a) Euskera. (b) Castellano.

Figura 5.6: Boootstrap de las categorías del idioma proyectadas sobre el plano principal como suplementarias y elipse de confianza al 95 %.

Clasificación sobre los factores principales

A continuación, se ha realizado una clasificación sobre los factores principales del ACM de las tablas apiladas. En este caso, del ACM sólo se extraen 3 factores, que son los que vamos a utilizar como variables de clasificación.

La clasificación es de tipo jerárquica, utilizando el criterio de Ward generalizado más una consolidación alrededor de centros móviles. La selección del número de clases se realiza tradicionalmente a partir del dendrograma o del histograma de índices de nivel, que aparece en la Figura 5.7. La caída de los índices de nivel es importante en los 3 últimos nodos, que dan lugar a una partición en 4 clusters. Se comprueba como la suma de los índices de nivel, equivale a la inercia total de la tabla inicial objeto del ACM, dado que no se ha reducido el número de factores.

```
Descripción de los nodos
Núm. Prim. Últ. Efec.  Peso  Índice   Histograma de índices de nivel
 7     3    6     2    47,15 0,11215  ****
 8     2    5     2    67,00 0,13828  ********
 9     4    1     2    85,05 0,20530  ******************
10     8    9     4   152,05 0,50322  ********************...**********
11     7   10     6   199,20 0,54105  ********************...**********
Suma de índices de nivel =    1,50000
```

Figura 5.7: Selección del número de clases.

Tras la selección del número de clases o clusters de la partición se realiza una consolidación de la misma a través de un algoritmo de centros móviles. Los detalles sobre composición inicial de los clusters, coordenadas y valores test antes y después de la consolidación, cambios en la proporción de inercia inter sobre la inercia total y distancias entre clusters están en el Apéndice B.1.2. La consolidación da exactamente el mismo resultado que la partición inicial, resultando en una partición muy estable. Esto se ve favorecido, sin duda, por el pequeño número de categorías y de factores seleccionados para la fase de clasificación en este estudio concreto.

La representación de los centros de los clusters de la partición en 4 clases sobre el plano principal del ACM puede observarse en la Figura 5.8. La superficie de los centros es proporcional al tamaño de los clusters. Los 4 clusters están claramente posicionados diferenciados unos de otros en los 4 cuadrantes del plano.

La interpretación de los diferentes clusters puede realizarse a partir de la interpretación de los ejes del ACM, en relación con la proyección de las variables suplementarias disponibles, de la posición de sus centros en el plano

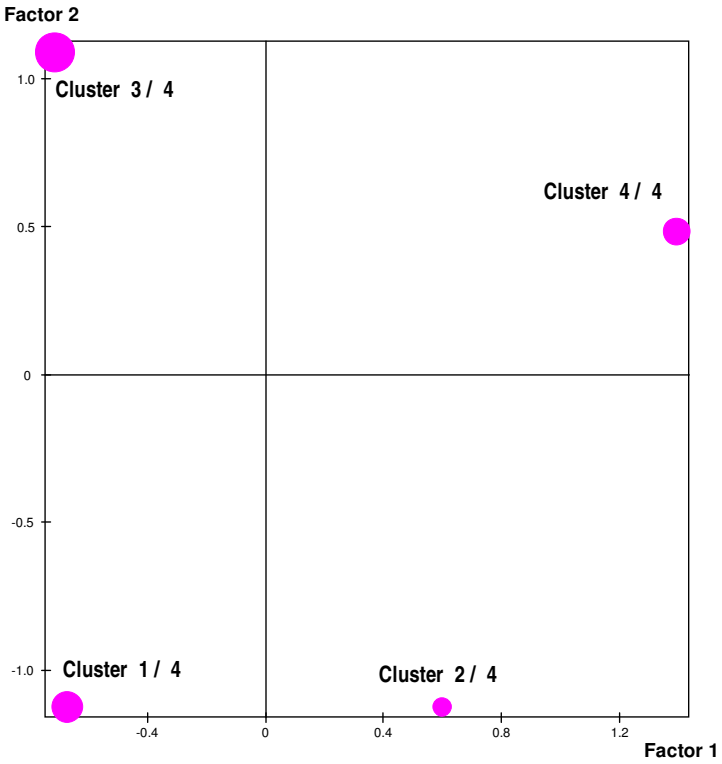


Figura 5.8: Plano principal del ACM sobre las variables categóricas activas con los centroides de la partición en 4 clusters.

y de las proporciones de las diferentes categorías en cada cluster en relación a las proporciones en la muestra total y a los valores test correspondientes, que aparecen en las Tablas B.8, B.9, B.10 y B.11 del apéndice B (ver págs. 224-227).

- El centro del **cluster 1** se posiciona en el tercer cuadrante, asociado a encuestados bastante satisfechos y con propensión declarada a comprar productos corporativos. Estarían asociados, en principio, a Personal Docente e Investigador y de mayor edad. Según la Tabla B.8, en base a valores test, sólo las categorías activas mencionadas describen sustancialmente este cluster, aunque sí hay una mayor proporción de PDI y de género masculino (por citar los 2 valores test mayores) que en la muestra. Es un cluster de tamaño medio, dada la partición escogida.

- Con respecto al **cluster 2**, éste se posiciona en el cuarto cuadrante, asociado a personas bastante satisfechas y con baja propensión a comprar productos con logotipo. La Tabla B.9 contiene las proporciones y valores

test que ayudan en su caracterización.

Se observa cómo, efectivamente, las categorías características son la escasa propensión a comprar y que son individuos bastante satisfechos con la universidad. Las categorías suplementarias más características (no desde el punto de vista estadístico, sólo desde el de la diferencia entre proporciones cluster/muestra total) son las del campus de Gipuzkoa y género femenino. Es un cluster relativamente pequeño (15 % del total de la muestra).

- El **cluster 3** (categorías características en Tabla B.10) tiene un centro que aparece situado en el cuadrante dos, es decir, donde existe propensión a comprar y la satisfacción es mayor. En efecto, esas categorías son las más características, con valores test mayores que el umbral 1,96 y nuevamente sin categorías suplementarias significativas. Sí hay una mayor proporción de género femenino y, menos, de personas entre 30 y 44 años. Es un cluster de tamaño mayor al que le correspondería si todos fuesen iguales (25 %).

- Finalmente el **cluster 4**, situado en el primer cuadrante, tiene las categorías características esperables por su posición en el plano principal. En la Tabla B.11 se observa cómo éstas son las correspondientes a satisfacción media o baja, escasa propensión a comprar y personas entre 23 y 29 años. De manera menos significativa, hay una proporción mayor que en la muestra de estudiantes y personas aún más jóvenes (entre 18 y 23 años). El tamaño del cluster es cercano al 24 %.

Conclusión

En esta sección se trata de explorar las opiniones de los encuestados atendiendo a sus respuestas a las preguntas cerradas, teniendo en cuenta la información sobre las variables suplementarias de caracterización disponibles. Conociendo, además, el idioma en que los individuos han respondido al enunciado de la encuesta (Euskera o Castellano) también se trata de ver si existe alguna diferencia estructural de respuesta entre ambos tipos de personas.

En el análisis efectuado, ACM de las categorías correspondientes a las preguntas cerradas más clasificación sobre los factores principales (en este caso todos los obtenidos, no solo los principales), permite concluir que no se aprecian diferencias significativas debidas al idioma empleado. Este resultado corresponde tanto al total de la muestra como a cada uno de los grupos que se obtienen en la fase de clasificación. En este último caso, sólo la variable edad tiene suficiente relevancia y sólo para uno de los grupos como elemento diferenciador.

Se aprecian algunas diferencias menores, dignas de comentar, pero de escasa significación estadística.

5.3.3. Análisis Factorial Múltiple de tablas mixtas: Análisis conjunto de respuestas cerradas y abiertas en dos idiomas

La información disponible, como se ha visto en la sección 3.3, consta no sólo de variables categóricas, ya se consideren como activas o como suplementarias en un análisis factorial, sino también de frecuencias provenientes de una tabla léxica que recoge las respuestas de los individuos a la pregunta abierta

- *¿Podrías escribir aquí por qué?*

que aparece inmediatamente después de la pregunta cerrada *¿Estarías interesado en comprar un producto con el logotipo de la UPV/EHU ... ?* por lo cual debería estar íntimamente relacionada con ella.

Los datos correspondientes a los distintos tipos de variables tienen la particularidad de que pueden considerarse como medidos sobre los mismos individuos si estructuramos y yuxtaponemos las tablas de datos de manera conveniente (véase el esquema de la Figura 5.9).

<table>
<tr><th></th><th>Variables categóricas</th><th>Tabla léxica de frecuencias</th><th>Tabla léxica de frecuencias</th><th>Variables categóricas</th></tr>
<tr><td></td><td>(activas)</td><td>(Euskera)</td><td>(Castellano)</td><td>(suplementarias)</td></tr>
<tr><td>1
$\vdots$
I_1</td><td rowspan="2">z_{ikj}</td><td>f_{ij1}</td><td>0</td><td rowspan="2">z^{+}_{ikj}</td></tr>
<tr><td>I_1+1
$\vdots$
I_1+I_2</td><td>0</td><td>f_{ij2}</td></tr>
</table>

Figura 5.9: Encuesta on-line en dos idiomas: esquema de tablas de categorías y tablas léxicas yuxtapuestas considerando ambos idiomas.

En este esquema aparecen yuxtapuestas las tablas de variables categóricas correspondientes a las preguntas cerradas y a las variables de caracterización, igual que la sección anterior. Esto no es posible realizarlo para las tablas léxicas de frecuencias, puesto que al obtenerse en dos idiomas diferentes, los términos no son los mismos y es difícil realizar una equivalencia entre ellos. La solución propuesta es utilizar ambas tablas léxicas rellenando con ceros las frecuencias

correspondientes a las palabras que pertenecen al idioma no utilizado por el encuestado.

Siguiendo la notación de la sección 5.2.5, disponemos de $P = 2$ grupos de individuos correspondientes a los dos idiomas, Euskera y Castellano, de tamaño $I_1 = 304$ e $I_2 = 1243$, respectivamente. El grupo 1 corresponde a individuos con respuestas en Euskera y el 2 a las respuestas en Castellano. Sobre ellos tenemos una tabla ($J_c = 1$) de $Q_1 = 2$ variables categóricas, con un total de $K_1 = 5$ variables indicadoras correspondientes a las categorías de las variables `BuyLogo` y `Satis` (2 y 3, respectivamente). Adicionalmente, tenemos $J_f = 2$ tablas léxicas de frecuencias correspondientes a las tablas léxicas de frecuencias de las respuestas en Euskera y Castellano, respectivamente. Estas tablas se completan con ceros para los individuos que han respondido en el idioma alternativo al de la tabla. El número de columnas de cada tabla de frecuencias es de $K_2 = 223$ para Euskera y $K_3 = 247$ para Castellano.

El número de columnas de las tablas de frecuencias coinciden con el número de términos (formas y segmentos repetidos) seleccionados de las respuestas de los individuos a la pregunta abierta. Se ha llegado al número de términos elegido a partir de un umbral mínimo de frecuencia de 4 ocurrencias de las formas o segmentos repetidos en el caso del Euskera y 15 en el de Castellano. Asimismo, se han reagrupado algunos términos de raíz común. La formación del diccionario de palabras se resume en la Tabla 5.6.

	Individuos	Palabras	Palabras distintas	%	Frecuencia umbral	Formas y seg. final
Euskera	304	4756	1637	34,4	4	223
Castellano	1243	28944	3170	11	15	247

Tabla 5.6: Formación de los diccionarios de palabras, Euskera y Castellano.

Extracción de factores globales mediante un AFM mixto

Se realiza un Análisis Factorial Múltiple sobre las 3 tablas yuxtapuestas de variables activas de la Figura 5.9. El software de cálculo utilizado es SPAD. En los análisis, tanto parciales como global, interviene como activa la tabla de variables categóricas correspondiente a las respuestas a las preguntas cerradas. En el caso de las tablas activas de frecuencias, la versión de SPAD utilizada no tiene un procedimiento para introducir directamente una tabla de frecuencias en el AFM, por lo que se ha realizado un análisis de correspondencias simples sobre cada tabla de frecuencias y se han extraído sus factores. Los grupos correspondientes a las tablas de frecuencias se han formado incluyendo todos los factores de los AC respectivos (sin reducción de dimensionalidad alguna) para,

posteriormente, realizar un ACP no normado sobre ellos, lo que proporciona las mismas distancias que un AC clásico.

Como se ha indicado en la sección anterior, página 142, falta por especificar los pesos de los individuos. En esta aplicación, hemos optado por usar los pesos $f_{i..}$ de las tablas léxicas de frecuencias, lo que da más peso a aquéllos individuos de respuestas de mayor riqueza de vocabulario. Estos pesos son los mismos que se han utilizado en el ACM de la sección anterior, en que se apilaban las tablas de variables categóricas para los dos idiomas, con el objetivo de facilitar la comparación entre ambos.

La tabla 5.7 contiene los resultados de los análisis factoriales parciales o separados de cada una de las tres tablas disponibles. El principal interés radica justo en el primer valor propio λ_1 de cada una de las tablas, puesto que es precisamente el valor empleado en la ponderación $1/\lambda_1^j$ del AFM. Según dicha tabla, los valores propios son, respectivamente, $\lambda_1^1 = 0{,}6451, \lambda_1^2 = 0{,}1848, \lambda_1^3 = 0{,}1937$, por lo que la primera tabla, de variables categóricas, se verá afectada de un peso inferior al resto. Esto es debido a la mayor importancia de su primer factor, como consecuencia de un menor número de columnas de la matriz de variables indicadoras (6) en comparación a las columnas de la matriz de frecuencias (tal y como se han realizado los cálculos, igual al número de factores totales extraídos) de las tablas léxicas en ambos idiomas (222 y 246 respectivamente[4]).

Una vez introducidos los pesos $1/\lambda_1^j$ en las subtablas, se realiza el análisis global, obteniéndose los valores propios globales reflejados en la tabla 5.8. Se aprecia cómo los porcentajes de inercia son bastante reducidos, debido al gran número de columnas de las tablas léxicas y a que se trata de tablas muy dispersas (véase pág. 120). A este tipo de matrices se les denomina *sparse* en Inglés.

La tabla 5.9 contiene las ayudas a la interpretación de los ejes globales correspondientes a las subtablas. Examinando tanto las coordenadas como las contribuciones a esos ejes globales se concluye que el primer eje se debe principalmente a la inercia existente en la primera subtabla de variables categóricas, mientras que las dos subtablas léxicas son las responsables del segundo eje. Esto indica que, aún con la ponderación adicional del AFM, el análisis de la tabla total no puede recaer en un eje únicamente sino en, al menos, los dos primeros. Vistas, sobre todo, las coordenadas y las contribuciones, parece evidente que los ejes globales 3, 4 y 5 están asociados, de forma individual, a cada una de las 3 subtablas existentes (variables categóricas y frecuencias de textos en castellano y euskera, respectivamente) por lo que estarían recogiendo

[4]Uno menos que el número de unidades lexicales de cada tabla, debido a la supresión del factor trivial, en cada caso.

Primera subtabla: variables categóricas

Número	Valor propio	%	% acumulado
λ_1	0,6451	43,0066	43,0066
λ_2	0,5000	33,3334	76,3400
λ_3	0,3549	23,6601	100,0000

Segunda subtabla: frecuencias de términos en Euskera

Número	Valor propio	%	% acumulado
λ_1	0,1858	2,0932	2,0932
λ_2	0,1685	1,8980	3,9911
λ_3	0,1544	1,7390	5,7301

Tercera subtabla: frecuencias de términos en Castellano

Número	Valor propio	%	% acumulado
λ_1	0,1937	1,8407	1,8407
λ_2	0,1679	1,5956	3,4363
λ_3	0,1530	1,4542	4,8905

Tabla 5.7: Principales valores propios de los análisis parciales de las 3 subtablas activas.

la estructura interna de las mismas y escasa comunalidad inter grupos.

Las Tablas 5.10, 5.11 y 5.12 contienen, respectivamente, la relación entre los factores globales y los factores parciales de las tablas separadas. Estas tablas muestran como, cuando se proyectan factores parciales sobre los ejes globales, las coordenadas y los cosenos cuadrados son grandes ($-0{,}8709$ y $0{,}7585$ respectivamente) para el primer eje global, con respecto al primer eje parcial de la primera tabla, de variables categóricas. Esto indica que el primer eje global se determina principalmente por la inercia existente en la tabla que contiene las respuestas a las preguntas de Satisfacción y de propensión a comprar productos con el logotipo.

El segundo eje global tiene una mayor relación con los primeros ejes de las tablas léxicas. Las coordenadas son de $0{,}7176$ y $-0{,}6386$ para las subtablas de Euskera y Castellano respecivamente y los cosenos cuadrado de $0{,}5150$ y de $0{,}4079$, mucho más elevados que el resto. Los signos opuestos de las coordenadas indican que este eje recoge las diferencias entre los dos idiomas, estando Euskera en el lado positivo y Castellano en el negativo.

Mientras tanto, los ejes globales 3, 4 y 5 están netamente relacionados con los segundos ejes parciales de las subtablas, respectivamente, por lo que son específicos de las mismas.

Número	Valor propio	%	% acumulado
1	1,4923	1,4290	1,4290
2	1,0077	0,9650	2,3940
3	0,9401	0,9002	3,2942
4	0,9221	0,8830	4,1772
5	0,8711	0,8342	5,0114
6	0,8603	0,8239	5,8353
7	0,8327	0,7974	6,6326
8	0,8152	0,7806	7,4133
9	0,7837	0,7505	8,1638
10	0,7767	0,7438	8,9076

Tabla 5.8: Valores propios del AFM mixto (global) de la tabla de variables categóricas y las tablas léxicas.

Interpretación de los ejes globales

El primer eje está mucho más relacionado con las variables categóricas de interés que con los términos de las tablas léxicas. Sin embargo, la contribución de las tablas léxicas a este primer eje no es despreciable (24,31 y 22,26 respectivamente; su suma es casi igual a la de la primera tabla). Esto es particularmente importante, puesto que, como se verá a continuación, este eje de máxima inercia es el que más fielmente relaciona ambos tipos de tablas en este caso; en concreto, relaciona la satisfacción y la propensión a comprar con el vocabulario empleado.

La Tabla 5.13 contiene los valores test correspondientes a las proyecciones de los centros de gravedad de las categorías activas (primera tabla sobre ejes globales y parciales). Las tablas correspondientes a coordenadas, contribuciones y cosenos cuadrado se incluyen en el apéndice B.

En esta tabla, se ve que las categorías más importantes corresponden, en el primer eje, a *Compraría* `(BuyLo=1)`, *No compraría* `(BuyLo=2)` (con signos opuestos en el primer eje) y *Nada, poco o medianamente satisfecho* `(Satis2=1)`, *Muy satisfecho* `(Satis2=3)` (con signos también opuestos). La satisfacción queda del mismo lado que la intención de comprar en este primer eje. Esto ocurre tanto desde el punto de vista global como del parcial, tabla por tabla.

Respecto al segundo eje, se observa como las categorías *Bastante satisfecho* `(Satis2=2)` y *Muy satisfecho* `(Satis2=3)` tienen valores test elevados y se proyectan en lados opuestos del eje. Esto está asociado principalmente a diferencias en la primera tabla (valores test de 12,9481 y −11,8673), más que a diferencias en las tablas léxicas, donde aparecen algunas otras categorías sig-

Tablas activas		eje 1	eje 2	eje 3	eje 4	eje 5
Tabla	Dist. O.	Coordenadas				
1	1,9034	0,7973	0,0661	0,5941	0,0136	0,0862
2	20,1651	0,3628	0,5301	0,2660	0,7806	0,0939
3	18,3991	0,3322	0,4115	0,0800	0,1279	0,6910
		Contribuciones				
1		53,4291	6,5556	63,1939	1,4793	9,8956
2		24,3101	52,6041	28,2954	84,6529	10,7772
3		22,2608	40,8403	8,5108	13,8678	79,3274
Todas		1,0000	1,0000	1,0000	1,0000	1,0000
		Cosenos cuadrado				
1		0,3340	0,0023	0,1854	0,0001	0,0039
2		0,0065	0,0139	0,0035	0,0302	0,0004
3		0,0060	0,0092	0,0003	0,0009	0,0260
Todas		0,8777	0,4547	0,4301	0,6258	0,4938

Tabla 5.9: Ayudas a la interpretación de los ejes globales (Coordenadas, contribuciones y cosenos cuadrado de las subtablas).

nificativas, pero que ya lo eran en el primer eje. Dados los signos de los valores test para este segundo eje de las dos tablas léxicas (grupos 2 y 3) parece que estén cambiados (hay que recordar que estas tablas están representadas por sus factores, de signo y orientación arbitrarios). Esto simplemente significa que en la parte negativa estarían los *Muy satisfecho* que *comprarían* del grupo 2 (Euskera) mientras que en la parte positiva los que *no comprarían*; el resto de categorías en ese grupo no son significativas. En la parte positiva de ese eje, estarían asimismo los integrantes del grupo 3 (castellano) que *comprarían* y en la negativa los que *no comprarían*; la satisfacción no es significativa en este grupo en el análisis global.

El tercer eje es un eje que diferencia entre encuestados *Bastante satisfechos*, en el lado positivo, y el resto de categorías que miden satisfacción (*Nada, poco o medianamente satisfecho* y *Muy satisfecho*), en una configuración de herradura que recuerda a un efecto Guttman. El lado negativo está asociado a estudiantes y jóvenes del grupo 2 (Euskera) mientras que el positivo a PDI y del grupo de edad 3 de ese mismo grupo. El lado positivo también está del lado de los que *comprarían* del grupo 3 (castellano) mientras que el negativo de los que no lo harían en ese grupo. En realidad, mirando a los valores test, este eje se parece bastante al eje 2, pero con valores test mayores en valor absoluto.

El examen de los valores test correspondientes a las variables categóricas suplementarias (Tablas 5.14 y 5.15) permite apreciar, en el primer eje, que las

Tablas	Eje Parcial	Pesos	eje 1	eje 2	eje 3	eje 4	eje 5
Var.	1	1,0000	-0,8709	0,0020	-0,1417	-0,1082	-0,1256
Categ.	2	0,7751	0,0620	-0,2824	-0,8516	-0,0482	0,0582
Activas	3	0,5502	-0,2552	-0,0878	0,1467	0,0156	0,3510
(ACM)	4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Tabla	1	1,0000	-0,4665	0,7176	-0,1518	0,2968	0,0670
Léxica	2	0,9067	0,2319	-0,0066	-0,1116	0,8666	0,0128
Eusk.	3	0,8308	0,0625	-0,0411	-0,2243	0,0263	-0,0087
(ACPnn)	4	0,7921	0,1240	0,0303	-0,0419	0,0487	-0,2773
	5	0,7787	-0,0509	0,0161	0,0921	-0,0168	0,0228
Tabla	1	1,0000	-0,5302	-0,6386	0,2090	0,3453	-0,0104
Léxica	2	0,8668	0,0542	0,0138	0,0200	0,0573	-0,8808
Cast.	3	0,7900	-0,0666	0,0184	0,0153	-0,0403	-0,1118
(ACPnn)	4	0,7121	0,1412	-0,0062	-0,0083	0,0536	0,0252
	5	0,6479	0,0189	0,0278	0,0711	0,0080	-0,0431

Tabla 5.10: Ayudas a la interpretación de la proyección los ejes parciales sobre los ejes globales: Coordenadas.

categorías con valor test estadísticamente significativo son las correspondientes a las clases de *Estudiantes, Edad entre 23 y 29 años* que se proyectan en el lado izquierdo (*No compraría, Satisfacción media/baja*) y a las de *PDI, Edad mayor de 45 años*, proyectadas sobre el lado derecho (junto con *Compraría* y *Satisfacción muy alta*).

Con respecto al segundo eje, no hay ninguna categoría suplementaria estadísticamente significativa. Las tablas correspondientes a coordenadas, contribuciones y cosenos cuadrado de las categorías suplementarias están disponibles en el apéndice B.

En definitiva, parece que los ejes globales del AFM tienen mucho que ver con los ejes extraídos del ACM de la tabla de variables categóricas en la que se apilan las dos subtablas correspondientes a las respuestas en los dos idiomas de la sección 5.3.2. El primer eje es un eje de valoración (positiva-negativa) de la satisfacción sobre la universidad, asociada a la probabilidad de compra de productos corporativos y relacionada principalmente con la edad y la vinculación que el encuestado tiene con la universidad. El segundo eje, de importancia relativamente menor, es un eje que distingue entre satisfacción elevada y muy elevada, sin apenas mayores elementos de distinción. Los ejes posteriores son más particulares o más difíciles de interpretar. Por ejemplo, el tercer eje parece estar relacionado únicamente con la satisfacción, el cuarto no está significativa-

Tablas	Eje Parcial	eje 1	eje 2	eje 3	eje 4	eje 5
Var.	1	50,8288	0,0004	2,1370	1,2697	1,8123
Categ.	2	0,1995	6,1348	59,7975	0,1951	0,3011
Activas	3	2,4007	0,4205	1,2594	0,0145	7,7823
(ACM)	4	0,0000	0,0000	0,0000	0,0000	0,0000
	10 ejes	53,4291	6,5556	63,1940	1,4793	9,8956
Tabla	1	14,5838	51,1021	2,4506	9,5557	0,5155
Léxica	2	3,2672	0,0039	1,2021	73,8532	0,0171
Eusk.	3	0,2178	0,1395	4,4446	0,0623	0,0073
(ACPnn)	4	0,8162	0,0723	0,1482	0,2039	6,9936
	5	0,1350	0,0199	0,7027	0,0239	0,0463
	10 ejes	20,8001	51,5956	14,4684	84,1190	8,9137
Tabla	1	18,8376	40,4751	4,6470	12,9277	0,0124
Léxica	2	0,1709	0,0163	0,0370	0,3092	77,2088
Cast.	3	0,2347	0,0266	0,0198	0,1391	1,1332
(ACPnn)	4	0,9514	0,0028	0,0053	0,2219	0,0519
	5	0,0155	0,0498	0,3485	0,0045	0,1381
	10 ejes	20,5678	40,5861	5,5831	13,6795	78,7263
	Contrib. acumulada	94,7970	98,7374	83,2454	99,2778	97,5356

Tabla 5.11: Ayudas a la interpretación de la proyección los ejes parciales sobre los ejes globales. Contribuciones.

mente relacionado con ninguna categoría a nivel global (salvo con la categoría de Edad=3) y el quinto nuevamente con la satisfacción a nivel global y con alguna categoría de algún grupo concreto (comprar/no comprar en el grupo 2), lo que aparentemente ya ha sido recogido en ejes anteriores.

Tablas	Eje Parcial	eje 1	eje 2	eje 3	eje 4	eje 5
Var.	1	0,7585	0,0000	0,0201	0,0117	0,0158
Categ.	2	0,0038	0,0798	0,7253	0,0023	0,0034
Activas	3	0,0651	0,0077	0,0215	0,0002	0,1232
(ACM)	4	0,0000	0,0000	0,0000	0,0000	0,0000
	10 ejes	0,3429	0,0284	0,2555	0,0059	0,0371
Tabla	1	0,2176	0,5150	0,0230	0,0881	0,0045
Léxica	2	0,0538	0,0000	0,0125	0,7510	0,0002
Eusk.	3	0,0039	0,0017	0,0503	0,0007	0,0001
(ACPnn)	4	0,0154	0,0009	0,0018	0,0024	0,0769
	5	0,0026	0,0003	0,0085	0,0003	0,0005
	10 ejes	0,0065	0,0109	0,0028	0,0162	0,0016
Tabla	1	0,2811	0,4079	0,0437	0,1192	0,0001
Léxica	2	0,0029	0,0002	0,0004	0,0033	0,7759
Cast.	3	0,0044	0,0003	0,0002	0,0016	0,0125
(ACPnn)	4	0,0199	0,0000	0,0001	0,0029	0,0006
	5	0,0004	0,0008	0,0051	0,0001	0,0019
	10 ejes	0,0056	0,0075	0,0010	0,0023	0,0126

Tabla 5.12: Ayudas a la interpretación de la proyección los ejes parciales sobre los ejes globales. Cosenos cuadrado.

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
BuyLogo		15,9320	0,7112	0,3708	1,0896	1,5213
	BuyLo=1	11,9882	0,7138	0,3723	1,0926	-1,5233
	Grupo 1	12,4796	1,7071	1,3616	10,0315	-0,8474
	Grupo 2	6,4901	-3,6878	0,4065	1,4098	-2,5342
	Grupo 3	5,7575	4,6936	-3,0577	-3,7687	-0,6008
	BuyLo=2	-11,9882	-0,7138	-0,3723	-1,0926	1,5233
	Grupo 1	-12,4796	-1,7071	-1,3616	-10,0315	0,8474
	Grupo 2	-6,4901	3,6878	-0,4065	-1,4098	2,5342
	Grupo 3	-5,7575	-4,6936	3,0577	3,7687	0,6008
Satis2		8,0624	3,5335	16,6482	0,4793	3,9886
	Satis2 = 1	-7,1848	-0,3447	-5,9686	-1,4873	-3,9810
	Grupo 1	-9,2417	-1,8816	-6,9224	-12,1040	-11,7379
	Grupo 2	-2,5974	1,3218	-2,3647	-0,3791	-1,9312
	Grupo 3	-1,9625	-1,3776	0,2917	0,9548	-0,2667
	Satis2 = 2	0,9120	3,6975	12,3196	0,9699	0,2003
	Grupo 1	1,4976	12,9481	14,0718	7,0384	0,3277
	Grupo 2	-0,2057	0,8864	4,0169	0,0952	0,4326
	Grupo 3	0,2764	0,2850	1,4690	0,0170	-0,0206
	Satis2 = 3	5,5406	-3,5629	-7,5133	0,3284	3,3896
	Grupo 1	6,7867	-11,8673	-8,4871	3,5685	10,2697
	Grupo 2	2,5640	-2,1240	-2,0713	0,2430	1,2928
	Grupo 3	1,4848	0,9469	-1,8031	-0,8811	0,2628

Tabla 5.13: Valores test de los centros de gravedad de las categorías activas.

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Vinculación		1,6205	-2,0518	1,1942	1,1305	-0,8653
	Estudiantes	-2,1719	-0,0784	-1,8693	1,5282	-0,5510
	Grupo 1	-2,2198	-0,3941	-1,0295	-2,4029	-1,4744
	Grupo 2	-0,5842	1,0674	-2,9183	1,6548	-0,0579
	Grupo 3	-1,7077	-1,1917	0,6005	0,8126	-0,1538
	PAS	-0,1908	-0,1526	-0,1405	0,6683	-0,1375
	Grupo 1	0,1959	-0,1329	-0,2806	-0,0226	-0,3195
	Grupo 2	-0,2813	0,6455	-0,2391	0,3002	-0,6691
	Grupo 3	-0,5376	-0,9233	0,6472	1,0320	0,1695
	PDI	2,2684	0,1941	1,9341	-2,0072	0,6440
	Grupo 1	2,0175	0,4873	1,2216	2,3646	1,6861
	Grupo 2	0,7871	-1,5396	3,0347	-1,8476	0,5716
	Grupo 3	2,0819	1,8748	-1,0848	-1,5882	0,0198
Género		0,2749	0,6684	0,7427	0,0990	0,4110
	Masc.	0,2760	0,6709	0,7454	-0,0995	0,4127
	Grupo 1	0,0161	1,2762	0,8411	0,1343	-1,2563
	Grupo 2	0,0196	-0,2427	0,6209	0,2966	-0,1986
	Grupo 3	0,6946	0,8838	-0,5233	-1,0099	0,9159
	Fem.	-0,2760	-0,6709	-0,7454	0,0995	-0,4127
	Grupo 1	-0,0161	-1,2762	-0,8411	-0,1343	1,2563
	Grupo 2	-0,0196	0,2427	-0,6209	-0,2966	0,1986
	Grupo 3	-0,6946	-0,8838	0,5233	1,0099	-0,9159

Tabla 5.14: Valores test de los centros de gravedad de las categorías suplementarias (`vinculación, género`).

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Prov		0,5510	-0,0054	-1,5891	-0,3058	-2,1630
	Araba	-0,8520	-0,0436	-0,0641	-0,5011	0,1800
	Grupo 1	-1,0364	0,1744	0,0211	-0,8652	-0,4159
	Grupo 2	-0,3900	0,0172	-0,3318	-0,4594	0,1147
	Grupo 3	-0,2440	-0,1491	0,2560	0,0709	0,2908
	Bizkaia	1,5658	-0,9925	0,3274	-0,4761	-0,0594
	Grupo 1	1,3623	0,0963	0,0622	1,0638	-0,0675
	Grupo 2	1,6870	-1,7235	0,8573	-0,5203	-0,1509
	Grupo 3	0,3268	0,3512	-0,3617	-0,3440	0,0047
	Gipuzkoa	-1,0159	1,1351	-0,3084	0,9486	-0,0857
	Grupo 1	-0,6354	-0,2533	-0,0866	-0,4492	0,4248
	Grupo 2	-1,5387	1,8930	-0,6695	0,9624	0,0705
	Grupo 3	-0,1564	-0,2632	0,1848	0,3210	-0,2500
Edad		1,4910	-1,2175	0,2143	1,0256	-0,8894
	Edad=1	-1,0170	0,2041	-1,5782	1,0974	-0,2995
	Grupo 1	-0,7511	-0,6097	-0,8223	-0,9927	-0,4066
	Grupo 2	-0,5756	1,0594	-2,5337	1,1193	0,1955
	Grupo 3	-0,9512	-0,6623	0,4943	0,5106	-0,2810
	Edad=2	-1,9017	-0,1890	-0,1443	1,5125	-0,7000
	Grupo 1	-2,4230	0,7823	-0,0415	-2,2699	-2,1365
	Grupo 2	0,1091	0,3769	-0,5658	1,6636	-0,3209
	Grupo 3	-1,3630	-0,9976	0,5176	0,7078	-0,0312
	Edad=3	0,6380	-0,6088	1,1492	-1,9811	0,8787
	Grupo 1	0,7812	-0,2572	0,4397	1,0915	1,7058
	Grupo 2	0,0433	-1,0238	2,0761	-2,2100	0,3926
	Grupo 3	0,4268	0,2885	-0,3073	-0,2429	0,3397
	Edad=4	1,9671	0,6751	0,3684	-0,1314	-0,1036
	Grupo 1	1,9970	0,2160	0,3478	1,7404	0,3156
	Grupo 2	0,4297	-0,2015	0,6187	-0,0202	-0,3694
	Grupo 3	1,6685	1,2152	-0,5974	-0,8583	-0,0875

Tabla 5.15: Valores test de los centros de gravedad de las categorías suplementarias (`campus, edad`).

5.3.4. Clasificación sobre los factores principales del AFM de tablas mixtas

Una vez realizada la extracción de factores globales de la tabla completa por el método de AFM aplicado a tablas mixtas, se procede a la clasificación sobre los mismos y a la posterior caracterización de la partición obtenida.

La clasificación se realiza sobre los 3 primeros factores del AFM. Hay dos razones para seleccionar este número de factores. La primera tiene que ver con el pequeño número de variables que existen en la subtabla primera del AFM. Dado que esta tabla sólo contiene 2 variables de 2 y 3 categorías respectivamente, del ACM de esta tabla sólo pueden extraerse 3 factores parciales. Además, el hecho de seleccionar tres factores favorece la comparabilidad con el ACM realizado sobre las tablas apiladas de la sección 5.3.2.

Se realiza una clasificación mixta combinando un algoritmo de medias móviles con 10 centros iniciales, una clasificación jerárquica por el método de Ward generalizado y una consolidación de la partición nuevamente mediante el algoritmo de medias móviles. La clasificación jerárquica da lugar al histograma de índices de nivel de la Figura 5.10, que sugiere una partición en 5 clases.

```
Núm.  Prim. Últ. Efec.    Peso    Índice    Histograma de índices de nivel
208   203   205    21     46,29   0,14733   ****
209   204   192     7     31,35   0,15509   ****
210   207   200    54     67,36   0,25982   ************
211   210   209    61     98,71   0,28640   **************
212   208   206    46    100,49   0,45856   ***************************
213   211   212   107    199,20   1,04693   ******************...*************
Suma de índices de nivel =    3,20679
```

Figura 5.10: Histograma de índices de nivel de la clasificación sobre 3 factores del AFM de tablas mixtas. Método generalizado de Ward.

La partición sugerida proporciona una clase más que la partición seleccionada en el ACM de tablas apiladas de la sección 5.3.2. En la figura 5.11 se incluyen, sobre el plano principal del AFM de tablas mixtas, las proyecciones de los centros de las clases de las particiones en 4 y 5 clases. En esta figura se observa cómo la partición en 5 clases (centros representados por cuadrados) en realidad no supone más que una subdivisión de la clase 4 correspondiente a la partición en 4 clases (centros representados por círculos), en dos clases (4 y 5). Dichas clases se sitúan en el mismo cuadrante del plano principal y tienen una interpretación similar, salvo porque una es más extrema que la otra. Por esta razón, se elige como partición final la de 4 clases. Esta partición resulta, además, similar a la obtenida en la sección 5.3.2.

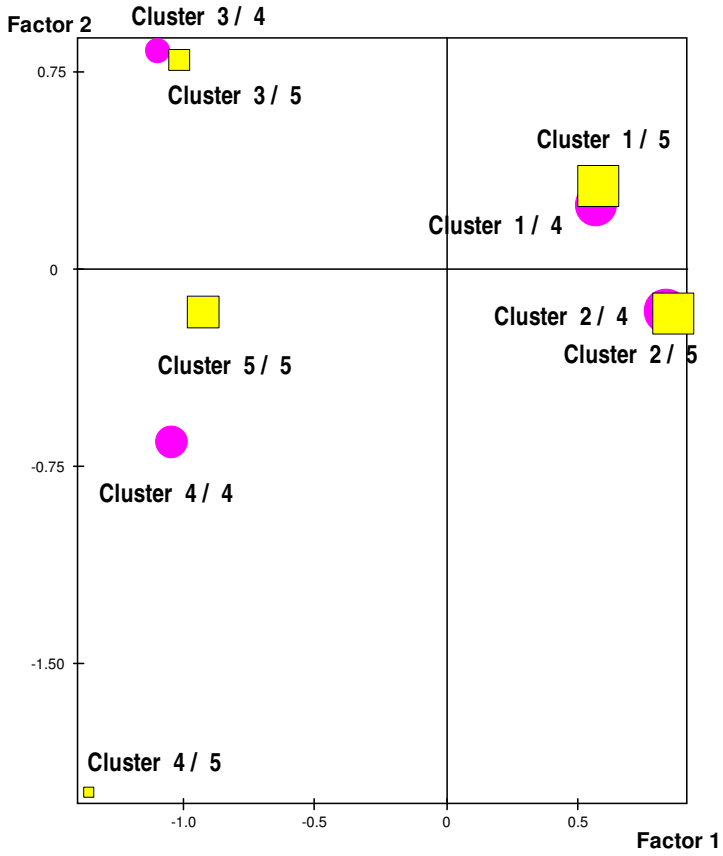


Figura 5.11: Proyección sobre el plano principal del AFM de las particiones en 4 clases (círculos) y 5 clases (cuadrados).

Los resultados correspondientes a la consolidación de la partición en 4 clases pueden consultarse en el apéndice B.2.2 (pág. 238). A continuación se procede a la caracterización de la partición escogida. En este caso, al tratarse de un análisis de tablas múltiples mixtas, la caracterización se realiza respecto de variables de diversa naturaleza, tanta como tipos de variables hay en las subtablas objeto de análisis en el AFM.

Primer cluster

El centro del primer cluster (o cluster número 1) se sitúa en el primer cuadrante del plano principal del AFM. Teniendo en cuenta la interpretación hecha de los ejes del AFM, en este área se situarían individuos propensos a comprar y bastante satisfechos con la institución, de mayor edad y personal docente e investigador. En este momento no sabemos nada de los términos

empleados por este colectivo en la pregunta abierta. La caracterización de dichos términos se realiza como cualquier otra categoría, mediante valores test que comparan la proporción del uso de un determinado término con respecto al total de la muestra.

Las tablas 5.16, 5.17 y 5.18 contienen, respectivamente, las categorías y términos en castellano y en euskera característicos del primer cluster. Si el valor test es positivo, indica que esa categoría o término se emplean significativamente más que en la muestra total y si es negativo, que se usan significativamente menos que en la muestra total.

El primer cluster se caracteriza por encuestados bastante satisfechos con la universidad `(Satis=2)`, que comprarían productos con su logotipo `(BuyLo=1)` siendo miembros del PDI (personal docente e investigador).

En cuanto a los términos empleados, son característicos, en Castellano, *que pertenece, bonito, de nuestra, calidad, orgullo, ayuda, conocer, buena, original, ...* entre otros. Estos términos evocan ideas positivas bien hacia la universidad o bien hacia productos. Por contra, se utilizan con pequeña frecuencia términos negativos como *no* o *no me gusta*.

En Euskera son característicos términos como *lan (trabajo), artikulu (artículo), aukera (oportunidad), korporatiboa (corporativo), indartzen (hacer más fuerte, mejorar), irudia (aspecto), ezaguna (conocido), instituzio (institución), polita (bonito), garrantzizko (de importancia), ...*, nuevamente términos en clave positiva hacia la universidad o hacia productos. Aparecen con menos frecuencia que el texto términos como *ez (no)* y *ez dut uste (no creo)*. Merece ser destacado el resultado de que las respuestas a las preguntas cerradas se corresponden en gran medida con los términos empleados y eso, aparentemente, para los dos idiomas.

Cluster 1/4 (Efectivo: 59 - Porcentaje: 29.63)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
Satis2 = 2	97,39	42,36	68,12	10,56	84
BuyLo=1	92,17	59,63	45,80	6,17	119
PDI	56,52	44,29	37,81	2,00	88
Male	50,88	44,45	33,92	0,97	89
Age=3	44,09	38,42	34,00	0,85	77
Bizkaia	59,59	54,44	32,43	0,77	108
Age=4	24,64	22,01	33,16	0,55	44
ARABA	16,27	17,35	27,77	-0,07	35
Age=2	15,27	16,34	27,69	-0,10	33
PAS	15,83	17,79	26,36	-0,34	35
Gipuzkoa	24,14	28,21	25,36	-0,72	56
Female	49,12	55,55	26,20	-1,07	111
Age=1	15,69	22,13	21,00	-1,34	44
Students	27,65	37,93	21,60	-1,95	76
Satis2 = 1	0,24	23,35	0,30	-5,77	47
BuyLo=2	7,00	39,35	5,27	-6,39	78
Satis2 = 3	1,32	32,94	1,19	-6,75	66

Tabla 5.16: Caracterización del cluster 1 por las categorías de las preguntas cerradas.

Cluster 1/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
que pertenece	0,48	0,19	16	23	3,975
bonito	1,11	0,62	37	75	3,888
depende	0,48	0,20	16	24	3,790
de nuestra	0,36	0,13	12	16	3,690
calidad	0,87	0,47	29	57	3,596
siempre	0,72	0,40	24	48	3,154
sobre todo	0,36	0,16	12	19	3,012
manera	0,72	0,41	24	50	2,936
fuera	0,63	0,35	21	42	2,932
orgullo	0,45	0,22	15	27	2,874
forma	1,11	0,73	37	88	2,819
ayuda	0,42	0,21	14	25	2,797
conocer	1,35	0,92	45	112	2,793
diseno	0,93	0,59	31	71	2,792
formacion	0,36	0,17	12	21	2,638
siempre que	0,36	0,17	12	21	2,632
ser	7,62	6,66	255	807	2,553
buena	1,35	0,96	45	116	2,531
precio	0,75	0,48	25	58	2,412
original	0,27	0,12	9	15	2,386
me gustan los productos	0,00	0,12	0	15	-2,414
regalar	0,90	1,30	30	158	-2,423
pagar	0,03	0,18	1	22	-2,426
en general	0,09	0,28	3	34	-2,442
publicidad	0,30	0,59	10	71	-2,556
general	0,12	0,34	4	41	-2,575
nada	0,09	0,31	3	37	-2,702
hacer publicidad	0,12	0,37	4	45	-2,877
con logotipos	0,03	0,24	1	29	-3,079
ningun tipo	0,00	0,18	0	22	-3,147
tipo	0,42	0,83	14	101	-3,173
cosas	0,24	0,61	8	74	-3,365
ningun	0,15	0,49	5	59	-3,465
no se	0,03	0,29	1	35	-3,570
gusta	1,26	1,97	42	239	-3,575
me gusta	0,90	1,57	30	190	-3,724
no me parece	0,03	0,31	1	38	-3,796
propaganda	0,09	0,46	3	56	-4,068
no me gusta	0,03	0,79	1	96	-6,986
no	3,35	6,08	112	737	-8,177

Tabla 5.17: Caracterización del cluster 1 por palabras y segmentos repetidos característicos de los encuestados en Castellano.

Cluster 1/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
lan	1,15	0,51	12	18	3,048
artikulu	0,57	0,20	6	7	2,709
aukera	0,76	0,31	8	11	2,678
zergaitik	1,34	0,70	14	25	2,589
on	1,24	0,65	13	23	2,518
korporatiboa	0,38	0,11	4	4	2,434
indartzen	0,38	0,11	4	4	2,434
irudia	0,76	0,34	8	12	2,397
nola	0,48	0,17	5	6	2,329
ezaguna	0,95	0,48	10	17	2,296
eman	0,95	0,48	10	17	2,296
dago	5,44	4,18	57	149	2,281
instituzio	0,57	0,25	6	9	1,998
lan egiten	0,38	0,14	4	5	1,898
ere	1,91	1,29	20	46	1,896
polita	1,15	0,70	12	25	1,775
garrantzizko	0,38	0,17	4	6	1,507
batez ere	0,38	0,17	4	6	1,504
beste	1,81	1,32	19	47	1,483
gonbidatu	0,29	0,11	3	4	1,410
iruditu	0,67	0,98	7	35	-1,047
uste	0,67	0,98	7	35	-1,047
ehu	2,48	3,00	26	107	-1,078
produktu	2,29	2,81	24	100	-1,101
euskal	0,00	0,20	0	7	-1,360
esan	0,19	0,48	2	17	-1,377
edukitzea	6,11	7,13	64	254	-1,475
agian	0,00	0,22	0	8	-1,544
logo	2,10	2,81	22	100	-1,566
propaganda	0,38	0,79	4	28	-1,613
egon	0,57	1,04	6	37	-1,642
ikasi	0,57	1,04	6	37	-1,642
ez dut uste	0,10	0,42	1	15	-1,759
erabil	0,76	1,38	8	49	-1,940
eraman	0,67	1,26	7	45	-1,969
zerbait	0,19	0,62	2	22	-1,991
ikusi	0,19	0,73	2	26	-2,420
gogoko	1,05	2,25	11	80	-3,175
ez zait	0,29	1,15	3	41	-3,232
ez	5,34	7,64	56	272	-3,355

Tabla 5.18: Caracterización del cluster 1 por palabras y segmentos repetidos característicos de los encuestados en Euskera.

Segundo cluster

El centro del segundo cluster se sitúa en el cuarto cuadrante del plano principal del AFM (Ver Figura 5.11). Al igual que el primer cluster, se sitúa en el lado correspondiente a los individuos más satisfechos y más propensos a comprar productos con el logotipo de la universidad. En efecto, las categorías características de esta clase corresponden (ver Tabla 5.19) a las categorías Muy satisfecho (`Satis2=3`) y Compraría (`BuyLo=1`), de acuerdo con los valores test de este cluster. Ésta sería la clase óptima en cuanto a sus características, ligeramente por encima de la anterior.

Con respecto a las palabras o segmentos empleados por los que han respondido en Castellano, algunos de los más característicos son (Tabla 5.20): *pertenecer*, *a la upv*, *para regalar*, *orgulloso*, *profesores*, *visitar*, *amigo*, *uso*, *regalo*, ... y los menos características están asociados a la palabra *no* (*no compraría*, *no me interesa*, ...).

En el caso del Euskera, son más representativos términos como *ikasi (aprender)*, *arro (orgulloso)*, *bertako (de aqui, expresado con orgullo o simpatía)*, *eskaini (ofrecer)*, *behar (necesitar)*, *zergatik ez (¿por qué no?)*, *lagun (amigo)*, ... y menos representativos que en el total de los términos empleados de nuevo términos asociados al no (ez) como *ez zait (no me...)*, *ez dut uste (no creo)*, *ezer (nada)*, *interesgarria (interesante)*, ...

También en este segundo cluster las respuestas positivas a las preguntas cerradas están asociadas a un vocabulario de connotaciones positivas respecto a la universidad y su productos corporativos.

Cluster 2/4 (Efectivo: 61 - Porcentaje: 30,51)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
Satis2 = 3	79,30	32,94	73,45	8,92	66
BuyLo=1	97,25	59,63	49,76	7,61	119
Age=4	28,58	22,01	39,61	1,11	44
Female	60,07	55,55	32,99	0,77	111
Bizkaia	59,31	54,44	33,24	0,74	108
Age=1	23,86	22,13	32,89	0,39	44
PAS	20,55	17,79	35,25	0,33	35
Gipuzkoa	28,43	28,21	30,75	0,22	56
Age=3	37,31	38,42	29,63	-0,03	77
PDI	43,81	44,29	30,19	-0,16	88
Students	35,64	37,93	28,67	-0,25	76
Male	39,93	44,45	27,41	-0,86	89
Satis2 = 1	18,67	23,35	24,40	-1,05	47
Araba	12,26	17,35	21,55	-1,32	35
Age=2	8,72	16,34	16,29	-1,97	33
BuyLo=2	1,94	39,35	1,51	-7,91	78
Satis2 = 2	1,14	42,36	0,82	-8,41	84

Tabla 5.19: Caracterización del cluster 2 por las categorías de las preguntas cerradas.

Cluster 2/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
pertenecer	1,95	1,02	69	124	6,070
de pertenecer	1,10	0,50	39	61	5,465
a la upv	1,36	0,69	48	84	5,191
pertenecer a la upv	0,73	0,30	26	36	5,152
para regalar	1,50	0,86	53	104	4,524
orgulloso	1,02	0,52	36	63	4,499
profesores	0,65	0,31	23	37	3,998
visitar	0,37	0,14	13	17	3,788
upv	5,28	4,19	187	507	3,742
estar	1,64	1,06	58	129	3,714
amigo	0,42	0,18	15	22	3,573
tambien	0,76	0,42	27	51	3,413
estar orgullosos	0,42	0,19	15	23	3,366
visitante	0,40	0,17	14	21	3,341
ehu	1,61	1,09	57	132	3,336
como regalo	0,34	0,15	12	18	3,055
dar	1,41	0,97	50	117	3,029
uso	0,73	0,43	26	52	3,018
tanto	0,51	0,27	18	33	2,874
regalo	1,47	1,03	52	125	2,874
ningun tipo	0,03	0,18	1	22	-2,573
ningun	0,23	0,49	8	59	-2,661
con logotipos	0,06	0,24	2	29	-2,682
depende	0,03	0,20	1	24	-2,783
me gusta	1,07	1,57	38	190	-2,785
parecer	1,24	1,77	44	215	-2,861
no es	0,06	0,26	2	31	-2,866
nada	0,08	0,31	3	37	-2,902
no me parece	0,08	0,31	3	38	-2,977
tampoco	0,00	0,16	0	19	-2,991
debe	0,25	0,58	9	70	-3,086
no compraria	0,00	0,17	0	20	-3,091
marca	0,40	0,79	14	96	-3,237
pagar	0,00	0,18	0	22	-3,295
dinero	0,03	0,25	1	30	-3,344
me interesa	0,00	0,19	0	23	-3,386
interesante	0,37	0,79	13	96	-3,499
propaganda	0,11	0,46	4	56	-3,898
no me gusta	0,11	0,79	4	96	-6,124
no	2,99	6,08	106	737	-9,700

Tabla 5.20: Caracterización del cluster 2 por palabras y segmentos repetidos característicos de los encuestados en Castellano.

Cluster 2/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
ikasi	2,39	1,04	19	37	3,722
arro	0,63	0,14	5	5	3,262
euskal	0,75	0,20	6	7	3,196
publikoa	0,63	0,17	5	6	2,782
egon	2,01	1,04	16	37	2,691
bat	3,89	2,61	31	93	2,361
bertako	0,88	0,37	7	13	2,226
unibertsitatea	3,02	1,97	24	70	2,191
eskaini	0,63	0,22	5	8	2,125
herriko	0,63	0,22	5	8	2,125
behar	2,14	1,32	17	47	2,030
zergatik ez	1,01	0,48	8	17	2,021
lagun	0,38	0,11	3	4	1,785
bizi	0,38	0,11	3	4	1,785
hemen	0,38	0,11	3	4	1,785
zure	0,38	0,11	3	4	1,785
zerbait	1,13	0,62	9	22	1,761
arrazoi	0,50	0,20	4	7	1,659
alde	0,75	0,37	6	13	1,652
harro	0,63	0,28	5	10	1,638
gainera	0,13	0,39	1	14	-1,058
egokia	0,13	0,39	1	14	-1,058
agian	0,00	0,22	0	8	-1,118
instituzio	0,00	0,25	0	9	-1,269
zalea	0,00	0,25	0	9	-1,269
etortzen	0,00	0,25	0	9	-1,269
ezaguna	0,13	0,48	1	17	-1,409
saldu	0,00	0,28	0	10	-1,409
gauza	0,38	0,84	3	30	-1,467
handi	0,00	0,34	0	12	-1,667
ezer	0,00	0,34	0	12	-1,667
erosi	1,88	2,81	15	100	-1,715
uste	0,38	0,98	3	35	-1,876
ez dut uste	0,00	0,42	0	15	-2,001
horrela	0,13	0,65	1	23	-2,008
gogoko	1,26	2,25	10	80	-2,098
interesgarria	0,00	0,53	0	19	-2,406
ez zait	0,25	1,15	2	41	-2,790
logo	1,38	2,81	11	100	-2,815
ez	4,40	7,64	35	272	-4,040

Tabla 5.21: Caracterización del cluster 2 por palabras y segmentos repetidos característicos de los encuestados en Euskera.

Tercer cluster

El tercer cluster se proyecta sobre el segundo cuadrante del plano principal y, a diferencia de los anteriores, está en el lado negativo del primer eje global, lo que implica estar asociado a una menor satisfacción con respecto a la universidad y una menor disposición a la compra de productos con el logotipo de la universidad.

Las categorías características de este cluster son (Tabla 5.22) *no compraría* y *bastante satisfecho*. Esto último se debe a que esta clase se sitúa en la parte positiva del segundo eje.

Los términos más característicos de esta clase son, en Castellano, *no me gusta, no, con logotipos, propaganda, ningún, logotipo, objetos, hacer publicidad, ...* términos principalmente orientados a un rechazo a llevar productos con el anagrama de la universidad, más por un tema de gustos (se sugiere que no gusta hacer publicidad de la universidad) que por un problema de insatisfacción con la universidad.

En el caso del idioma Euskera, se usan significativamente los términos *zalea (persona aficionada a algo), ez (no), asko (mucho), propaganda, logo, ikusi (ver), batere (ningún), atsegin (agradable), diseinua (diseño), ez zait (no me..), gutxi (poco), kontsumismoa (consumismo), ...* que son términos que se pueden asociar preferentemente a los productos y en sentido negativo, al igual que en el caso de las respuestas en Castellano.

Este cluster representa, por tanto, un grupo de encuestados a los que no le gusta en absoluto la idea de llevar productos que supongan algún tipo de *propaganda* (al menos, la que estaría dirigida a la universidad, o a esta universidad en concreto).

Cluster 3/4 (Efectivo: 32 - Porcentaje: 16,23)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
BuyLo=2	94,19	39,35	38,84	6,90	78
Satis2 = 2	67,74	42,36	25,95	3,11	84
Students	48,04	37,93	20,56	1,30	76
Gipuzkoa	37,98	28,21	21,85	1,07	56
Age=1	27,84	22,13	20,41	0,68	44
Araba	22,73	17,35	21,26	0,47	35
Age=2	19,08	16,34	18,96	0,14	33
Male	46,46	44,45	16,96	0,08	89
Satis2 = 1	25,24	23,35	17,54	0,00	47
PAS	14,94	17,79	13,63	-0,03	35
Female	53,54	55,55	15,64	-0,14	111
Age=4	17,04	22,01	12,57	-0,24	44
Age=3	34,95	38,42	14,76	-0,34	77
PDI	37,02	44,29	13,57	-0,64	88
Bizkaia	39,28	54,44	11,71	-1,50	108
Satis2 = 3	5,07	32,94	2,50	-3,64	66
BuyLo=1	4,61	59,63	1,25	-7,28	119

Tabla 5.22: Caracterización del cluster 3 por las categorías de las preguntas cerradas.

Cluster 3/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
no me gusta	3,80	0,79	85	96	14,753
no	12,87	6,08	288	737	13,523
me gusta	4,92	1,57	110	190	11,879
gusta	5,50	1,97	123	239	11,574
con logotipos	1,03	0,24	23	29	6,926
propaganda	1,43	0,46	32	56	6,340
ningun tipo	0,80	0,18	18	22	6,237
me gustan los productos	0,63	0,12	14	15	6,058
ningun	1,39	0,49	31	59	5,782
en general	0,98	0,28	22	34	5,760
tipo	1,97	0,83	44	101	5,739
general	1,03	0,34	23	41	5,246
logotipo	4,69	2,93	105	355	5,103
objetos	1,74	0,81	39	98	4,845
llevar	2,19	1,19	49	144	4,390
hacer publicidad	0,94	0,37	21	45	4,169
comprar	3,44	2,24	77	271	3,978
nada	0,76	0,31	17	37	3,679
marca	1,48	0,79	33	96	3,622
principio	0,45	0,14	10	17	3,484
ehu	0,49	1,09	11	132	-3,142
pertenecer a la upv	0,00	0,30	0	36	-3,216
detalle	0,00	0,30	0	36	-3,223
todos	0,00	0,31	0	37	-3,281
tambien	0,04	0,42	1	51	-3,378
buena	0,36	0,96	8	116	-3,421
a la upv	0,18	0,69	4	84	-3,497
bonito	0,13	0,62	3	75	-3,535
forma	0,18	0,73	4	88	-3,684
dar	0,31	0,97	7	117	-3,769
de pertenecer	0,04	0,50	1	61	-3,852
recuerdo	0,04	0,51	1	62	-3,912
orgulloso	0,04	0,52	1	63	-3,958
manera	0,00	0,41	0	50	-3,980
a conocer	0,00	0,42	0	51	-4,017
ser	4,74	6,66	106	807	-4,142
upv	2,59	4,19	58	507	-4,327
conocer	0,09	0,92	2	112	-5,384
universidad	0,80	2,20	18	267	-5,461
pertenecer	0,09	1,02	2	124	-5,781

Tabla 5.23: Caracterización del cluster 3 por palabras y segmentos repetidos característicos de los encuestados en Castellano.

Cluster 3/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
zalea	0,82	0,25	7	9	3,089
ez	9,85	7,64	84	272	2,652
asko	0,94	0,37	8	13	2,637
propaganda	1,52	0,79	13	28	2,431
logo	3,99	2,81	34	100	2,203
ikusi	1,29	0,73	11	26	1,890
batere	0,59	0,25	5	9	1,739
atsegin	0,35	0,11	3	4	1,695
ez zait	1,76	1,15	15	41	1,652
diseinua	0,47	0,20	4	7	1,545
interesgarria	0,94	0,53	8	19	1,540
gauza	1,29	0,84	11	30	1,396
gutxi	0,35	0,14	3	5	1,325
karpeta	0,35	0,14	3	5	1,325
kontsumismoa	0,35	0,14	3	5	1,325
logotipoa duen	0,35	0,14	3	5	1,322
gainera	0,70	0,39	6	14	1,318
hartu	0,59	0,31	5	11	1,292
horren	0,47	0,22	4	8	1,284
gogoko	2,81	2,25	24	80	1,144
guzti	0,00	0,20	0	7	-1,050
euskal	0,00	0,20	0	7	-1,050
publizitatea	0,00	0,20	0	7	-1,050
eskaini	0,00	0,22	0	8	-1,218
kanpoko	0,00	0,22	0	8	-1,218
edozein	0,23	0,59	2	21	-1,342
zabaldu	0,00	0,25	0	9	-1,374
atzerrian	0,00	0,25	0	9	-1,374
bat	1,88	2,61	16	93	-1,447
beste edozein	0,00	0,28	0	10	-1,514
zergatik ez	0,12	0,48	1	17	-1,547
oparitu	1,17	1,88	10	67	-1,652
lan	0,12	0,51	1	18	-1,666
beste	0,70	1,32	6	47	-1,703
zergaitik	0,23	0,70	2	25	-1,744
irudia	0,00	0,34	0	12	-1,784
unibertsitatea	1,17	1,97	10	70	-1,838
laguntza	0,00	0,39	0	14	-2,025
irakaslea	0,00	0,39	0	14	-2,025
on	0,12	0,65	1	23	-2,172

Tabla 5.24: Caracterización del cluster 3 por palabras y segmentos repetidos característicos de los encuestados en Euskera.

Cuarto cluster

El cuarto cluster se sitúa en el tercer cuadrante y, al igual que en el caso del tercero, va asociado a opiniones negativas sobre la universidad y los productos corporativos. Respecto a su tamaño (23,63 %), comentar que es inferior a los dos primeros y superior al tercero.

Las categorías características de este cluster son *no compraría* y *Muy poco, poco o medianamente satisfecho* con la universidad.

Los términos empleados por los que han respondido en Castellano son, con mayor frecuencia, *no me parece, dinero, no se, no es, pagar, interesante, atractivo, marketing, servicio, me interesa, pública, cosas, taza, ...* que son términos en su mayoría con sentido negativo y asociados sobre todo a productos, pero también a la universidad, como los términos *servicio* o *pública*.

En el caso de las respuestas en Euskera, predominan *ez (no), gogoko (agradable o resultar agradable), ez dut uste (no creo), ez zait (no me...), produktuak gustatzen (gustan los productos), arropa (ropa), eraman (llevar), ekintza (acción, actuación), ez zaizkidalako (porque no me ...), ez zaizkit gustatzen (porque no me gustan ...), logo (logotipo), marka (marca), joera (tendencia), tresna (aparato), ...* Son todos términos y segmentos en sentido negativo principalmente asociados a los productos. En estos términos más frecuentes no aparece ninguno que se pueda relacionar fácilmente con la universidad directamente.

Este cluster, en definitiva está asociado a un rechazo generalizado a los productos corporativos, tanto en las respuestas a las preguntas cerradas como en el vocabulario empleado en la respuesta abierta. Aparece también, de forma más clara en las preguntas cerradas que en el léxico empleado en la abierta, una menor satisfacción con respecto a la universidad. Hay que recordar, sin embargo, que las respuestas asociadas a una muy baja o baja satisfacción con la universidad eran bastante escasas y por ello se han incluido en la categoría de satisfacción media. En ese sentido, es posible que haya cierto riesgo de autoselección adversa en cuanto que las personas que decidieron no responder a la pregunta textual de la encuesta podrían ser precisamente aquéllas que sienten una menor satisfacción con la universidad, pero no nos es posible comprobar este extremo.

Cluster 4/4 (Efectivo: 47 - Porcentaje: 23,63)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
BuyLo=2	90,57	39,35	54,38	8,46	78
Satis2 = 1	57,06	23,35	57,75	5,74	47
Age=2	25,62	16,34	37,06	1,63	33
Students	46,82	37,93	29,17	1,22	76
Araba	21,61	17,35	29,42	0,56	35
Female	59,17	55,55	25,17	0,43	111
PAS	18,63	17,79	24,75	0,13	35
Age=1	24,07	22,13	25,70	0,06	44
Bizkaia	52,09	54,44	22,61	-0,01	108
Satis2 = 3	31,87	32,94	22,86	-0,02	66
Age=3	35,14	38,42	21,61	-0,23	77
Gipuzkoa	26,30	28,21	22,03	-0,26	56
Male	40,83	44,45	21,71	-0,51	89
PDI	34,54	44,29	18,43	-1,45	88
Age=4	13,66	22,01	14,66	-1,61	44
Satis2 = 2	9,16	42,36	5,11	-5,55	84
BuyLo=1	8,03	59,63	3,18	-8,28	119

Tabla 5.25: Caracterización del cluster 4 por las categorías de las preguntas cerradas.

Cluster 4/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
no me parece	0,97	0,31	29	38	6,512
parecer	3,04	1,77	91	215	5,649
me parece	2,07	1,12	62	136	5,177
no	7,73	6,08	231	737	4,185
dinero	0,60	0,25	18	30	3,940
no se	0,67	0,29	20	35	3,924
no es	0,60	0,26	18	31	3,779
pagar	0,47	0,18	14	22	3,668
interesante	1,34	0,79	40	96	3,567
atractivo	0,70	0,35	21	43	3,284
marketing	0,37	0,15	11	18	3,062
servicio	0,33	0,14	10	17	2,773
debe	0,94	0,58	28	70	2,719
me interesa	0,40	0,19	12	23	2,636
publica	0,37	0,17	11	21	2,527
querer	0,57	0,32	17	39	2,434
mas	1,71	1,27	51	154	2,290
preferir	0,30	0,14	9	17	2,280
cosas	0,90	0,61	27	74	2,156
taza	0,43	0,25	13	30	2,067
en general	0,10	0,28	3	34	-2,084
uso	0,20	0,43	6	52	-2,161
me gustan los productos	0,00	0,12	0	15	-2,189
siempre	0,17	0,40	5	48	-2,273
profesores	0,10	0,31	3	37	-2,330
depende	0,03	0,20	1	24	-2,334
otras universidades	0,10	0,31	3	38	-2,401
regalo	0,64	1,03	19	125	-2,467
ehu	0,67	1,09	20	132	-2,558
orgullo	0,03	0,22	1	27	-2,602
estar	0,64	1,06	19	129	-2,651
personal	0,30	0,64	9	78	-2,741
pertenecer a la upv	0,03	0,30	1	36	-3,302
para regalar	0,37	0,86	11	104	-3,469
de pertenecer	0,10	0,50	3	61	-3,905
a la upv	0,17	0,69	5	84	-4,349
no me gusta	0,20	0,79	6	96	-4,581
pertenecer	0,23	1,02	7	124	-5,513
gusta	0,67	1,97	20	239	-6,466
me gusta	0,40	1,57	12	190	-6,529

Tabla 5.26: Caracterización del cluster 4 por palabras y segmentos repetidos característicos de los encuestados en Castellano.

Cluster 4/4					
Palabra o segmento	% Interno	% Global	Frecuencia interna	Frecuencia global	Valor test
ez	11,23	7,64	97	272	4,326
gogoko	4,05	2,25	35	80	3,756
ez dut uste	1,27	0,42	11	15	3,746
ez zait	2,43	1,15	21	41	3,548
produktuak gustatzen	0,58	0,14	5	5	3,138
horrela	1,50	0,65	13	23	3,127
uste	1,97	0,98	17	35	2,976
arropa	0,46	0,11	4	4	2,702
eraman	2,20	1,26	19	45	2,526
ekintza	0,46	0,14	4	5	2,200
ez zaizkidalako	0,46	0,14	4	5	2,195
ez zaizkit gustatzen	0,46	0,14	4	5	2,195
agian	0,58	0,22	5	8	1,979
logo	3,82	2,81	33	100	1,905
marka	1,39	0,81	12	29	1,868
kasuan	0,35	0,11	3	4	1,678
joera	0,35	0,11	3	4	1,678
baizik	0,35	0,11	3	4	1,678
tresna	0,35	0,11	3	4	1,678
edukitzea	8,22	7,13	71	254	1,339
laguntza	0,12	0,39	1	14	-1,218
eskaini	0,00	0,22	0	8	-1,237
zein	0,00	0,22	0	8	-1,237
behar	0,81	1,32	7	47	-1,367
batere	0,00	0,25	0	9	-1,394
kalitate	0,00	0,28	0	10	-1,539
aukera	0,00	0,31	0	11	-1,677
prezio	0,00	0,31	0	11	-1,677
arazorik	0,00	0,31	0	11	-1,677
bertako	0,00	0,37	0	13	-1,930
bat	1,62	2,61	14	93	-2,050
gisa	0,12	0,65	1	23	-2,203
egin	0,81	1,66	7	59	-2,204
bada	0,35	1,01	3	36	-2,205
identifikatzen	0,00	0,45	0	16	-2,270
egon	0,35	1,04	3	37	-2,283
zergatik ez	0,00	0,48	0	17	-2,366
izan	2,55	4,07	22	145	-2,612
zergaitik	0,00	0,70	0	25	-3,110
polita	0,00	0,70	0	25	-3,110

Tabla 5.27: Caracterización del cluster 4 por palabras y segmentos repetidos característicos de los encuestados en Euskera.

Respuestas modales

Se ha realizado un análisis de las respuestas modales según los criterios de selección de la sección 5.1.1, página 125. Las Tablas 5.28 a 5.35 contienen las respuestas modales en cada uno de los clusters para los dos idiomas, utilizando el criterio de los elementos característicos, basados en valores test. En el apéndice B.2.3 están las respuestas modales obtenidas en base al criterio χ^2. Es evidente la diferente longitud de respuestas que favorecen uno y otro criterio, mucho mayor en el caso del criterio χ^2.

A la vista de estas respuestas modales, parece más claro que el primer cluster es un cluster de valoración positiva de la universidad y de los productos, pero más enfocado a los productos. En cambio, el segundo cluster está más enfocado, en cuanto al léxico utilizado al menos, hacia la percepción sobre la universidad. Respecto a los clusters 3 y 4, en el tercero se aprecia un rechazo general a la idea de comprar productos corporativos en general, mientras que en el cuarto parece más un rechazo en concreto a los productos disponibles. Esta distinción es, en cualquier caso, un poco arriesgada, puesto que se están considerando sólo 10 respuestas modales, aún siendo éstas las más representativas. Si se consideran las respuestas modales obtenidas por el criterio χ^2, éstas van en una línea similar. Parece apreciarse que las respuestas más largas tienden a desviarse del objetivo de la pregunta conteniendo términos que, en varios casos, es probable que no se repitan en otras preguntas *(contubernio, mercachifles, calefacción, ...)* siendo por tanto de escaso aprovechamiento estadístico. En cualquier caso, la observación de respuestas reales ayuda mucho en la interpretación de las palabras y segmentos repetidos que aparecen como característicos tras la extracción de factores de las tablas léxicas y de la posterior clasificación.

Criterio	Respuesta
3,154	por moda. la moda de lo de siempre. como los llaveros del athletic.
2,893	porque se supone que son de buena calidad
2,797	para ayudar a invertir en educacion
2,773	si la calidad es buena, y el precio razonable, ¿por que no?
2,560	porque es una manera de que la sociedad conozca la u.p.v.
2,432	porque formo parte de ella de una forma u otra.
2,256	es representativo de una parte de nuestra sociedad
2,222	por que pertenezco a ella.
2,194	si, pero depende de los disenos. ¿se van a implicar las chicas/os de bbaa?
2,194	dependeria del diseno de dicho producto

Tabla 5.28: Respuestas modales para individuos que responden en Castellano. Criterio de selección según elementos característicos (Valores test). Cluster 1.

Criterio	Respuesta
4,223	porque pertenezco a la upv/ehu.
4,223	por pertenecer a la upv/ehu
4,184	porque indicaria que estoy orgullosa de pertenecer a la upv/ehu
4,048	porque me siento orgullosa de pertenecer a la upv.
3,905	porque pertenezco a la uvp/ehu, y me siento orgullosa de ello.
3,878	por que me siento orgullosa de pertenecer a la upv/ehu
3,742	porque se favorece la expansion de la upv
3,607	por apoyar a la upv/ehu
3,459	por sentirme parte de la upv y estar orgulloso de ella.
3,434	por estar muy satisfecho de mi pertenencia a la upv/ehu.

Tabla 5.29: Respuestas modales para individuos que responden en Castellano. Criterio de selección según elementos característicos (Valores test). Cluster 2.

Criterio	Respuesta
13,523	no
13,523	no
13,523	no acostumbro
13,523	no
13,523	porque no me llama la atencion.
13,523	no se
13,523	y porque no!
12,549	no me gustan
11,574	me gusta el merchandising
10,067	no me gustan los logotipos

Tabla 5.30: Respuestas modales para individuos que responden en Castellano. Criterio de selección según elementos característicos (Valores test). Cluster 3.

Criterio	Respuesta
5,649	porque me parece una tonteria
5,649	sinceramente, me parece una ridiculez.
5,649	me parece tonto.
4,917	no me parece atracctivo
4,373	no parece atractivo
4,373	no me parecen atractivos.
4,373	no me parecen llamativos ni atractivos
3,810	porque no me parece que sean interesantes
3,773	no me interesan los envoltorios, me interesan los contenidos
3,734	no me atrae en absoluto

Tabla 5.31: Respuestas modales para individuos que responden en Castellano. Criterio de selección según elementos característicos (Valores test). Cluster 4.

Criterio	Respuesta
2,028	polita delako *(porque es bonito)*
1,610	unibertsitatearen irudia indartu eta suspertzeko *(para fortalecer y revitalizar la imagen de la universidad)*
1,599	konnotazioak orokorrean onak direlako *(porque las connotaciones en general son buenas)*
1,597	lankide batentzat opari egokia delako *(porque es un regalo adecuado para un trabajador)*
1,499	irudi korporatiboa indartzen duelako. opar egokia izan daitekeelako eremu profesionalean *(Porque fortalece la imagen corporativa. Porque puede ser un regalo adecuado en el ambito profesional)*
1,482	laguntzeko modu on bat delako *(porque es una buena forma de ayuda)*
1,373	bere irudi instituzionala erakargarria delako, pozik sentitzen naizelako kolektibo honetan lan egiteaz *(porque su imagen institucional es atractiva, me siento contento de trabajar en este colectivo)*
1,339	publizitatea baita eta hori da empresa baten helburua, ezaguna izatea. *(Ya que es publicidad y ese es el objetivo de una empresa, el ser conocida)*
1,233	korporaziorekin identifikatzen naizelako *(porque me identifico con la corporación)*
1,097	egokiak izan daitezke ikerketetetan (batez ere soziologian, psikologian, etab.) elkarrizketatzen diren pertsonei oparitzeko edo gonbidatzen diren kanpoko irakasleei oparitzeko *(Podrían ser adecuadas en las investigaciones (sobre todo en sociología, en psicología, etc.) para regalar a los entrevistados o a los profesores invitados de fuera)*

Tabla 5.32: Respuestas modales para individuos que responden en Euskera. Criterio de selección según elementos característicos (Valores test). Cluster 1.

Criterio	Respuesta
2,090	unibertsitate honen partaide izatea, ikastegi batean ikastea baino haratago egon behar lukeelako *(El ser miembro de esta universidad, tendría que estar más allá que el aprender en uno de sus centros)*
1,935	ikasten dudan lekuaz arro nagoelako *(Porque estoy orgulloso del lugar en el que estudio)*
1,597	unibertsitatea gizarteari zabaltzeko *(para abrir la universidad a la sociedad)*
1,523	laguntza bezala *(Como ayuda)*
1,407	unibertsitatea beharrezkoa dela aldarrekatzeko *(para pregonar que la universidad es necesaria)*
1,280	euskal herrian kokatutako erakunde kultural baten produktua litzatekeelako
1,268	zure unibertsitateko garaiko oroigarri bat izango litzatekeelako *(Porque sería un recuerdo de la época de tu universidad)*
1,243	unibertsitate publikoari propaganda egiteko *(Para hacer propaganda de la universidad pública)*
1,241	ni oso arro nago ehu-n ikasten ari naizelako,izan ere betiko nire nahia bertan ikastea baitzen.behin ehu-n ikasten egonik,oso arro egongo nitzeke holako produktua nere eskutan edukitzea eta jendeak ikustea ni ehu-ko ikasle bat,eta beraz bere parte bat naizela *(porque estoy muy orgulloso de estudiar en la ehu, de hecho siempre desee estudiar aqui. una vez de estar estudiando en la ehu, estaría muy orgulloso de tener un producto en mis manos y de que la gente viera que soy un estudiante de la ehu y por tanto una parte de ella)*
1,112	azkenean hemengo produktu bat erosiko nuke eta horrekin unibertxitateari lagunduz euskal gizarte osoari laguntzeko ahalegin txiki bat izango litzateke *(Al final compraria un producto de aquí y al ayudar con ello a la universidad haría un pequeño esfuerzo para ayudar a toda la sociedad vasca)*

Tabla 5.33: Respuestas modales para individuos que responden en Euskera. Criterio de selección según elementos característicos (Valores test). Cluster 2.

Criterio	Respuesta
1,931	ez zaidalako inoren ez ezeren propagandarik egitea. ez nioke begiratu logoa duen edo ez *(No me gusta hacer propaganda de nadie ni de nada. No miraría si tiene logo o no)*
1,914	ez nahiz 'marka' zale *(no soy aficionado a las marcas)*
1,816	ez zaidalako gustatzen propaganda egitea *(porque no me gusta hacer propaganda)*
1,816	ez zaidalako propaganda egitea gustatzen *(Porque no me gusta hacer propaganda)*
1,733	propaganda egiteko *(para hacer propaganda)*
1,589	ez naiz llogotipo instituzionalen zalea *(no soy aficionado a logotipos institucionales)*
1,557	ez zaidalako gustatzen inolako eratako propagandarik ibiltzea *(porque no me gusta llevar ningún tipo de propaganda)*
1,535	ez naiz horrelako propaganda egin zalea *(No soy aficionado a hacer propagandas así)*
1,520	ez dudalako interesgarririk ikusten *(porque no lo veo interesante)*
1,500	ez zaizkit produktu logotipodunak gustatzen *(No me gustan los productos que tienen logotipo)*

Tabla 5.34: Respuestas modales para individuos que responden en Euskera. Criterio de selección según elementos característicos (Valores test). Cluster 3.

Criterio	Respuesta
3,329	logotipoak ez zaizkit gustatzen *(no me gustan los logotipos)*
3,110	ez zaizkidalako horrelako produktuak gustatzen *(porque no me gustan los productos así)*
2,830	logotipoa gustatzen zait *(me gusta el logotipo)*
2,514	logotipoa daramaten produktuak gustokoak ez ditudalako *(porque los productos que llevan logotipo no son de mi gusto)*
2,511	ez zaizkit logotipoak dauzkaten produktuak gustatzen *(No me gustan los productos que tienen logotipos)*
2,355	logotipoak ez eramateko joera daukat *(tengo tendencia a no llevar logotipos)*
2,297	ez dudalako gustoko horrelako propaganda eramatea gainean *(porque no me gusta llevar encima ese tipo de propaganda)*
2,259	horrelako gauzak erostea ez dut gustoko *(no me gusta comprar este tipo de cosas)*
2,137	halakoak erabiltzea ez dudalako gustuko *(No me gusta llevar ese tipo de cosas)*
2,122	ez zait gustatzen edozein publizitate eramaten *(no me gusta llevar ninguna publicidad)*

Tabla 5.35: Respuestas modales para individuos que responden en Euskera. Criterio de selección según elementos característicos (Valores test). Cluster 4.

5.3.5. Conclusiones de la aplicación

Los datos provenientes de la encuesta on-line sobre satisfacción acerca de la Universidad del País Vasco y sobre la aceptación de la tienda de productos corporativos muestran, a partir de las variables consideradas, la estrecha relación entre ambas. Para ello, se han utilizado exclusivamente herramientas de tipo descriptivo del ámbito del análisis factorial y de los métodos de clasificación.

El hecho de disponer de dos tipos de respuestas, unas obtenidas a partir de preguntas cerradas y otra de una pregunta abierta, sugiere la conveniencia de equilibrar ambos tipos de preguntas. Esto se puede conseguir, entre otras posibles técnicas, a partir del Análisis Factorial Múltiple de Tablas Mixtas, que se ha aplicado aquí. El hecho de disponer de datos en dos idiomas diferentes, supone por un lado, una adaptación del método empleado y, por otro, plantea la posibilidad de que la diferencia en el idioma produzca alguna divergencia en el comportamiento de los encuestados. La adaptación del método empleado, así como la técnica del Análisis de Correspondencias Múltiples sobre las tablas apiladas, no muestra diferencias significativas entre ambos idiomas, aún cuando los términos empleados en ambos idiomas no sean siempre traducciones literales unos de otros. A esta conclusión se llega tras obtener una clasificación de los individuos y una descripción de estas clases de una manera muy detallada, teniendo incluso en consideración el vocabulario y las frases que incluyen el vocabulario más característico de cada clase.

Aparte de la ausencia de diferencias significativas en el idioma, este trabajo, en cierta medida, cuantifica y finalmente describe las clases de los individuos tanto favorables como opuestos a la implantación de la tienda de productos corporativos. Esta información es de suma utilidad para la orientación de una política de marketing de dichos productos corporativos y caracteriza los segmentos hacia los cuales habría que seguir una política, o utilizar una línea de productos, diferente.

5.4. Conclusiones

Este capítulo de la tesis complementa, por un lado, la presentación de técnicas de análisis de tablas múltiples iniciado en el capítulo 4 y, por otro lado, completa el estudio empírico que sobre la viabilidad de la tienda corporativa de la UPV/EHU se ha presentado en el capítulo 3.

Sin embargo, tiene entidad propia ya que se enmarca en una de las variaciones de la Minería de Datos: el Text Mining o Minería de textos. En este caso, las técnicas de análisis de datos se aplican para descubrir patrones ocultos en textos, esto es, para extraer conocimiento a partir de la información contenida

en bases de datos textuales.

El análisis textual es la herramienta estadística exploratoria por excelencia para el tratamiento estadístico de textos. Generalmente su tratamiento viene asociado a la técnica del Análisis de Correspondencias de una tabla léxica de frecuencias generada a partir del texto de interés.

En ocasiones, como en el caso de una encuesta, existe información no sólo de tipo textual sino también medida en torno a variables cuantitativas y/o categóricas. En este caso, dada la diferente naturaleza de las variables disponibles es probablemente preferible equilibrar estos diferentes tipos de variables. Aquí es donde el Análisis Factorial Múltiple de Tablas Mixtas es de utilidad, como se ha puesto de manifiesto en este estudio.

El hecho de que los individuos disponibles pertenezcan a dos o más clases diferentes (en este caso, dos idiomas distintos) hace que, en ocasiones, parte de las variables sean comunes y parte no lo sean. Es decir, parte del espacio de referencia es común y parte específico a cada clase. En este caso, una alternativa viable es la propuesta en este capítulo.

Finalmente, como en otros estudios, una clasificación sobre los factores principales, en este caso del Análisis Factorial Múltiple, ayuda a la descripción de los individuos y a su caracterización. El hecho de que algunas de las variables sean, en realidad, frecuencias provenientes de una tabla léxica, no dificulta tal descripción, puesto que existen herramientas suficientes para describir las clases obtenidas, teniendo en cuenta las específicas a la clasificación posterior a un análisis textual.

CAPÍTULO 6

Pasado, presente y futuro de esta tesis doctoral

6.1. Pasado

En el camino recorrido desde los comienzos de esta tesis hasta su culminación son muchos los trabajos de investigación realizados. Todos ellos han contribuido, en mayor o menor medida, al entramado final de la misma, constituyendo, por tanto, su pasado, reciente en algunos casos. Y, como no puede entenderse de otra manera, han sido presentados a la comunidad científica utilizando las diferentes vías disponibles: seminarios, comunicaciones y/o póster en congresos, tanto de carácter nacional como internacional y publicaciones en revistas especializadas.

A continuación se presenta en orden cronológico una selección de los trabajos, relacionados con los contenidos de esta tesis, que han sido publicados y que, desde el punto de vista del autor de esta tesis, merecen ser destacados:

- M. I. Landaluce, K. Fernández & J. I. Modroño (1999), 'Reflexiones sobre el uso comparativo del análisis factorial múltiple y de la metodología statis para el análisis factorial múltiple', *Methodologica* **7**, París, 37–65.

- E. Abascal, K. Fernández, M. I. Landaluce & J. I. Modroño (2001), 'Diferentes aplicaciones de las técnicas factoriales de análisis de tablas múltiples en las investigaciones mediante encuestas', *Metodología de Encuestas* **3**(2), Sevilla, 251–279.

- E. Abascal, K. Fernández, J. I. Modroño, M. I. Landaluce (2001), 'Técnicas factoriales de análisis de tablas múltiples: nuevos desarrollos empíricos', Documentos de trabajo BILTOKI D. T. 2001.06, EA II, EAIII, Fundamentos del Análisis Económico e Instituto de Economía Pública, UPV/EHU, Bilbao, https://addi.ehu.es/bitstream/10810/5765/1/2001.06.pdf.

- J. I. Modroño, K. Fernández, M. I. Landaluce (2001), 'A Multivariate Two Step Method for Tables of Categorical Variables', en V. Esposito Vinzi, C. Lauro, A. morineau, M. Tenenhaus Eds., PLS AND RELATED METHODS Proceedings of the PLS 01, CISIA-CERESTA, Montreuil, 211-224.

- J. I. Modroño, K. Fernández, M. I. Landaluce (2002), A Two Step Experimental Method for the Analysis of Multiple Tables of Categorical Variables, en Sigbert Klinke, Patricia Ahrend, Luise Richter Eds., Proceedings of the Conference CompStat 2002, Humboldt-Universität zu Berlin, Berlin.

- K. Fernández, I. Gallastegui, J. I. Modroño, V. A. Núñez (2003), 'Clients characteristics and marketing of products: some evidence from a financial institution', The International Journal of Bank Marketing, Vol. 21 No. 5, 243-254.

- J. I. Modroño, K. Fernández, M. I. Landaluce (2003), Una propuesta para el análisis de tablas múltiples de variables cualitativas, Documentos de trabajo BILTOKI D. T. 2003.11, AE II, EA III, Fundamentos del Análisis Económico e Instituto de Economía Pública, UPV/EHU, Bilbao, https://addi.ehu.es/bitstream/10810/5712/1/2003.11.pdf.

- K. Fernández, M. A. Garín, J. I. Modroño (2004), 'Motivación de los estudiantes de LE y LADE ante el estudio de la Estadística', Documentos de trabajo BILTOKI No 2004-03, AE II, EA III, Fundamentos del Análisis Económico e Instituto de Economía Pública, UPV/EHU, Bilbao. https://addi.ehu.es/bitstream/10810/5683/1/2004.03.pdf

- K. Fernández, J. I. Modroño, M. Isabel Landaluce (2004), 'ACM y Statis Dual ponderado: Dos técnicas complementarias para analizar una visión de la cultura de la Universidad', Estadística Española, Vol. 46 No. 156, 205-228.

- K. Fernández, M. I. Landaluce, J. I. Modroño (2005), PLS in Market Research. The Opening a Gift Shop in an Educational Institution, en T.

Aluja, J. Casanovas, V. Esposito Vinci, A. Morineau, M. Tenenhaus eds., PLS AND RELATED METHODS Proceedings of the PLS 05, TEST & GO, Paris, 243-250.

- K. Fernández, A. M. Martín, J. I. Modroño & Zorrilla, P. (2005), Análisis de la viabilidad de la tienda universitaria de la UPV/EHU "EHUdenda" de acuerdo con la imagen percibida de la institución, Informe Técnico, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Bilbao.

- M. A. Garín, K. Fernández, E. Ferreira, B. Goitisolo, J. I. Modroño, J. M. Orbe, J. A. Rubio, M. J. Bárcena, J. Virto, A. Zarraga, M. A. Zarraga, (2005), 'Motivación de los estudiantes universitarios: El caso de la estadística en una facultad de Económicas', en A. Goñi Ed., Innovación educativa en la universidad, Servicio editorial de la UPV/EHU, Bilbao, 169-180.

- T. Palomares, K. Fernández, J. I. Modroño (2005), 'Las tecnologías de la información y comunicación como factor de aprendizaje en la docencia universitaria', en A. Goñi Ed., Innovación educativa en la universidad, Servicio editorial de la UPV/EHU, Bilbao, 145-156.

- K. Fernández, J. I. Modroño, T. Palomares (2006), 'Las Tecnologías de la Información y Comunicación en la Docencia Universitaria Presencial. Aplicación en Distintas Titulaciones y Áreas de Conocimiento', Documentos de trabajo BILTOKI No 2006-01, AE II, EA III, Fundamentos del Análisis Económico e Instituto de Economía Pública, UPV/EHU, Bilbao. `https://addi.ehu.es/bitstream/10810/5646/1/2006.01.pdf`

- T. Palomares, P. Bilbao, J. González, J. I. Modroño, K. Fernández, F. J. Saez, Y. Chica, A. L. Torres, M. J. Chomon (2006), 'Las tecnologías de la información y comunicación en la enseñanaza universitaria: estudio multidepartamental de la influencia de su utilización sobre la motivación, el autoaprendizaje y la participación activa del alumno', en J. Guisasola y T. Nuño Eds., La educación universitaria en tiempos de cambio, Servicio editorial de la Universidad del País Vasco UPV/EHU, Bilbao, 339-350.

- K. Fernández, M. Isabel Landaluce, J. I. Modroño (2007), 'Exploración Textual en el contexto del Modelo de Valores en Competencia. Aplicación al tipo de cultura de la UPV-EHU', *Estadística Española*, **49**(166), (tercer cuatrimestre 2007), 501-530.

- T. Palomares, K. Fernández, J. I. Modroño, J. Gonzalez, F. J. Sáez, Y. Chica, A. L. Torres, M. J. Chomón, P. Bilbao (2007), 'Las Tecnologías

de la Información y la Comunicación en la enseñanza universitaria: influencia sobre la motivación, el autoaprendizaje y la participación activa del alumno', Revista de Psicodidáctica, Vol. 12, No. 1, 51-78

- K. Fernández, M. I. Landaluce, A. M. Martín, J. I. Modroño (2008), 'Data Mining of an On-Line Survey. A Market Research Application', en Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, Reinhold Decker Eds., Data Analysis, Machine Learning and Applications, Springer-Verlag, Berlín-Heidelberg, 183-191.

- M. Bécue, K. Fernández, J. I. Modroño (2009), 'Analysis of a mixture of closed and open-ended questions in the case of a multilingual survey', en C. H. Skiadas Ed., Advances in Data Analysis, Springer/Birkhäuser, Berlin, 23-34.

- K. Fernández, M.I. Landaluce, A. Martín, J.I. Modroño (2011), 'Knowledge extraction from a large on-line survey: a case study for a higher education corporate marketing', Journal of Applied Statistics 38, TAYLOR & FRANCIS, United Kingdom, 2661-2679.

- K. Fernández, M. A. Garín, J. I. Modroño (2012), 'Visual Displays. Some evidence through artificial and real data', enviado y pendiente de evaluación.

La lista de congresos en los que se han presentado resultados totales o parciales de los artículos publicados es la siguiente:

XXIV Congreso Nacional de Estadística e Investigación Operativa AFM versus STATIS: unas reflexiones (K. Fernández, M. I. Landaluce, J. I. Modroño), SEIO, Almeria, Octubre 1998.

VI Congreso de Economía Regional de Castilla y León Estudio Comparativo de la Metodología STATIS y el Análisis Factorial Múltiple (K. Fernández, M. I. Landaluce, J. I. Modroño), Junta de Castilla y León, Zamora, Noviembre 1998.

Large Scale Data Analysis An Approximation to the Analysis of Multiple Tables Of Categorical Variables (K. Fernández, M. I. Landaluce, J. I. Modroño), Zentralarchiv für Empirische Sozialforschung, Colonia, Mayo 1999.

XIII Reunión ASEPELT-España Una extensión del análisis de tablas múltiples de variables cualitativas (K. Fernández, M. I. Landaluce, J. I. Modroño), ASEPELT, Burgos, Junio 1999.

XXV Congreso Nacional de Estadística e Investigación Operativa Propuesta metodológica para el análisis de tablas múltiples de naturaleza cualitativa (K. Fernández, M. I. Landaluce, J. I. Modroño), SEIO, Vigo, Abril 2000.

COMPSTAT 2000 An approximation to the analysis of multiple tables of categorical variables (K. Fernández, M. I. Landaluce, J. I. Modroño), Utrecht University, Agosto 2000.

1er Congreso de Investigación mediante Encuestas Diferentes aplicaciones de las técnicas factoriales de análisis de tablas múltiples en las investigaciones mediante encuestas (E. Abascal, K. Fernández, M. I. Landaluce, J. I. Modroño), Facultad de Psicología. Universidad de Sevilla. Sociedad Internacional de Profesionales de la Investigación en Encuestas, Septiembre 2000.

XV Reunión ASEPELT-España Un Método Experimental en Dos Etapas para el Análisis de Tablas Múltiples de Variables Categóricas (K. Fernández, M. I. Landaluce, J. I. Modroño), ASEPELT, Madrid, Junio 2002.

COMPSTAT 2002 A 2 Step Experimental Method for the Analysis of Multiple Tables of Categorical Variables (K. Fernández, M. I. Landaluce, J. I. Modroño), Berlin, Agosto 2002.

XXVI Congreso Nacional de Estadística e Investigación Operativa ACM y STATIS DUAL : Dos Métodos Complementarios para el Análisis de Tablas Cualitativas (K. Fernández, M. I. Landaluce, J. I. Modroño), SEIO, Lleida, Abril 2003.

International Conference on Correspondence Analysis and Related Methods (CARME 2003) New Proposals in the Exploratory Analysis of Joint Tables of Categorical Variables (K. Fernández, M. I. Landaluce, J. I. Modroño), Universitat Pompeu Fabra, Barcelona, Julio 2003.

4th Internacional Symposium on PLS and Related Methods PLS in Market Research. The opening of a gift shop in an educational institution (K. Fernández, M. I. Landaluce, J. I. Modroño), Barcelona, Septiembre 2005.

IV Congreso de Metodología de Encuestas Intención de compra de artículos con el logotipo de la UPV/EHU. Diseño de una encuesta online y análisis preliminares. (K. Fernández, A. Martín, J. I. Modroño, P. Zorrilla), Pamplona, Septiembre 2006.

IV Congreso de Metodología de Encuestas Artículos con el logotipo de la UPV/EHU. Análisis factoriales multivariantes y análisis de clasificación (K. Fernández, A. Martín, J. I. Modroño), Pamplona, Septiembre 2006.

IV Congreso de Metodología de Encuestas Análisis de la intención de compra de artículos con el logotipo de la UPV/EHU. Una aplicación de la Metodología PLS (K. Fernández, M. I. Landaluce, A. Martín, J. I. Modroño), Pamplona, Septiembre 2006.

IV Congreso de Metodología de Encuestas Análisis de la intención de compra de artículos con el logotipo de la UPV/EHU. Modelos de regresión Logit (K. Fernández, A. Martín, J. I. Modroño), Pamplona, Septiembre 2006.

31st Annual Conference on the German Classification Society on Data Analysis, Machine Learning and Applications Data Mining of an on-line survey. A market research application (K. Fernández, M. I. Landaluce, A. Martín, J. I. Modroño), Friburgo (Alemania), Marzo 2007.

Applied Stochastics Models and Data Analysis, XIIth ASMDA 2007 Analysis of a mixture of closed and open-ended questions in the case of a multilingual survey (M. Bécue, K. Fernández, J. I. Modroño), Chania, Creta, Junio 2007.

RC33 Logic and Methodology in Sociology. 7th International Conference on Social Science Methodology Principal axes methods for categorical variables. Some evidence from student motivation surveys (K. Fernández, M. A. Garín, J. I. Modroño), Naples, Italia, Septiembre 2008.

RC33 Logic and Methodology in Sociology. 7th International Conference on Social Science Methodology Simultaneous analysis of open-ended and closed questions by means of canonical correspondence analysis (M. Bécue, K. Fernández, J. I. Modroño), Naples, Italia, Septiembre 2008.

Symposiun on Learning and Data Science SLDS 2009 Visualization based on first singular values. CA versus PCA (K. Fernández, M. A. Garín, J. I. Modroño), University of Paris-Dauphine, Abril de 2009.

Correspondence Analysis and Related Methods CARME 2011 Visual Displays. Some evidence through artificial and real data (K. Fernández, M. A. Garín, J. I. Modroño), Rennes, Francia, Febrero 2011.

6.2. Presente

En esta sección se reflejarían las principales conclusiones alcanzadas en los distintos capítulos de la tesis.

El presente de esta tesis queda claramente reflejado a través de todas y cada una de las aportaciones y conclusiones alcanzadas en las dos partes y los capítulos que la integran. Sirva esta sección del capítulo final como radiografía de las mismas.

6.2.1. Capítulo 3: Aplicación de las técnicas multivariantes a una encuesta on-line: enfoque desde el Data Mining

En el capítulo 3 se han utilizado técnicas de data mining para la extracción de conocimiento de una encuesta on line. Las encuestas de tipo on line son cada vez más frecuentes en los métodos de investigación de encuestas y marketing. La automatización del proceso alcanza incluso al propio diseño de la misma, además del proceso y ejecución de la encuesta, por lo que es cada vez más relevante trabajar con este tipo de cuestionarios. Sin olvidar, por supuesto, que las características y problemática de este tipo de encuestas son, sin descuidar sus peculiaridades, a menudo similares a otras, como por ejemplo, las enviadas por correo ordinario.

Las ventajas de estas encuestas son evidentes desde el punto de vista del procesamiento, tanto en tiempo necesario como en el coste del mismo, a la vez que permite que el tamaño de la información acumulada sea potencialmente enorme. Es en este punto donde las técnicas de Data Mining son de particular utilidad.

Las variables obtenidas en la encuesta son mayoritariamente cualitativas y, dentro de ellas, generalmente ordinales. En su mayor parte muestran estados de opinión sobre la universidad, productos corporativos con su imagen y características generales de los mismos, además de reflejar algunas características personales de los encuestados.

El tipo de variables seleccionadas da lugar al uso tanto de técnicas exploratorias multivariantes clásicas como el Análisis de Componentes y Correspondencias Múltiples, complementadas con un análisis de clasificación sobre los factores como técnicas de carácter predictivo como el Partial Least Squares (PLS) path modelling y los modelos logit. El objetivo perseguido en cualquier caso es el estudio de la viabilidad de la tienda corporativa de la Universidad del País Vasco (UPV/EHU).

Con respecto a las conclusiones obtenidas con cada una de las técnicas

hay que decir que las divergencias entre las conclusiones están asociadas a los específicos objetivos de cada una de las metodologías. Así, los métodos exploratorios ayudan a describir la información dentro de un conjunto relativamente grande de datos, mientras que los métodos predictivos, como el PLS o los modelos logit, permiten modelizar el comportamiento de los individuos, utilizando herramientas inferenciales para buscar y seleccionar un modelo mejor o para establecer claramente las características de los individuos.

En términos prácticos, y desde un punto de vista de la difusión de la información, los métodos predictivos, sea cual sea la técnica detrás de ellos, permiten presentar las relaciones entre variables a través de coeficientes, cuya interpretación es similar, con relativa independencia del método de estimación empleado, con lo que resultan directamente asimilables para personas familiarizadas con métodos de este tipo. Los métodos multivariantes son particularmente ricos en cuanto a su capacidad descriptiva, y son particularmente interpretables para una persona quizás no experta, aunque sí entrenada, en cuanto a la información gráfica presentada en los planos principales de las diversas técnicas que se emplean y convenientemente apoyada en los coeficientes de ayudas a la interpretación que están disponibles habitualmente.

Los resultados particulares de cada método permiten extraer conclusiones que en parte son comunes y en parte se van complementando entre todos ellos. Es posible, en particular, establecer las características tanto de los individuos que son los más probables compradores de los productos corporativos como las de los que no están interesados en los mismos, así como algunas características de tales productos. Esto permite tanto dirigir los productos existentes hacia su público objetivo como considerar otros productos diferentes para su incorporación a la gama para tratar de acceder al tipo de clientes menos receptivos al elenco actual. Estas conclusiones son valiosas desde un punto de vista de objetivos de mercado y proporciona pautas útiles de marketing en este caso concreto.

6.2.2. Capítulo 4: Tablas Múltiples de tablas de efectivo diferente

Este capítulo se refiere a una nueva aportación enmarcada en el contexto de adaptaciones, extensiones y aplicaciones del Análisis Factorial Múltiple, dentro del ámbito del Análisis de Tablas Múltiples. Una combinación de técnicas factoriales de tabla única (ACM) y de tabla múltiple (AFM), que permite el tratamiento simultáneo de grupos de individuos diferentes en los que se ha medido la misma información a través de variables nominales.

A modo de resumen, a continuación se resaltan las características de la

extensión propuesta:

1. Los datos son los que proporcionan los factores más relevantes detrás de cada grupo de individuos, al igual que otras técnicas factoriales.

2. Se proporciona una visión global de los datos, sin que ninguno de los grupos considerados tenga excesivo protagonismo ni quede eclipsado por el resto. De entre los resultados proporcionados destacan las distintas medidas de relación entre los grupos analizados. Estas medidas ayudan a tomar decisiones sobre los grupos a mantener en un análisis global, por su similitud, o a analizar por separado dado su comportamiento específico y distinto al resto.

3. Los individuos originales pierden su identidad. No es, por tanto, una técnica adecuada cuando el interés se centre en examinar su comportamiento con respecto a las variables medidas. En el caso no obstante habitual en que los individuos no tengan interés en sí mismos, esta técnica aparece como útil ya que cuando menos preserva las clases definidas a través de ellos en su diferente caracterización, por medio de las categorías de las variables originales.

6.2.3. Capítulo 5: Tablas múltiples en el análisis textual

Este capítulo revisita el problema de las tablas múltiples mixtas, que son aquéllas que pueden contener tres tipos de variables o elementos obtenidos de un mismo conjunto de individuos: variables cuantitativas, categorías definidas a partir de variables cualitativas y/o frecuencias como las correspondientes a unidades léxicas empleadas en un corpus textual como las respuestas a una pregunta abierta de una encuesta. La particularidad de este capítulo es cuando el texto (*e.g.,* la pregunta abierta de la encuesta) aparece en dos idiomas diferentes, en cuyo caso las variables medidas, cuantitativas o cualitativas, son las mismas, pero no así los términos empleados ni, por tanto, las frecuencias observadas de los mismos. El objetivo en este tipo de tablas es doble: por un lado interesa el análisis de las variables y de la interestructura de las tablas y, por otro, examinar posibles diferencias entre los idiomas, es decir, entre los individuos que utilizan diferentes idiomas y asociadas a este hecho.

La configuración en tablas múltiples propuesta deja en el aire la elección de los pesos de los individuos, a elegir entre uniforme, relativa a las frecuencias marginales del número de términos empleados en el texto (pregunta abierta) con respecto a todos los individuos o relativa a las mismas frecuencias marginales pero manteniendo constante el peso de las subtablas. Cualquiera de las

ponderaciones no uniformes favorece las respuestas más largas, equilibrando adicionalmente la última las respuestas en los dos idiomas.

Este técnica se aplica a datos provenientes de la encuesta on-line sobre satisfacción acerca de la Universidad del País Vasco y sobre la aceptación de la tienda de productos corporativos. Se trata de determinar una posible diferencia de idiomas tanto tratando esta característica desde un punto de vista ilustrativo en un ACM utilizando solo variables cualitativas como mediante tablas mixtas, utilizando las frecuencias de los respectivos corpus textuales de las respuestas.

Ninguno de los métodos empleados muestran finalmente diferencias significativas entre ambos idiomas, aún cuando los términos empleados en ambos idiomas no sean siempre traducciones literales unos de otros, que es cuando este tipo de técnicas son particularmente relevantes. En este sentido, y al igual que las conclusiones de la aplicación realizada en el capítulo 3, este resultado, junto con los demás que se obtienen sobre las otras variables, es de suma utilidad para la orientación de una política de marketing de los productos corporativos de la universidad, en este caso.

6.3. Futuro

En la primera parte de esta tesis se ha hablado de diferentes técnicas de Data Mining que pueden ser de utilidad en el análisis de encuestas, en particular, para la extracción de información en la encuesta sobre satisfacción hacia la universidad y sus productos corporativos.

Son muchas las técnicas susceptibles de ser aplicadas a este conjunto de datos, y sólo algunas, quizás las más relevantes, han sido utilizadas aquí. Dos variantes de las mismas nos parece interesantes de considerar:

- En el caso de la estimación del modelo Logit, la variable endógena no es dicotómica, sino que se transforma en tal para el análisis. La variable, en realidad presenta cuatro categorías, y además es ordinal. Cabe, por tanto, realizar el análisis mediante un logit multinomial ordenado.

- En el caso del PLS path modelling, el tratamiento de las variables categóricas en la construcción de las variables latentes no es sustancialmente diferente de las variables continuas. Puede ser un problema menor mientras que las variables manifiestas sean al menos ordinales, pero es claro que deja de serlo en el caso de variables puramente nominales. Se hace preciso prestar más atención a este caso, manteniendo las ventajas de los modelos PLS, como la independencia distribucional. Dos alternativas

recientes son el algoritmo de Máxima Verosimilitud Parcial (Jacobowicz & Derquenne 2007) y el algoritmo de Trinchera & Russolillo (2010).

En el capítulo 4 se ha tratado el tema del análisis de tablas múltiples, desde la perspectiva del AFM, de subtablas con individuos diferentes. Nuestras futuras aportaciones se enmarcarán, por un lado, y al igual que la presente tesis doctoral, en este contexto de adaptaciones, extensiones y aplicaciones del AFM. En particular:

- Queremos seguir los pasos de Le Dien & Pagés (2003) y profundizar en una nueva extensión o adaptación del AFM, el Análisis Factorial Múltiple Dual, diseñada para el tratamiento de tablas múltiples cuantitativas referidas a distintos conjuntos de individuos. Permite la representación simultánea de las variables, valoradas en cada uno de los grupos de individuos, sobre unos ejes comunes en los que la participación de los grupos en la obtención de estos factores está equilibrada. La representación de las variables correspondientes a los distintos grupos de individuos se interpreta en términos de correlación entre variables. El estudio de las nubes de variables correspondientes a cada grupo permite estudiar las correlaciones entre las variables. En este contexto nos planteamos un doble objetivo; el primero será poner de manifiesto la posibilidad de extender esta metodología al análisis de un conjunto de variables cualitativas construyendo la tabla de variables indicadoras asociadas a estas últimas siguiendo el mismo razonamiento empleado en AFM de variable categóricas, basado en la equivalencia entre el ACM y el ACP normado de la tabla disyuntiva completa cuando se han ponderado las variables indicadoras de manera adecuada (este trabajo de investigación ya ha sido iniciado en Abascal et al. (2008)). El segundo objetivo, y no menos interesante, es realizar un minucioso estudio comparativo de esta nueva adaptación y la presentada en esta tesis, ya que se trataría de dos enfoques metodológicos distintos apropiados para el mismo tipo de información de partida.

- Por otro lado, en la Escuela Francesa de Análisis de Datos, y de forma casi paralela al AFM, surgió otra filosofía metodológica para el análisis de tablas múltiples: la técnica denominada STATIS (Structurations des TAbleaux a Trois Indices de la Statistique), con la que ya hemos tenido alguna toma de contacto, por lo que somos conscientes de las similitudes y diferencias que ambas metodologías presentan. En un futuro próximo queremos continuar con esta línea de investigación y profundizar sobre la utilidad de la filosofía STATIS en el estudio de tablas múltiples de naturaleza cualitativa y, especialmente, referidas a distintos conjuntos de individuos.

Finalmente, en el capítulo 5 se trata el problema del análisis de tablas mixtas cuando una de las subtablas, la de frecuencias, está *dividida* en dos o más partes, lo que ocurre, por ejemplo, cuando los individuos usan idiomas distintos para expresarse de forma que hay corpus distintos para una misma temática, distinción debida al idioma empleado.

Con respecto a esta temática, creemos que es interesante proseguir en las siguientes líneas:

- Una comparativa entre los distintos pesos posibles para los individuos de las subtablas es conveniente, aunque no esperamos encontrar una solución claramente mejor que las demás.

- La relación entre una tabla de variables categóricas y una tabla de frecuencias puede establecerse por medio de otras técnicas, como, por ejemplo, el Análisis Canónico Parcial de Correspondencias de Ter Braak (1988), técnica relacionada al menos en concepto con el Análisis de Correlación Canónica de Hotelling.

Apéndices

APÉNDICE A

Encuesta web EHUdenda

Encuesta EHUdenda Inkesta

Encuesta EHUdenda Inkesta

Ongi etorri orri honetara. Itaun batzuk egin gura dizkizugu UPV/EHUz, haren irudiaz eta haren logotipoa daramaten zenbait artikuluz. Asko estimatuko dizugu galderok irakurtzeko tartetxo bat hartzea, eta ahalik eta zintzoen erantzutea.

Inkesta hau behin bakarrik bete daiteke. Mesedez, behin hasiz gero, ahalegindu zaitez amaitzen. 10 minutu baino ez dira. (Gainera, komeni da orri bakoitzaren behealdean ageri den next botoia ez sakatzea orria guztiz amaituta eduki arte; bestela, atzera joz gero nabigatzaileko atzera botoiaren bidez, berriz bete beharko duzu orri osoa, sistemak ez dizu-eta ezer gordeko).

Bienvenido a esta página. Vamos a realizar una serie de preguntas sobre la UPV/EHU, su imagen y algunos artículos con su logotipo. Le agradecemos que dedique un poco de su tiempo a contestar estas preguntas y que lo haga de la manera más veraz posible a su parecer.

Esta encuesta sólo se puede *rellenar una sóla vez*. Por favor, una vez que empiece, trate de disponer del tiempo suficiente para terminarla, unos 10 minutos. (Es conveniente, además, no pulsar el botón "next" del final de cada página hasta que ésta esté completamente rellenada; si no es así y después se usa el botón "back" del navegador para volver, es necesario rellenar toda la página otra vez.)

1 Aukeratu, zure iritziz, zelakoa den UPV/EHU:
En su opinión, la UPV/EHU es:

2 Aukeratu, zure iritziz, zelakoa den Deustuko Unibertsitatea:
En su opinión, la Universidad de Deusto es:

3 Aukeratu, zure iritziz, zelakoa den Mondragon Unibertsitatea:
En su opinión, la Universidad de Mondragón es:

Hurrengo galderok (4tik 13ra) Euskal Herriko Unibertsitatearen gainekoak dira, erakundea den aldetik begiratuta. Proposatzen dizkizugun puntuazioetako bat aukeratu behar duzu, 1etik 5era. 1 zenbakiak esan gura du ez nago batere ados eta 5 zenbakiak, berriz, guztiz ados nago. N/A da ez dakit edo erantzunik ez.

Las preguntas siguientes (4-13) son relativas a la Universidad del País Vasco como organización. Son preguntas de elección múltiple en una escala de 1 a 5, donde 1 significa *completamente en desacuerdo* y 5 *completamente de acuerdo*. N/A significa *no sabe* o *no contesta*.

4	UPV/EHU gizarteari irekita dagoen erakundea da La UPV/EHU es una institución abierta a la sociedad	1	2	3	4	5	N/A
5	UPV/EHUk zerbitzu baliagarria ematen dio gizarteari La UPV/EHU ofrece un servicio útil a la sociedad	1	2	3	4	5	N/A
6	UPV/EHUk kultura zabaltzen laguntzen du La UPV/EHU contribuye a la difusión de la cultura	1	2	3	4	5	N/A
7	UPV/EHUk aurrerabidean eta berrikuntzan laguntzen du La UPV/EHU contribuye al progreso e innovación	1	2	3	4	5	N/A
8	Irakaskuntza funtsezko zeregina da UPV/EHUn La docencia es una función fundamental en la UPV/EHU	1	2	3	4	5	N/A
9	Ikerketa funtsezko zeregina da UPV/EHUn La investigación es una función fundamental en la UPV/EHU	1	2	3	4	5	N/A
10	UPV/EHUri ematen zaizkion baliabideak urriak dira Los recursos asignados a la UPV/EHU son escasos	1	2	3	4	5	N/A
11	Hedabideek UPV/EHUren jardunari ematen dioten arreta urriegia da La cobertura de noticias sobre la actividad de la UPV/EHU en los medios de comunicación es escasa	1	2	3	4	5	N/A
12	Oro har, pozik nago UPV/EHUko kidea izateaz En general, estoy satisfecho de pertenecer a la UPV/EHU	1	2	3	4	5	N/A

Figura A.1: Página 1 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

13 Zelan uste duzu hobetu litekeela UPV/EHU, edozein alderditatik?
¿Cómo crees que podría mejorar la UPV/EHU en cualquiera de sus aspectos?

14 Erosiko zenuke UPV/EHUren logotipoa daraman produkturen bat, zeuk erabiltzeko edo beste inori oparitzeko? (Adibidez, iduneko zapiak, giltzak-eta uzteko platertxoa, kamisetak, erlojuak, katiluak...) Bai - Sí Ez - No

¿Estarías interesado en comprar un producto con el logotipo de la UPV/EHU (tal como pañuelos, vacíabolsillos, camisetas, relojes, tazas, ...) para uso personal o para regalo?

15 Azaldu hemen zergatik, mesedez.
¿Podrías escribir aquí por qué?

Orrialde honetan eta hurrengoan artikulu batzuk erakutsiko dizkizugu, UPV/EHUren logoa daramatenak. Bakoitzerako, adierazi, mesedez, ea zelako probabilitatea dagoen zuk hori erosteko, zeuretzat edo beste inori oparitzeko: oso aukera txikia, aukera txikia, aukera handia edo oso aukera handia.

A continuación, en esta página y en las siguientes, mostramos una serie de artículos, con el logotipo de la UPV/EHU. Para cada uno de ellos, le pedimos que valore como muy poco probable, poco probable, probable, y muy probable la probabilidad de que adquiera alguno de estos productos, ya sea para uso personal o para regalar.

16 Goardasola - Paraguas: (6,22) Oso aukera txikia - Muy poco probable | Aukera txikia - Poco probable | Aukera handia - Probable | Oso aukera handia - Muy probable

17 Giltzatakoa - Llavero: (4,55) Oso aukera txikia - Muy poco probable | Aukera txikia - Poco probable | Aukera handia - Probable | Oso aukera handia - Muy probable

Figura A.2: Página 2 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

18	Gorbata - Corbata: (21,82)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

19	Euritarako kapela - Gorro para lluvia: (5,10)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

20	Iduneko zapia - Pañuelo de señora: (35,17)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

21	Iduneko zapia - Pañuelo de señora: (27,38)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.3: Página 3 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

22	Giltzak-eta uzteko platertxoa, plastikozkoa - Vacía bolsillos de plástico (3,70)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable
23	Giltzak-eta uzteko platertxoa, larruzkoa - Vacía bolsillos de piel (6)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

24	Kamiseta - Camisetas (6,07)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

25	Kamiseta pikoduna, andreentzat - Camisetas (de cuello en V - chica) (7,91)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.4: Página 4 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

26	Txandaleko jaka - Chandal - sudadera (19,98)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

27	Kotoizko bisera - Gorras de algodón (2,32)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

28	Metxeroa, zilar kolorekoa - Mechero plateado (2,84)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

29	Zilarrezko pina - Pin de plata (14,57)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.5: Página 5 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

30	Erlojua, larruzko uhala daukana - Reloj de correa de piel (16,44)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

31	Erlojua, metalezko uhala daukana - Reloj de correa metálica (22,16)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

32	Diruzorroa - Billetero (18,95)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.6: Página 6 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

33	Motxila - Mochila (13,32)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

34	Eskegitzeko poltsa - Mochila de bandolera (11,37)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

35	Boligrafoa, urdina eta zilar kolorekoa - Bolígrafo azul y plata (2,55)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.7: Página 7 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

36	Boligrafo beltza bere kutxan - Bolígrafo negro con estuche (5,45)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

37	Boligrafoa, zilar kolorekoa - Bolígrafo plateado (14,22)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

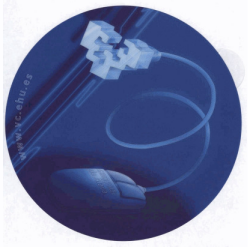
38	Zilar koloreko boligrafoa egurrezko kutxan - Bolígrafo plateado en estuche de madera (21,34)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.8: Página 8 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

39	Portzelanazko katiluak - Tazas de porcelana (2,16 c/u)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

40	Sagurako tapiza - Alfombrilla ratón	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable



41	Eskultura (UPV/EHUren logoa) - Escultura UPV/EHU (70,76)	Oso aukera txikia - Muy poco probable	Aukera txikia - Poco probable	Aukera handia - Probable	Oso aukera handia - Muy probable

Figura A.9: Página 9 de la encuesta EHUdenda.

Encuesta EHUdenda Inkesta

49tik 56ra bitarteko galderak orokorrak dira. UPV/EHUren logotipoa daraman produktu batek izan litzakeen ezaugarri batzuk aipatzen dizkizugu. Eman puntuak bakoitzari (gutxienez 1 eta gehienez 7), bakoitzari ematen diozun garrantzia adierazteko.

Las preguntas de la 49 a la 56 son de tipo general. Por favor, de una valoración con puntuaciones del 1 (valor mínimo) al 7 (valor máximo) las características que, de entre las siguientes, debería tener un producto con el logotipo de la UPV/EHU.

42	Orijinala - Original	1	2	3	4	5	6	7
43	Ausarta - Audaz o atrevido	1	2	3	4	5	6	7
44	Praktikoa - Práctico	1	2	3	4	5	6	7
45	Tradizionala - Tradicional	1	2	3	4	5	6	7
46	Artistikoa - Artístico	1	2	3	4	5	6	7
47	Dotorea - Elegante	1	2	3	4	5	6	7
48	Serioa - Serio	1	2	3	4	5	6	7
49	Modernoa - Moderno	1	2	3	4	5	6	7

Eskerrik asko, ia-ia amaitu duzu. Bukatzeko, erantzun mesedez, zure ezaugarriei buruzko galdera hauei. Informazio hau inkesta aztertzeko baino ez dugu erabiliko, ez dugu inora eramango eta ez dugu beste ezertarako baliatuko.

Muchas gracias, está llegando al final de la encuesta. Para finalizar, le rogamos que nos conteste a unas pocas preguntas de caracterización. Esta información sólo será usada en el análisis de la encuesta, sin ser transferida a ningún otro sitio ni utilizada para otro propósito que el mencionado.

50 Generoa - Género: Gizonezkoa - Masculino Emakumezkoa - Femenino

51 Adina (aukeratu tarte bat):
Edad (elija un intervalo):

52 Eginak (amaituta) dituzun maila handieneko ikasketak hauek dira:
Su máximo nivel de estudios alcanzado (terminado) es:

Mesedez, *ikaslea bazara*, erantzun hurrengo galdera bi hauei (AZPkoa edo irakaslea bazara, ez erantzun)

Por favor responda a las dos preguntas siguientes sólo *si es alumno* (y en ningún caso PAS o PDI):

53 Ikasle elkarteren bateko kidea zara? (Ikasleen Kontseilua, Fecem junior, aiesec, besterik...):
¿Perteneces a alguna asociación estudiantil? (Consejo de estudiantes, Fecem junior, aiesec, otras, ...):

54 Ikasteaz gain, ordaindutako lanen bat egiten duzu?
Además de estudiar, ¿tiene un trabajo remunerado?

Figura A.10: Página 10 de la encuesta EHUdenda.

APÉNDICE B

Tablas y figuras del capítulo 5

B.1. Análisis de Correspondencias Múltiples de las tablas apiladas

B.1.1. ACM

BuyLogo				
Categorías	Efectivo inicial	Peso inicial	Efectivo final	Peso final
BuyLo=1	963	118,78	972	119,73
BuyLo=2	567	78,39	575	79,47
ausentes	17	2,03	reasig.	
Satis2				
Categorías	Efectivo inicial	Peso inicial	Efectivo final	Peso final
Satis2 = 1	373	46,51	377	47,15
Satis2 = 2	660	84,38	664	85,05
Satis2 = 3	499	65,62	506	67,00
ausentes	15	2,69	reasig.	

Tabla B.1: Variables categóricas activas: efectivos y pesos, antes y después de la reasignación aleatoria de valores ausentes.

Categoría	Efectivo	Peso absoluto	Dist. al orig.	Eje 1	Eje 2	Eje 3
Students	509	75,55	1,63667	0,22	0,05	0,03
Admin Staff	371	35,43	4,62236	-0,03	0,03	0,06
Teaching Staff	667	88,22	1,25800	-0,18	-0,06	-0,05
Masc.	705	88,54	1,24984	0,02	-0,08	0,09
Femen.	842	110,66	0,80011	-0,01	0,07	-0,07
Araba	232	34,57	4,76223	0,16	-0,03	-0,03
Bizkaia	928	108,44	0,83696	-0,09	0,00	0,06
Gipuzkoa	387	56,19	2,54512	0,07	0,01	-0,10
Age=1	287	44,09	3,51804	0,11	0,09	0,02
Age=2	208	32,54	5,12170	0,41	-0,09	0,13
Age=3	571	76,54	1,60256	-0,09	-0,01	-0,11
Age=4	457	43,85	3,54277	-0,26	-0,02	0,08
missing cat.	24	2,18	90,37630	0,19	0,42	-0,07
Euskera	304	100,06	0,99081	0,03	-0,06	0,00
Cast.	1243	99,14	1,00928	-0,03	0,06	0,00

Tabla B.2: Coordenadas de las categorías suplementarias.

B.1.2. Clasificación sobre los ejes del ACM

Cluster	Efectivo	Peso
1	441	54,37
2	223	30,68
3	506	67,00
4	377	47,15

Tabla B.3: Composición inicial de clusters.

Clusters			Valores test			Coordenadas			Distancia
Cluster	Efec.	Peso Ab.	1	2	3	1	2	3	al origen
1/4	441	54,37	16,6	28,0	11,6	0,54	0,80	0,28	1,00
2/4	223	30,68	9,7	18,2	20,3	0,48	0,80	0,74	1,42
3/4	506	67,00	19,5	29,9	14,4	0,57	0,77	0,31	1,02
4/4	377	47,15	31,1	10,8	22,9	1,12	0,34	0,61	1,74

Tabla B.4: Coordenadas y Valores test antes de la consolidación. Ejes 1 a 3.

Iteración	Inercia Total	Inercia Inter-clusters	Proporción
0	1,50000	1,24957	0.83305
1	1,50000	1,24957	0.83305
2	1,50000	1,24957	0.83305

Tabla B.5: Fase de consolidación de la partición en 4 clusters.

Inercias	Inercias		Efectivo		Pesos		Distancias	
	Antes	Desp.	Antes	Desp.	Antes	Desp.	Antes	Desp.
Inter Clusters	1,2496	1,2496						
Intra Cluster								
Cluster 1/4	0	0	441	441	54,37	54,37	1,0030	1,0030
Cluster 2/4	0	0	223	223	30,68	30,68	1,4244	1,4244
Cluster 3/4	0,1383	0,1383	506	506	67,00	67,00	1,0212	1,0212
Cluster 4/4	0,1121	0,1121	377	377	47,15	47,15	1,7448	1,7448
Inercia Total	1,5000	1,5000						

Proporción (Inercia Inter / Inercia Total): Antes 0,8330
Después 0,8330

Tabla B.6: Descomposición de la inercia computada sobre los 3 ejes.

Clusters			Valores test			Coordenadas			Distancia
Cluster	Efec.	Peso Ab.	1	2	3	1	2	3	al origen
1/4	441	54,37	16,6	28,0	11,6	0,54	0,80	0,28	1,00
2/4	223	30,68	9,7	18,2	20,3	0,48	0,80	0,74	1,42
3/4	506	67,00	19,5	29,9	14,4	0,57	0,77	0,31	1,02
4/4	377	47,15	31,1	10,8	22,9	1,12	0,34	0,61	1,74

Tabla B.7: Valores test y coordenadas tras la consolidación, ejes 1 a 3.

Cluster 1/4 (Efectivo: 54 - Porcentaje: 27.29)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
Satis2 = 2	99,63	42,36	64,20	10,77	84
BuyLo=1	99,37	59,63	45,49	7,91	119
Teaching Staff	53,83	44,29	33,18	1,48	88
Male	52,99	44,45	32,54	1,39	89
Bizkaia	58,27	54,44	29,21	0,70	108
Age=4	26,17	22,01	32,45	0,61	44
Age=3	40,02	38,42	28,43	0,20	77
Age=2	16,57	16,34	27,69	0,17	33
Euskera	51,48	50,23	27,97	0,12	100
Cast	48,52	49,77	26,61	-0,12	99
Gipuzkoa	24,04	28,21	23,26	-0,59	56
Age=1	16,85	22,13	20,78	-0,94	44
Female	47,01	55,55	23,10	-1,16	111
Students	29,58	37,93	21,28	-1,36	76
Satis2 = 1	0,00	23,35	0,00	-5,43	47
Satis2 = 3	0,00	32,94	0,00	-6,88	66
BuyLo=2	0,00	39,35	0,00	-7,77	78

Tabla B.8: Cluster 1: Categorías características.

Cluster 2/4 (Efectivo: 31 - Porcentaje: 15.40)					
Categorías Características	% categ. en grupo	% categ. en total	% del grupo en categ.	Valor test	Peso
BuyLo=2	99,28	39,35	38,86	7,24	78
Satis2 = 2	98,47	42,36	35,80	6,87	84
Gipuzkoa	34,58	28,21	18,88	0,78	56
Female	61,41	55,55	17,03	0,47	111
Admin Staff	17,86	17,79	15,47	0,46	35
Eusk	55,96	50,23	17,16	0,36	100
Students	41,66	37,93	16,92	0,28	76
Age=3	40,94	38,42	16,41	0,21	77
Age=1	23,83	22,13	16,58	0,14	44
ARABA	17,08	17,35	15,16	-0,06	35
Age=4	18,48	22,01	12,93	-0,14	44
Age=2	15,94	16,34	15,03	-0,23	33
Cast	44,04	49,77	13,63	-0,36	99
Teaching Staff	40,48	44,29	14,08	-0,47	88
Bizkaia	48,34	54,44	13,68	-0,52	108
Male	38,59	44,45	13,37	-0,53	89
Satis2 = 1	0,00	23,35	0,00	-3,71	47
Satis2 = 3	0,00	32,94	0,00	-4,76	66
BuyLo=1	0,00	59,63	0,00	-7,69	119

Tabla B.9: Cluster 2: Categorías características.

Cluster 3/4 (Efectivo: 67 - Porcentaje: 33.63)					
Categorías	% categ.	% categ.	% del grupo	Valor	Peso
Características	en grupo	en total	en categ.	test	
Satis2 = 3	97,94	32,94	100,00	15,21	66
BuyLo=1	72,07	59,63	40,66	2,30	119
Female	62,31	55,55	37,73	1,25	111
Age=3	42,55	38,42	37,25	0,79	77
Cast	53,88	49,77	36,41	0,65	99
Age=4	24,76	22,01	37,83	0,62	44
Teaching Staff	48,27	44,29	36,66	0,57	88
Bizkaia	56,31	54,44	34,79	0,34	108
Gipuzkoa	28,69	28,21	34,21	0,11	56
Age=1	22,36	22,13	33,98	0,10	44
Admin Staff	18,33	17,79	34,66	0,10	35
Araba	15,00	17,35	29,07	-0,49	35
Eusk	46,12	50,23	30,88	-0,65	100
Students	33,40	37,93	29,62	-0,95	76
Male	37,69	44,45	28,52	-1,35	89
Age=2	8,84	16,34	18,19	-1,91	33
BuyLo=2	26,34	39,35	22,52	-2,41	78
Satis2 = 1	0,00	23,35	0,00	-6,32	47
Satis2 = 2	0,00	42,36	0,00	-9,56	84

Tabla B.10: Cluster 3: Categorías características.

Cluster 4/4 (Efectivo: 47 - Porcentaje: 23.67)					
Categorías	% categ.	% categ.	% del grupo	Valor	Peso
Características	en grupo	en total	en categ.	test	
Satis2 = 1	98,64	23,35	100,00	99,99	47
BuyLo=2	64,22	39,35	38,63	3,75	78
Age=2	26,98	16,34	39,09	2,05	33
Students	51,56	37,93	32,18	1,89	76
Age=1	26,81	22,13	28,67	0,85	44
Araba	20,49	17,35	27,94	0,56	35
Male	48,02	44,45	25,57	0,50	89
Admin Staff	18,35	17,79	24,41	0,13	35
Eusk	50,90	50,23	23,99	0,04	100
Cast	49,10	49,77	23,35	-0,04	99
Gipuzkoa	28,19	28,21	23,65	-0,12	56
Female	51,98	55,55	22,15	-0,24	111
Bizkaia	51,33	54,44	22,32	-0,34	108
Age=4	15,61	22,01	16,78	-1,17	44
Age=3	29,08	38,42	17,91	-1,27	77
Teaching Staff	30,10	44,29	16,08	-2,13	88
BuyLo=1	34,91	59,63	13,86	-3,93	119
Satis2 = 3	0,00	32,94	0,00	-6,25	66
Satis2 = 2	0,00	42,36	0,00	-7,47	84

Tabla B.11: Cluster 4: Categorías características.

B.2. Análisis Factorial Múltiple de tablas mixtas

B.2.1. AFM de las tablas completadas con ceros

Coeficientes $\mathcal{L}_g$ de asociación entre subtablas

	Tabla 1	Tabla 2	Tabla 3	Tabla 4	Todas
Tabla 1	1,9034				
Tabla 2	0,3060	20,1651			
Tabla 3	0,1857	0,0000	18,3991		
Tabla 4	0,0576	0,4155	0,1737	2,6276	
Todas	1,6050	13,7180	12,4539	0,4334	18,6137

Coeficientes RV de asociación entre subtablas

	Tabla 1	Tabla 2	Tabla 3	Tabla 4	Todas
Tabla 1	1,0000				
Tabla 2	0,0494	1,0000			
Tabla 3	0,0314	0,0000	1,0000		
Tabla 4	0,0258	0,0571	0,0250	1,0000	
Todas	0,2696	0,7081	0,6730	0,0620	1,0000

Tabla B.12: Medidas de asociación entre las tablas del AFM mixto, incluída la tabla de variables categóricas suplementarias.

Variable	Categorías y grupo	Peso rel.	Dist. al orig.	eje 1	eje 2	eje 3	eje 4	eje 5
BuyLogo								
	BuyLo=1	60,0400	0,8183	0,8486	0,0415	0,0209	0,0608	-0,0824
	Grupo 1		5,0337	1,5690	0,0663	0,1660	0,2085	-0,0371
	Grupo 2		1,2927	0,5208	-0,4619	0,0300	0,2155	-0,1238
	Grupo 3		1,0386	0,4561	0,5202	-0,1333	-0,2416	-0,0863
	BuyLo=2	39,9598	1,8474	-1,2751	-0,0624	-0,0314	-0,0913	0,1238
	Grupo 1		11,3637	-2,3575	-0,0996	-0,2494	-0,3132	0,0557
	Grupo 2		2,9183	-0,7824	0,6940	-0,0451	-0,3238	0,1860
	Grupo 3		2,3447	-0,6853	-0,7816	0,2003	0,3630	0,1297
Satis2								
	Satis2 = 1	23,9307	3,0704	-1,1115	-0,0438	-0,7329	-0,1809	-0,4705
	Grupo 1		23,9047	-2,5393	-0,1596	-1,8443	-0,5497	-1,1217
	Grupo 2		2,5961	-0,4555	0,3618	-0,3820	-0,1267	-0,2062
	Grupo 3		1,1327	-0,3398	-0,3336	0,0278	0,1338	-0,0837
	Satis2 = 2	42,6003	1,1897	0,0919	0,3060	0,9848	0,0768	0,0154
	Grupo 1		9,4403	0,2679	0,7152	2,4409	0,2081	0,0204
	Grupo 2		0,9899	-0,0235	0,1580	0,4225	0,0207	0,0301
	Grupo 3		0,2768	0,0312	0,0449	0,0911	0,0016	-0,0042
	Satis2 = 3	33,4688	1,8495	0,6778	-0,3582	-0,7295	0,0316	0,3168
	Grupo 1		14,3316	1,4746	-0,7962	-1,7881	0,1282	0,7761
	Grupo 2		1,6972	0,3556	-0,4598	-0,2646	0,0642	0,1091
	Grupo 3		0,6167	0,2033	0,1814	-0,1358	-0,0976	0,0652

Tabla B.13: Coordenadas de los centros de gravedad de las categorías activas.

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
BuyLogo		72,5113	0,2570	0,0699	0,6023	1,1708
	BuyLo=1	0,2898	0,0010	0,0003	0,0024	0,0047
	Grupo 1	0,0776	0,0001	0,0033	0,0024	0,0003
	Grupo 2	0,0161	0,0264	0,0000	0,0026	0,0002
	Grupo 3	0,0230	0,0239	0,0038	0,0099	0,0000
	BuyLo=2	0,4354	0,0015	0,0004	0,0036	0,0070
	Grupo 1	0,1166	0,0001	0,0050	0,0035	0,0004
	Grupo 2	0,0241	0,0397	0,0000	0,0039	0,0003
	Grupo 3	0,0346	0,0359	0,0056	0,0149	0,0000
Satis2		30,3576	8,2661	76,5752	1,1576	9,9510
	Satis2 = 1	0,1981	0,0005	0,1367	0,0085	0,0608
	Grupo 1	0,1215	0,0006	0,0778	0,0059	0,0211
	Grupo 2	0,0256	0,0068	0,0078	0,0001	0,0035
	Grupo 3	0,0355	0,0035	0,0364	0,0043	0,0074
	Satis2 = 2	0,0024	0,0396	0,4395	0,0027	0,0001
	Grupo 1	0,0033	0,0124	0,2377	0,0013	0,0000
	Grupo 2	0,0014	0,0016	0,0355	0,0002	0,0000
	Grupo 3	0,0004	0,0050	0,0896	0,0004	0,0000
	Satis2 = 3	0,1030	0,0426	0,1895	0,0004	0,0386
	Grupo 1	0,0529	0,0112	0,0987	0,0006	0,0147
	Grupo 2	0,0087	0,0006	0,0190	0,0001	0,0030
	Grupo 3	0,0188	0,0169	0,0311	0,0010	0,0044

Tabla B.14: Contribuciones de los centros de gravedad de las categorías activas.

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
BuyLogo						
	BuyLo=1	0,8801	0,0021	0,0005	0,0045	0,0083
	Grupo 1	0,4891	0,0009	0,0055	0,0086	0,0003
	Grupo 2	0,2098	0,1650	0,0007	0,0359	0,0119
	Grupo 3	0,2003	0,2605	0,0171	0,0562	0,0072
	BuyLo=2	0,8801	0,0021	0,0005	0,0045	0,0083
	Grupo 1	0,4891	0,0009	0,0055	0,0086	0,0003
	Grupo 2	0,2098	0,1650	0,0007	0,0359	0,0119
	Grupo 3	0,2003	0,2605	0,0171	0,0562	0,0072
Satis2						
	Satis2 = 1	0,4024	0,0006	0,1749	0,0107	0,0721
	Grupo 1	0,2697	0,0011	0,1423	0,0126	0,0526
	Grupo 2	0,0799	0,0504	0,0562	0,0062	0,0164
	Grupo 3	0,1019	0,0983	0,0007	0,0158	0,0062
	Satis2 = 2	0,0071	0,0787	0,8153	0,0050	0,0002
	Grupo 1	0,0076	0,0542	0,6311	0,0046	0,0000
	Grupo 2	0,0006	0,0252	0,1804	0,0004	0,0009
	Grupo 3	0,0035	0,0073	0,0300	0,0000	0,0001
	Satis2 = 3	0,2484	0,0694	0,2878	0,0005	0,0543
	Grupo 1	0,1517	0,0442	0,2231	0,0011	0,0420
	Grupo 2	0,0745	0,1245	0,0413	0,0024	0,0070
	Grupo 3	0,0670	0,0533	0,0299	0,0154	0,0069

Tabla B.15: Cosenos cuadrado de los centros de gravedad de las categorías activas.

Variable	Categorías y grupo	Peso rel.	Dist. al orig.	eje 1	eje 2	eje 3	eje 4	eje 5
Vinculación								
	Estudiantes	37,9266	0,4024	-0,2411	-0,0072	-0,1647	0,1333	-0,0467
	Grupo 1		0,4289	-0,4376	-0,0240	-0,1968	-0,0783	-0,1011
	Grupo 2		2,1056	-0,0735	0,2096	-0,3383	0,3967	-0,0044
	Grupo 3		1,0870	-0,2121	-0,2071	0,0410	0,0817	-0,0347
	PAS	17,7861	0,4389	-0,0356	-0,0234	-0,0208	0,0980	-0,0196
	Grupo 1		0,0426	0,0649	-0,0136	-0,0901	-0,0012	-0,0368
	Grupo 2		1,4896	-0,0595	0,2131	-0,0466	0,1209	-0,0861
	Grupo 3		2,4178	-0,1122	-0,2697	0,0743	0,1743	0,0642
	PDI	44,2871	0,3283	0,2208	0,0155	0,1494	-0,1536	0,0479
	Grupo 1		0,3186	0,3487	0,0260	0,2047	0,0676	0,1014
	Grupo 2		1,8232	0,0868	-0,2651	0,3084	-0,3883	0,0384
	Grupo 3		0,8128	0,2268	0,2857	-0,0650	-0,1400	0,0039
Género								
	Masc.	44,4477	0,1726	0,0268	0,0535	0,0574	-0,0076	0,0306
	Grupo 1		0,0938	0,0028	0,0679	0,1405	0,0038	-0,0753
	Grupo 2		1,0504	0,0022	-0,0417	0,0629	0,0621	-0,0133
	Grupo 3		0,4095	0,0754	0,1342	-0,0313	-0,0887	0,1803
	Fem.	55,5521	0,1105	-0,0214	-0,0428	-0,0459	0,0061	-0,0245
	Grupo 1		0,0600	-0,0022	-0,0543	-0,1124	-0,0031	0,0602
	Grupo 2		0,6724	-0,0017	0,0333	-0,0503	-0,0497	0,0106
	Grupo 3		0,2621	-0,0603	-0,1074	0,0250	0,0710	-0,1443

Tabla B.16: Coordenadas de los centros de gravedad de las categorías suplementarias (`vinculación, género`).

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Vinculación		2,9389	0,0222	2,1540	2,0492	0,2195
	Estudiantes	0,0148	0,0000	0,0109	0,0073	0,0010
	Grupo 1	0,0036	0,0000	0,0001	0,0031	0,0002
	Grupo 2	0,0027	0,0031	0,0030	0,0047	0,0001
	Grupo 3	0,0001	0,0026	0,0042	0,0002	0,0000
	PAS	0,0002	0,0001	0,0001	0,0019	0,0001
	Grupo 1	0,0004	0,0000	0,0002	0,0003	0,0000
	Grupo 2	0,0000	0,0017	0,0000	0,0000	0,0002
	Grupo 3	0,0003	0,0019	0,0004	0,0002	0,0003
	PDI	0,0145	0,0001	0,0105	0,0113	0,0012
	Grupo 1	0,0018	0,0000	0,0004	0,0039	0,0003
	Grupo 2	0,0020	0,0061	0,0029	0,0044	0,0000
	Grupo 3	0,0000	0,0056	0,0054	0,0000	0,0002
Género		0,0384	0,2271	0,2803	0,0050	0,0859
	Masc.	0,0002	0,0013	0,0016	0,0000	0,0005
	Grupo 1	0,0001	0,0000	0,0008	0,0000	0,0010
	Grupo 2	0,0001	0,0007	0,0000	0,0004	0,0002
	Grupo 3	0,0003	0,0005	0,0009	0,0005	0,0021
	Fem.	0,0002	0,0010	0,0012	0,0000	0,0004
	Grupo 1	0,0001	0,0000	0,0006	0,0000	0,0008
	Grupo 2	0,0001	0,0006	0,0000	0,0003	0,0001
	Grupo 3	0,0002	0,0004	0,0007	0,0004	0,0017

Tabla B.17: Contribuciones de los centros de gravedad de las categorías suplementarias (`vinculación, género`).

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Vinculación						
	Estudiantes	0,1445	0,0001	0,0674	0,0442	0,0054
	Grupo 1	0,4466	0,0013	0,0903	0,0143	0,0238
	Grupo 2	0,0026	0,0209	0,0544	0,0747	0,0000
	Grupo 3	0,0414	0,0395	0,0015	0,0061	0,0011
	PAS	0,0029	0,0012	0,0010	0,0219	0,0009
	Grupo 1	0,0989	0,0043	0,1907	0,0000	0,0318
	Grupo 2	0,0024	0,0305	0,0015	0,0098	0,0050
	Grupo 3	0,0052	0,0301	0,0023	0,0126	0,0017
	PDI	0,1485	0,0007	0,0680	0,0718	0,0070
	Grupo 1	0,3816	0,0021	0,1316	0,0143	0,0322
	Grupo 2	0,0041	0,0385	0,0522	0,0827	0,0008
	Grupo 3	0,0633	0,1004	0,0052	0,0241	0,0000
Género						
	Masc.	0,0042	0,0166	0,0191	0,0003	0,0054
	Grupo 1	0,0001	0,0491	0,2105	0,0002	0,0604
	Grupo 2	0,0000	0,0017	0,0038	0,0037	0,0002
	Grupo 3	0,0139	0,0440	0,0024	0,0192	0,0794
	Fem.	0,0042	0,0166	0,0191	0,0003	0,0054
	Grupo 1	0,0001	0,0491	0,2105	0,0002	0,0604
	Grupo 2	0,0000	0,0017	0,0038	0,0037	0,0002
	Grupo 3	0,0139	0,0440	0,0024	0,0192	0,0794

Tabla B.18: Cosenos cuadrado de los centros de gravedad de las categorías suplementarias (`vinculación, género`).

Variable	Categorías y grupo	Peso rel.	Dist. al orig.	eje 1	eje 2	eje 3	eje 4	eje 5
Prov	ARABA	17,3544	0,4934	-0,1613	-0,0068	-0,0096	-0,0746	0,0260
	Grupo 1		0,2318	-0,3486	0,0181	0,0069	-0,0481	-0,0487
	Grupo 2		3,3916	-0,0837	0,0058	-0,0656	-0,1878	0,0150
	Grupo 3		0,8172	-0,0517	-0,0442	0,0299	0,0122	0,1118
	BIZKAIA	54,4376	0,1081	0,1243	-0,0647	0,0206	-0,0297	-0,0036
	Grupo 1		0,0749	0,1921	0,0042	0,0085	0,0248	-0,0033
	Grupo 2		0,7224	0,1518	-0,2421	0,0711	-0,0892	-0,0083
	Grupo 3		0,1752	0,0290	0,0436	-0,0177	-0,0247	0,0008
	GIPUZKOA	28,2078	0,3191	-0,1406	0,1291	-0,0339	0,1032	-0,0091
	Grupo 1		0,0753	-0,1562	-0,0192	-0,0206	-0,0183	0,0363
	Grupo 2		2,3318	-0,2414	0,4636	-0,0968	0,2877	0,0067
	Grupo 3		0,4650	-0,0242	-0,0570	0,0158	0,0402	-0,0702
Edad	Edad=1	22,5502	0,6075	-0,1635	0,0270	-0,2014	0,1387	-0,0368
	Grupo 1		0,1575	-0,2145	-0,0538	-0,2277	-0,0469	-0,0404
	Grupo 2		4,0744	-0,1049	0,3014	-0,4255	0,3887	0,0217
	Grupo 3		1,2354	-0,1712	-0,1667	0,0490	0,0743	-0,0917
	Edad=2	16,4558	0,9307	-0,3718	-0,0304	-0,0224	0,2324	-0,1046
	Grupo 1		1,6476	-0,8414	0,0839	-0,0140	-0,1303	-0,2580
	Grupo 2		5,0829	0,0242	0,1304	-0,1155	0,7023	-0,0433
	Grupo 3		1,6459	-0,2982	-0,3054	0,0623	0,1253	-0,0124
	Edad=3	38,6446	0,2567	0,0698	-0,0547	0,0997	-0,1703	0,0734
	Grupo 1		0,1560	0,1517	-0,0154	0,0828	0,0350	0,1152
	Grupo 2		1,6749	0,0054	-0,1981	0,2371	-0,5218	0,0296
	Grupo 3		0,4798	0,0522	0,0494	-0,0207	-0,0240	0,0754
	Edad=4	22,3494	0,4674	0,3181	0,0897	0,0473	-0,0167	-0,0128
	Grupo 1		0,6270	0,5737	0,0192	0,0969	0,0826	0,0315
	Grupo 2		1,9266	0,0788	-0,0577	0,1045	-0,0071	-0,0412
	Grupo 3		1,6534	0,3020	0,3077	-0,0595	-0,1257	-0,0287

Tabla B.19: Coordenadas de los centros de gravedad de las categorías suplementarias (`campus, edad`).

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Prov		1,2402	0,6939	0,0608	0,4828	0,0170
	ARABA	0,0030	0,0000	0,0000	0,0010	0,0001
	Grupo 1	0,0015	0,0000	0,0000	0,0000	0,0002
	Grupo 2	0,0003	0,0000	0,0001	0,0004	0,0000
	Grupo 3	0,0005	0,0000	0,0001	0,0002	0,0003
	BIZKAIA	0,0056	0,0023	0,0002	0,0005	0,0000
	Grupo 1	0,0006	0,0004	0,0000	0,0003	0,0000
	Grupo 2	0,0001	0,0030	0,0004	0,0003	0,0000
	Grupo 3	0,0012	0,0011	0,0002	0,0000	0,0000
	GIPUZKOA	0,0037	0,0047	0,0003	0,0033	0,0000
	Grupo 1	0,0000	0,0011	0,0000	0,0008	0,0001
	Grupo 2	0,0007	0,0055	0,0003	0,0017	0,0000
	Grupo 3	0,0010	0,0017	0,0002	0,0002	0,0002
Edad		3,5705	0,3246	1,4440	2,6565	0,4849
	Edad=1	0,0040	0,0002	0,0097	0,0047	0,0004
	Grupo 1	0,0001	0,0003	0,0000	0,0014	0,0000
	Grupo 2	0,0002	0,0030	0,0030	0,0025	0,0002
	Grupo 3	0,0000	0,0015	0,0037	0,0002	0,0001
	Edad=2	0,0152	0,0002	0,0001	0,0096	0,0021
	Grupo 1	0,0090	0,0004	0,0000	0,0039	0,0008
	Grupo 2	0,0064	0,0007	0,0004	0,0066	0,0001
	Grupo 3	0,0002	0,0022	0,0003	0,0003	0,0003
	Edad=3	0,0013	0,0011	0,0041	0,0121	0,0024
	Grupo 1	0,0006	0,0001	0,0000	0,0029	0,0001
	Grupo 2	0,0004	0,0014	0,0019	0,0086	0,0002
	Grupo 3	0,0000	0,0007	0,0015	0,0015	0,0000
	Edad=4	0,0152	0,0018	0,0005	0,0001	0,0000
	Grupo 1	0,0036	0,0002	0,0001	0,0004	0,0001
	Grupo 2	0,0032	0,0008	0,0002	0,0000	0,0000
	Grupo 3	0,0000	0,0018	0,0007	0,0005	0,0000

Tabla B.20: Contribuciones de los centros de gravedad de las categorías suplementarias (`campus, edad`).

Variable	Categorías y grupo	eje 1	eje 2	eje 3	eje 4	eje 5
Prov	ARABA	0,0527	0,0001	0,0002	0,0113	0,0014
	Grupo 1	0,5240	0,0014	0,0002	0,0100	0,0102
	Grupo 2	0,0021	0,0000	0,0013	0,0104	0,0001
	Grupo 3	0,0033	0,0024	0,0011	0,0002	0,0153
	BIZKAIA	0,1430	0,0388	0,0039	0,0082	0,0001
	Grupo 1	0,4923	0,0002	0,0010	0,0082	0,0001
	Grupo 2	0,0319	0,0811	0,0070	0,0110	0,0001
	Grupo 3	0,0048	0,0109	0,0018	0,0035	0,0000
	GIPUZKOA	0,0620	0,0522	0,0036	0,0334	0,0003
	Grupo 1	0,3242	0,0049	0,0057	0,0044	0,0175
	Grupo 2	0,0250	0,0922	0,0040	0,0355	0,0000
	Grupo 3	0,0013	0,0070	0,0005	0,0035	0,0106
Edad	Edad=1	0,0440	0,0012	0,0668	0,0317	0,0022
	Grupo 1	0,2922	0,0184	0,3293	0,0139	0,0104
	Grupo 2	0,0027	0,0223	0,0444	0,0371	0,0001
	Grupo 3	0,0237	0,0225	0,0019	0,0045	0,0068
	Edad=2	0,1485	0,0010	0,0005	0,0581	0,0117
	Grupo 1	0,4296	0,0043	0,0001	0,0103	0,0404
	Grupo 2	0,0001	0,0033	0,0026	0,0970	0,0004
	Grupo 3	0,0540	0,0567	0,0024	0,0095	0,0001
	Edad=3	0,0190	0,0117	0,0387	0,1129	0,0210
	Grupo 1	0,1475	0,0015	0,0440	0,0079	0,0851
	Grupo 2	0,0000	0,0234	0,0336	0,1625	0,0005
	Grupo 3	0,0057	0,0051	0,0009	0,0012	0,0118
	Edad=4	0,2165	0,0172	0,0048	0,0006	0,0004
	Grupo 1	0,5249	0,0006	0,0150	0,0109	0,0016
	Grupo 2	0,0032	0,0017	0,0057	0,0000	0,0009
	Grupo 3	0,0552	0,0573	0,0021	0,0096	0,0005

Tabla B.21: Cosenos cuadrado de los centros de gravedad de las categorías suplementarias (`campus, edad`).

B.2.2. Clasificación sobre los ejes del AFM de tablas mixtas

Cluster	Efectivo	Peso
1	436	54.20
2	355	46.29
3	187	31.35
4	569	67.36

Tabla B.22: Partición en 4 clases. Formación de clusters.

Clusters			Valores test			Coordenadas			Distancia
Cluster	Efec.	Peso Ab.	1	2	3	1	2	3	al origen
1/4	436	54,20	15,2	9,3	24,5	0,62	0,38	1,00	1,65
2/4	355	46,29	22,4	-6,1	-16,6	1,05	-0,29	-0,77	2,28
3/4	187	31,35	-15,2	10,7	8,1	-1,04	0,74	0,56	2,46
4/4	569	67,36	-21,9	-13,5	-15,8	-0,73	-0,45	-0,53	1,26

Tabla B.23: Valores test y coordenadas antes de la consolidación.

Iteración	Inercia total	Inercia entre clusters	Proporción
0	3,44004	1,79188	0,52089
1	3,44004	1,93220	0,56168
2	3,44004	1,95118	0,56720
3	3,44004	1,95534	0,56841
4	3,44004	1,95634	0,56870
5	3,44004	1,95653	0,56875

Tabla B.24: Consolidación de la partición en 4 clases. 10 iteraciones.

Inercias	Inercias		Efectivo		Pesos		Distancias	
	Antes	Desp.	Antes	Desp.	Antes	Desp.	Antes	Desp.
Inter Clusters	1,7919	1,9565						
Intra Cluster								
Cluster 1/4	0,2827	0,4436	436	461	54,20	59,02	1,6464	1,5658
Cluster 2/4	0,4004	0,3928	355	485	46,29	60,78	2,2780	1,7389
Cluster 3/4	0,3095	0,2845	187	184	31,35	32,33	2,4613	2,6136
Cluster 4/4	0,6555	0,3625	569	417	67,36	47,07	1,2633	2,2761
Inercia Total	3,4400	3,4400						

Proporción (Inercia Inter / Inercia Total): Antes 0,5209
Después 0,5688

Tabla B.25: Descomposición de la inercia computada sobre los 3 ejes.

Clusters			Valores test			Coordenadas			Distancia
Cluster	Efec.	Peso Ab.	1	2	3	1	2	3	al origen
1/4	461	59,02	14,6	6,1	26,8	0,57	0,24	1,05	1,57
2/4	485	60,78	22,2	-4,3	-22,4	0,84	-0,16	-0,84	1,74
3/4	184	32,33	-15,8	12,0	5,5	-1,09	0,83	0,38	2,61
4/4	417	47,07	-24,9	-15,8	-11,5	-1,04	-0,66	-0,48	2,28

Tabla B.26: Valores test y coordenadas después de la consolidación.

	1	2	3	4
1	0			
2	1,904	0		
3	2,211	2,820	0	
4	2,621	2,373	1,716	0

Tabla B.27: Matriz de distancias entre clusters.

B.2.3. Caracterización de la partición sobre los ejes principales del AFM. Respuestas modales

Respuestas modales a partir del criterio χ^2: Castellano

Cluster 1/4	
Criterio	Respuesta
0,904	me parece que es una bonita forma de que otros colegas de fuera de esta universidad tengan un recuerdo de aqui. ademas siempre es un recurso con el que quedas bien en cualquier situacion
0,913	si en la medida que estos fueran de buena calidad y diseno atractivo, donde el distintivo - logotipo de la upv-ehu, segun para el articulo en que se aplique, tenga una referencia mas directa o sutil, referencia desarrollada en un buen analisis y estudio de diseno grafico, industrial e imagen corporativa para cada producto.(por ejemplo en la imagen del paraguas el tamano del logotipo a mi parecer es demasiado grande para utilizarla tal cual, por lo que la referencia es desproporcionada. otra cosa es que tubiera menos tamano, o que teniendo mas tamano la referencia fuera no ta directa, es decir mas sutil. como todo la idea de generar una serie de articulos a nuestra universidad, es una idea muy buena, pero se puede llevar a cabo bien, regular o mal.
0,915	es un detalle, un recuerdo del sitio donde trabajo, un lugar al que estimo y aprecio, siempre que sean bonitos, asequibles y bien disenados p.e. con aportaciones de los alumnos de bellas artes.
0,922	si fuese un producto bonito, y de buena calidad, a un precio razonable, yo creo que se podria comprar porque es un recuerdo de la upv, la empresa donde trabajas.
0,926	el logotipo de la upv/ehu, siempre que se coloque de forma adecuada en un producto aporta al producto un elemento diferencial, que deberia siempre estar unido a calidad y diseno.
0,927	porque seria un buen recuerdo del paso por la universidad, ademas creo que los universitarios nos sentimos identificados con la upv y llevar algo con su logotipo seria un orgullo.

0,928	estaria interesada, siempre que este disenando con buen gusto, con buena calidad -muy importante- y sea discreto a la par que practico, por mi respuesta a la pregunta 12; mi grado de satisfaccion de pertenecer a esta organizacion (aunque todo es mejorable).
0,929	porque pienso que puede ser una manera de publicitar nuestra institucion. no obstante, creo que nuestra mejor carta de presentacion es realizar un trabajo excelente tanto en docencia como en investigacion. en la consecucion de este objetivo todos somos responsables, los profesionales docentes, el personal pas, etc. pero de una forma muy importante las instituciones pertinentes. no hay mas que fijarse en el estado del campus de leioa como para pensar que no es la mejor manera de dar a conocer nuestra universidad el estado en el que se encuentra, por ejemplo.
0,931	ya lo he hecho de otras universidades en las que he estado. a pesar de sus problemas no siento verguenza por pertenecer a la upv/ehu, mas bien orgullo.
0,931	porque, a pesar de los pesares, siento a la upv/ehu como algo propio. me ha dado formacion y trabajo. ademas, es una bonita manera de poseer un chillida.

Cluster 2/4	
Criterio	Respuesta
0,885	porque indicaria que estoy orgullosa de pertenecer a la upv/ehu
0,889	para regalar a las personas que estan vinculadas con el departamento sobre todo temporalmente, (alumnos de doctorado nacionales y extranjeros) profesores visitantes, investigadores. para el personal de la upv/ehu en diversas situaciones por ejemplo, jubilaciones, 25 anos de pertenencia a la upv/ehu, etc.
0,890	porque me gustaria poder disponer de productos con el logotipo de la upv/ehu -tal como ocurre con los de otras universidades que uno tiene ocacion de visitar- tanto para uso personal como para regalar a familiares, conocidos, amigos o colegas.
0,895	cuando visito otras instituciones, museos, etc., suelo comprar algun producto como recuerdo, con mayor razon compraria productos con el logotipo de la upv/ehu para uso personal y tambien para regalar a otras personas.

0,900	por que me siento orgullosa de pertenecer a la upv/ehu
0,901	lo aceptaria como regalo de la upv-ehu, pero nunca compraria un producto de este tipo para regalarselo a nadie de mi familia o amigos. lo veria bien como regalo institucional. por ejemplo un profesor visitante al que se le regala algo como recuerdo de su estancia en nuestra universidad.
0,901	por estar muy satisfecho de mi pertenencia a la upv/ehu.
0,901	cuando visito otras universidades siempre compro algun objeto con el logotipo como recuerdo, siendo donde trabajas, lo logico es que te guste llevar algo que te identifique con ella. en epocas de regalos, como navidad, estas deseando tener algo caracteristico para obsequiar a familiares y amigos.tambien cumplir con compromisos. cuando nos visitan de otras universidades quedariamos muy bien regalandoles algo que sirva para que recuerden nuestra universidad.
0,903	como he dicho, estoy orgulloso de pertenecer a la institucion, y por tanto demostrarlo con el logo. tambien creo que pueden ser un buen regalo-recuerdo, especialmente para colegas de otras universidades.
0,903	de hecho, estoy muy orgulloso de ser de la upv, y ya tengo algunos productos, que los luzco permanentemente: los relojes, el llavero, etc... creo que este tipo de productos son fundamentales como detalles para los profesores visitantes.
Cluster 3/4	
Criterio	Respuesta
0,770	no me gusta exteriorizar mis sentimientos con un logotipo
0,770	no me gustan los logotipos
0,770	porque el logotipo no me gusta
0,796	no me gusta llevar ningun tipo de marca, logotipo o escudo
0,803	no, no son el tipo de productos que compro porque si. no me gusta hacer publicidad y en caso de hacerlo seria por una causa social.
0,807	no, porque no me gusta nada llevar propaganda de ningun tipo
0,807	no me gustan
0,813	no me gustan los objetos que llevan un logotipo,ir haciendo propaganda
0,813	en general, no suelo comprar productos que lleven el nombre de la marca como motivo principal. otra cosa es que el dibujo o el diseno del producto me guste, e ineviatblmente lleve un pequeno logotipo que no pueda esconder.

0,818	no me gustan los objetos con logotipos.
Cluster 4/4	
Criterio	Respuesta
0,867	porque la universidad no es el puto toys?r us. yo, en mi cortedad, entiendo la universidad por un sitio donde se va a aprender cosas y hay gente que las ensena, no un contubernio de mercachifles. ademas yo creo que mis amigos podria llegar a la agresion fisica si aparezco un dia con un vaciabolsillos (¿¡!?) con un anagrama de la upv. por otra parte, y entrando en el terreno del marketing, del q soy un ignorante, si bien cuento con nociones instintivas a nivel de usuario, yo creo que esos productos son los clasicos productos que se regalan. me compro un bote de colacao y me regalan un panuelo, me compro un pack doble de nivea y me dan un reloj muy guapo de nivea. no voy a ir a comprar un reloj de la upv (cuya imagen de marca es, entiendanme, bastante ambigua como poco) por que parece de propaganda. y si, ya esto es que es de mofa, cojo y le regalo a mi chavala una taza de la upv me la parte en la cara, por manguan. espero que se hagan cargo.
0,871	porque no me parece que sea un servicio de la universidad. podiais gastar ese dinero que parece que vais a invertir en marketing en otro tipo de cosas mas urgentes como, por ejemplo, mejorar el servicio editorial, apoyar en serio la ensenanza en euskera promocionando una asignatura obligatoria de euskera durante los cuatro anos de carrera para todos... en algo que sirva para algo productivo. ¿quereis prestigio? pues ganaroslo sin marketing y chapuzas de ese estilo. perdon pero no me parece muy normal vender tacitas para que tus padres beban el cafe por la manana pensando en que parte de la matricula de sus hijos se utiliza para ese tipo de actividades tan dignas de una universidad.
0,881	porque no compraria nada o no pagaria dinero por hacer propaganda de ninguna entidad, en cambio si te lo regalan pues si te lo pondrias pero no se una taza si puedes comprar, pero un reloj de mano...no se, creo que debes de ser muy creyente de la upv.no se, creo que ya no se estila eso.

0,887	entre otras cosas porque creo que el dinero del que dispone la universidad deberia invertirse en otras cosas de mayor importancia. tampoco se si es asunto suyo o no pero como decia antes, ya que por una vez en toda mi vida en la universidad se me da la oportunidad de explayarme, voy a hacerlo a gusto. estudio una carrera, educacion social, que se financio con presupuesto cero. se cogio a profesores de aqui y de alla y se les dijo, tu tienes que impartir esta materia y tu tal otra. eso en lo que se refiere a las asignaturas troncales que marca el mec y que es obligatorio impartir porque son las que definen la carrera, porque en cuanto a las optativas el tema es otro; estudio educacion social y las asignaturas optativas que deberian servirme para definir mi futuro profesional, es decir, para especializarme, son tan interesantes para dicho futuro como lo son las siguientes: matematica recreativa; fisica y quimica de la vida cotidiana; la educacion en la historia; etc. la que mas relacion tiene con la educacion social es la ultima, como pueden apreciar, y esa relacion se encuentra en el nombre de la asignatura, que incluye la palabra educacion, pero no va mas alla. asi es que me plantean a mi que padezco la falta total de interes y de presupuesto hacia lo propio por parte de los organismos pertinentes que si compraria un vaciabolsillos (?!), que ni siquiera se lo que es, con el logotipo de la upv para financiar no se que causa y les digo que se metan ese vaciabolsillos por donde les quepa. no se de quien ha sido la brillante idea de promocionarse a traves de estos articulos, pero le auguro un negro futuro. personalmente no pienso comprar nada.
0,896	ya tengo 2 carpetas y algun boligrafo con el logotipo de la universidad que podre guardar de recuerdo si quiero.ademas, ¿esto que es?, ¿una universidad o una tienda?¿otra manera de querer sacarnos dinero?¿esto es lo mas importante y en lo que debe trabajar la universidad?¿que mas da estudiar en la upv o en otra universidad? ni que fueramos la mejor universidad del mundo y nos debamos sentir orgullosisimos de ser miembros de la upv y llevar cositas que lo demuestren

0,901	porque no me parece que sea atractivo para el trabajador comprar pproductos con el logotipo de tu empresa, porque la mayoria de la gente pensaria que te lo han regalado. encima de tener que aguantar que nos regalan las cosas, encima nos sale dinero. asi que con rotundidad no.
0,909	porque no me parece que sean interesantes
0,910	porque no me interesa tener una taza de la universidad, no me parece importante, prefiero tener calefaccion en clase, la verdad. entiendo que sea una forma de promocion, pero es que no me parece muy importante.
0,916	porque no me parece atractivo tener ningun tipo de producto con el logo de la upv y si lo regalas igual la gente piensa que te lo han regalado a ti en la universidad y como no te gusta lo regalas. no se para poder vender tendria que tener algun elemento atractivo algo que no te ofrezca un producto similar (precio, donativo,calidad...algo) pero planteado asi no me parece atractivo.
0,919	en general, estoy bastante orgullosa de mi universidad. sin embargo, si bien es cierto que no tendria ningun problema en llevar productos de la upv, preferiria gastar mi dinero en regalar una camiseta del m.i.t. probablemente porque el m.i.t. produce numerosas noticias cientificas, es muy prestigiosa en muchos campos de la investigacion. obviamente, no es una universidad publica, y por tanto, maneja mucho mas dinero en investigacion, lo que a su vez le permite dar muchas mas noticias, lo que a su vez le da mucha mas publicidad. asi que, deduzco que es una cuestion de publicidad.

Tabla B.28: Respuestas modales para individuos respondiendo en Castellano. Criterio de selección χ^2.

Respuestas modales a partir del criterio χ^2: Euskera

Cluster 1/4	
Criterio	Respuesta
0,824	publizitatea egiteko modu bat delako. jendeak ikusten zaitu eta artikulu hori zergatik daramazun galdetzen dizu, horrek unibertsitateari buruz eta bere onurari buruz hitzegiteko aukera paregabea emanez. lehen azaldu bezala, unibertsitatean ikasi ez duen gendeak ez dauka ezta ideiarik bez unibertsitatea zer den, nola lan egiten duen edo zer irakasten duen ere (eta ez naiz matematiketaz bakarrik ari). ez dakite mundu berri bati ateak irekitzen dituela, atzerrian ikasteko aukera, jende eta lagun mordoa ezagutzeko aukera, munduan zeure kabuz mugitzeko eta murgitzeko aukera eta askatasuna, eta noski, lan egiteko nahiko diren ezagutza oinarrizkoak ere ematen dituela.
0,870	oparitzeko batez ere. logoa era 'diskretoan' jarriz gero, niretzat ere erosiko nituzke. produktua ona bada nahiago dut ehu-ri erostea. oparitzean, ehu ezagunagoa egin daiteke.
0,907	azken finean ehu beste edozein instituzio modukoa da, eta halan nola athletik edo bizkaialde fundazioko logotipoa dauketen gauzak erosten dtudan, ehukoak ere erosteko prest nago, politak edo erabilgarriak diren bitarten.
0,912	unibertsitateari laguntzeko eta niretzako ere produktu horiek erostea interesgarria izan daitekelako
0,915	bai-ez beste aukerarik ez bada, bai aukeratu behar, noiz edo noiz erosiko nukeelakoan. hala ere, ez dut uste garrantzizko gauza denik logotipodun produktuen kontua. ez nau horrek zirraratzen. hau abiapuntu hartuta erantzungo ditut behekoak. oparia egiteko bada, tira!, baina momentuko aukera izanda.
0,919	erosiko nuke, baino ziur aski unibertsitate munduarekin erlazionatuta dagoen norbaiti oparitzeko; beste unibertsitate baterko irakasleren bati... edo horrelakoak. egia da, produktua bereziki interesgarria izango balitz... niretzako edo etxekoentzako erosiko nuke. oparitzeko... ez dakit, badirudi, logotipoa edukita oparitu egin dizula unibertsitateak... nahiz eta horrela ez izan. niri oraindik ez didate ezer oparitu logotipoduna.

0,920	irudi polita iruditzen zait eta ispiritu korporatiboa bultzatu beharra daukagu. dena den, bertako langile bezela, erosi baino, insittuzioak oparitua gehiago nahi nuke.
0,920	ez unibertsitateko logotipoa daramalako, baina kalitatea ona bada eta prezioa ez bada oso altua zergatik ez erosi? hori bai, logotipoa ez da izan behar oso andia, bestela propagandako produktua izango litzatekelako, baina adibidez kamixeta bat logotipo txiki batekin bularrean eta atzean edozein marrazki eta listo.
0,921	logotipoa polita izateaz gain, ezaguna delako. eta zergatik ez, beste marka batzuko produktuak erosten baditugu ere.
0,921	bertako irakaslea naiz eta eguneroko disfuntzioak, beharrak eta mixeriak ere ezagutzen ditut. hala ere, pentsatzen dut oso zerbitzu ona eskaintzen duela kontuan hartuta eusko jaurlaritzak ematen dion baliabide eskasia, eta baita azken bolada honetan unibertsitate-kontseilariarengandik jasan behar izan dugun desprestigio mediatikoa ere.

Cluster 2/4

Criterio	Respuesta
0,829	ni oso arro nago ehu-n ikasten ari naizelako,izan ere betiko nire nahia bertan ikastea baitzen.behin ehu-n ikasten egonik,oso arro egongo nitzeke holako produktua nere eskutan edukitzea eta jendeak ikustea ni ehu-ko ikasle bat,eta beraz bere parte bat naizela.
0,876	ez daukadalako inongo arazorik e.h.uko edozein produktu erosteko, beti ere, arrazoizko prezio eta balioa aterako badiot. izan ere, oso orgullosa nago e.h.u.-n ikasten dudalako.
0,876	zerbitzu publiko duin eta, neurri batean, harro egoteko modukoa eskaintzen du ehuk. ez nuke ikusiko marka komertzial bat bezala.
0,877	egia esan, ehuren ikurra edo logotipoa oso polita da, baita lema ere. bata zein bestea sustatzen eta gehiago gizarteratzen saiatu beharko ginateke. ildo horretan, ez genuke ahaztu behar ehu, izenez bederen, 'euskal herriko' unibertsitatea dela, baina euskal herritartasun horri ez diola behar adina (eta merezi adina) arreta eskaintzen.

0,885	nik harro nago ehu-en ikasteaz, nik ikasketa publikoetan sinisten dut. unibertsitate publikoa ona bada, sistema ona dagoela pentsatzen dut. arlo publikoa ona bada, bizimodu, giro eta harreman onak egongo dira. arlo pribatua zuk ordaintzen duzu beraz zu zeuk eskatzen duzu ona izatea eta berez saiatzen da hala izaten. bere hobe beharrez, aurrera segitu nahi baldin badu.
0,892	unibertsitate honen partaide izatea, ikastegi batean ikastea baino haratago egon behar lukeelako.
0,897	pare bat arrazoi eman ditzaket. alde batetik, atzerriko unibertsitateetara joaten garenean, gustoko izaten dugu hango zerbait ekartzea, eta ondorioz, hemen ere berdina egin dezakegu. bestalde, oparitxoren bat egin nahi denean, ideia aproposa iruditzen zait unibertsitateko logotipoa daukan zerbait oparitzea.
0,905	beste edozein modukoa ikusten dodalako eta euskal herriko unibertsitatea gurea dogulako.
0,909	ez daukadalako inolako arazorik propaganda daramaten erabilgarriak, jantzi edo bitxiak eramateko. baldin eta horrek upv/ehu gehiago ezagutarazten lagunduko badu. beti nago horren alde, laguntza eskaintzeko: egunkaria, ikastolak eta abarren alde, eta beste hainbaten alde agertzen naizen moduan, upv/ehuren alde baita. beraz ez dut salbuespena egiten upv/ehukin.
0,917	azkenean hemengo produktu bat erosiko nuke eta horrekin unibertxitateari lagunduz euskal gizarte osoari laguntzeko ahalegin txiki bat izango litzateke.

Cluster 3/4	
Criterio	Respuesta
0,862	sekula ez nuelako unibertsitate honen propagandarik eginen honela! zergatik duzue horrenbesterainoko ardura logotiko madarikatuarekin? kalitatezko hezkuntza lortzea da propaganda egiteko modua, eta ez soilik logotipo ponpoxo bat eginez. nik uste dut unibertsitate honek logotipoa eta kontsumismoa ez diren beste hainbat premia zein lehentasun dituela. ez dut egoki ikusten ehu-ko produktuetan dirua inbertitzea.
0,873	ez zaidalako inoren ez ezeren propagandarik egitea. ez nioke begiratu logoa duen edo ez.

0,874	logotipoa leku askotan ikusten delako (liburuetan, orrietan, karpetak, agendak, etab...) eta ez nuke upv/ehuko logotipoa duen beste zerbait erosiko. logoa ere zaharkitua dagoela iruditzen zait baina ikusi bazain laster, upv/ehurekin identifikatzen da.
0,887	logotipoa gustoko dudan arren, produktuen diseinuak hobetu beharko lirateke,gaur egun gazteok erabiltzen dugun estiloan erreparatuz.produktua saldu ahal izateko, ez dut uste beharrezkoa denik logotipoa horren handia izaterik. gainera, ehuak salgai dituen produktuak ezagutzera emango nituzke.
0,890	nire ustez harro eraman ahal izateko ehuko edozein produktu unibertsitate honek gauza askotan hobetu beharko luke. gainera ez naiz oso propaganda zalea.
0,893	ez naiz oro har logotipo zale, ez ehuri dagokionez, ez bestela.
0,894	ez nagoelako batere pozik unibertsitatea dena eta egiten duenarekin eta naiz identifikatuta ikusten. beraz, zer egingo nuke nik logotipoa daukan zerbaitekin?
0,894	ez dudalako inolako interesik ez produktu horietan ez upv/ehuko logotipoa daraman ezertan
0,897	ez zaizkidalako gustatzen logotipoak dituzten gauzak.
0,899	propaganda egitea guztokoa ez dudalako

Cluster 4/4	
Criterio	Respuesta
0,797	logotipoa daramaten produktuak gustokoak ez ditudalako
0,806	ez zaizkit logotipoak dauzkaten produktuak gustatzen.
0,818	printzipioz ez ditudalako produktuak ikusi, beraz ez dakit gustokoak izango ditudan edo ez, bestalde gustokoak izango banitu agian zer edo zer erosiko nuke, baina ez dut uste, ez zait propaganda eramatea gustatzen.
0,840	ez da upv/ehukoa delako baizik eta ez zaizkidalako logotipoak eramaten dituzten produktuak gustatzen ez niretzako ez inorri oparitzeko.
0,842	nire kasuan, ez ditudalako interesgarritzat jotzen halako tresnak. ezta ehuren logotipoa barne ere. gainera ez dut uste ehuren logotipoak ematen dienik balio handiagorik. azkenik ez dut uste honelako ekintzak bultzatzen duenik ehuri buruzko ezagupenik.

0,861	logotipoak ez zaizkit gustatzen
0,872	ez dut uste hori bidea denik, eta, nahiz eta ez esan inolaz ere ez nukeela erosiko, ez dut uste berehala erosiko nukeenik.
0,877	horrelako gauzak erostea ez dut gustoko
0,880	ez zaizkidalako horrelako produktuak gustatzen
0,900	horrelako logotipoak soinean eramatea ez zait asko gustatzen. ez da arropa marka bat bezala.

Tabla B.29: Respuestas modales para individuos respondiendo en Euskera. Criterio de selección χ^2.

Bibliografía

Abascal, E., Díaz de Rada, V., García Lautre, I. & Landaluce, I. (2008), Factorial analysis of three-way categorical tables relating to different populations: A compared analysis of one survey carried out by two different ways, *in* '7th International Conference on Social Science Methodology (RC33 Logic and Methodology in Sociology)', International Sociological Association, ISA, Naples, Italy.

Abascal, E., Fernández, K., Landaluce, M. & Modroño, J. (2001), 'Diferentes aplicaciones de las técnicas factoriales de análisis de tablas múltiples en las investigaciones mediante encuestas', *Metodología de Encuestas* **3**(2), 251–279.

Abascal, E., García Lautre, I. & Landaluce, I. (2004), 'Análisis de la evolución a través de encuestas. trayectoria electoral de las comunidades autónomas españolas en el período 1977-2004.', *Metodología de Encuestas* **6**(2), 147–162.

Abascal, E., García Lautre, I. & Landaluce, I. (2006), Multiple factor analysis of mixed tables of metric and categorical data, *in* M. Greenacre & J. Blasius,

eds, 'Multiple correspondence analysis and related methods', Chapman-Hall, Boca-Raton, FL, pp. 351–367.

Abdessemed, L. & Escofier, B. (1996), 'Analyse factorielle multiple de tableaux de fréquences: comparaison avec l'analyse canonique des correspondances', *JSSP* **137**(2), 3–17.

Amato, S., Esposito Vinzi, V. & Tenenhaus, M. (2005), A global goodness-of-fit test index for pls structural equation modelling, Technical report, HEC School of Management, France.

Amemiya, T. (1981), 'Qualitative response models: A survey', *Journal of Economic Literature* **19**(4), 481–536.

Amemiya, T. (1985), *Advanced Econometrics*, Blackwell, Oxford.

Benzécri, J. (1983), 'Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence.', *Les Cahiers de l'Analyse des Données* **8**(3), 351–358.

Benzécri, J. P. (1973), *L'analyse des données (tome 1 et 2)*, Dunod, Paris.

Benzécri, J. P. (1977), 'Analyse discriminante et analyse factorielle', *Les Cahiers de l'Analyse des Données* **4**, 369–406.

Benzécri, J. P. (1979), 'Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire', *Cahiers de l'Analyse des Données* **4**, 377–378.

Benzécri, J. P. (1982), *Histoire et préhistoire de l'analyse des données*, Dunod.

Berkson, J. (1944), 'Application of the logistic function to bio-assay', *Journal of the American Statistical Association* **39**, 357–365.

Berndt, E., Hall, B., Hall, R. & Hausman, J. (1974), 'Estimation and inference in nonlinear structural models', *Annals of Economic and Social Measurement* **3**(4), 653–665.

Burt, C. (1950), 'The factorial analysis of qualitativa data', *British Journal of Statist. psychol.* **3**(3), 166–185.

Bárcena, M. J. (2001), Técnicas multivariantes para el enlace de encuestas, PhD thesis, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Bilbao.

Bécue-Bertaut, M. & Pagès, J. (2001), 'Analyse simultanée de questions ouvertes et de questions fermées. méthodologie, exemple.', *Journal de la Société Française de Statistique* **42**(4), 91–104.

Bécue-Bertaut, M. & Pagès, J. (2004), 'A principal axes method for comparing contingency tables: MFACT', *Computational Statistics & Data Analysis* **45**(3), 481–503.

Bécue-Bertaut, M. & Pagès, J. (2008), 'Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data', *Computational Statistics & Data Analysis* **52**, 3255 – 3268.

Carrol, J. (1968), Generalization of canonical correlation to three or more sets of variable, *in* 'Proceedings of the American Psychological Association', pp. 227–228.

Cazes, P. & Moreau, J. (1991), Analysis of a contingency table in which the rows and the columns have a graph structure, *in* E. Diday & Y. Leche-

vallier, eds, 'Symbolic-Numeric Data Analysis and Learning', Nova Science Publishers, New York, pp. 271–280.

Cazes, P. & Moreau, J. (2000), Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique, *in* J. Moreau, P. A. Doudin & P. Cazes, eds, 'L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données', Springer, Berlin-Heidelberg, pp. 87–103.

Chateau, F. & Lebart, L. (1996), Assessing sample variability in the visualization techniques related to principal components analysis: bootstrap and alternative simulation methods, *in* A. Prats, ed., 'XII Symposium on Computational Statistics COMPSTAT '96', Physica-Verlag, Heidelberg, pp. 205–210.

Chin, W. W. (1998), The partial least squares approach to structural equation modelling, *in* G. A. Marcoulides, ed., 'Modern Methods for Business Research', Lawrence Erlbaum Associates, London, pp. 295–336.

Chin, W. W. (2001), *PLS-Graph user's guide*, C. T. Bauer College of Business, University of Houston, USA.

Daudin, J., Duby, C. & Trecourt, P. (1988), 'Stability of principal component analysis studied by the bootstrap method', *Statistics* **19**(2), 241–258.

Dhrymes, P. (1984), Limited dependent variables, *in* Z. Griliches & M. Intriligator, eds, 'Handbook of Econometrics', Vol. 2, North Holland, Amsterdam.

Efron, B. (1979), 'Bootstraps methods: another look at the jacknife', *Ann. Statist.* **7**, 1–26.

Efron, B. & Tibshirani, R. (1993), *An introduction to the bootstrap*, number 57 *in* 'Monographs on Statistics and applied Probability', Chapman & Hall, New York.

Escofier, B. (1965), l'Analyse des correspondances, PhD thesis, Faculté des Sciences de Rennes.

Escofier, B. (1984), 'Analyse factorielle en référence à un modèle: application à l'analyse d'un tableau d'échanges', *Revue de statistique apliquée* **32**(4), 25–36.

Escofier, B. & Drouet, D. (1983), 'Analyse des différences entre plusieurs tableaux de fréquence', *Les Cahiers de l'Analyse des Données* **8**(4), 491–499.

Escofier, B. & Pagés, J. (1986), 'Le traitement des variables qualitatives et tableaux mixtes par analyse factorielle multiple', *Data Analysis and Informatics* **IV**(2), 179–191.

Escofier, B. & Pagés, J. (1992), *Análisis Factoriales Simples y Múltiples*, Servicio Editorial de la Universidad del País Vasco, Bilbao.

Escofier, B. & Pagés, J. (1994), 'Multiple factor analysis (afmult package)', *Computational Statistics and Data Analysis* **18**, 121–140.

Escofier, B. & Pagés, J. (1998), *Analyses factorielles simples et multiples: objectifs, méthodes et interpretation*, 3 edn, Dunod, Paris.

Escoufier, Y. (1973), 'Le traitement des variables vectorielles', *Biometrics* **29**, 751–760.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From Data Mining to Knowledge Discovery in Databases', *Artificial Intelligence Magazine* .
URL: *http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf*

Fernández-Aguirre, K. & Modroño, J. I. (2008), Teoría de la asignatura análisis de datos: Métodos exploratorios multivariantes. Apuntes de clase.

Fernández, K., Martín, A. M., Modroño, J. I. & Zorrilla, P. (2005), Análisis de la viabilidad de la tienda universitaria de la UPV/EHU "EHUdenda" de acuerdo con la imagen percibida de la institución, Technical report, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Bilbao.

Fornell, C. (1992), 'A national customer satisfaction barometer: the swedish experience', *Journal of Marketing* **56**, 6–21.

Fornell, C. & Bookstein, F. (1982), 'Two structural equation models: LISREL and PLS applied to consumer exit-voice theory', *Journal of Marketing research* **19**, 440–452.

García Lautre, I. (2001), Medición y análisis de las infraestructuras. Una nueva metodología basada en el análisis factorial múltiple, PhD thesis, Universidad Pública de Navarra.

García Lautre, I. & Abascal, E. (2003), 'Una metodología para el estudio de la evolución de variables latentes. análisis de las infraestructuras de ca-

rreteras de las comunidades autónomas (1975-2000)', *Estadística Española* **45**(153), 193–210.

Gifi, A. (1990), *Non Linear Multivariate Analysis*, Wiley, Chichester.

Goitisolo, B. (2002), El análisis simultáneo. Propuesta y aplicación de un nuevo método de análisis factorial de tablas de contingencia, PhD thesis, Universidad del País Vasco (UPV/EHU).

Goitisolo, B. & Zárraga, A. (2008), *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, chapter Factorial Analysis of a Set of Contingency Tables, pp. 219–226.

Gower, J. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* **27**(4), 857–871.

Grabmeier, J. & Rudolph, A. (2002), 'Techniques of cluster algorithm in data mining', *Data Mining and Knowledge Discovery* **6**, 303–360.

Greenacre, M. J. (1987), 'Graphical analysis of readership data', *The American Statistical Association* . Section on Statistical Graphics.

Greenacre, M. J. (1993), *Correspondence Analysis in Practice*, Academic Press, London.

Guiraud, P. (1954), *Les caractères statistiques du vocabulaire*, P.U.F., Paris.

Guiraud, P. (1960), *Problèmes et méthodes de la statistique linguistique*, P.U.F., Paris.

Guttman, L. (1941), The quantification of a class of attributes: a thery and method of a scale construction, *in* P. Horst, ed., 'The prediction of personal adjustment', SSCR, New York, pp. 251–264.

Hand, D., Mannila, H. & Smyth, P. (2001), *Principles of Data Mining*, Adaptative Computation and Machine Learning, MIT press, Cambridge, Mass.

Hayashi, C. (1956), 'Theory and examples of quantification (II)', *Proc. of the Institute of Stat. Math.* **4**(2), 19–30.

Husson, F. & Pagès, J. (2006*a*), 'Aspects méthodologiques du modèle indscal', *rsa* **54**(2), 83–100.

Husson, F. & Pagès, J. (2006*b*), 'Indscal model: geometrical interpretation and methodology', *Computational Statistics and Data Analysis* **50**(2), 358–378.

Hwang, H., Dillon, W. R. & Takane, Y. (2006), 'An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents', *Psychometrika* **71**(1), 161 – 171.

Jacobowicz, E. & Derquenne, C. (2007), 'A modified pls path modeling algorithm handling reflective categorical variables and a new model building strategy', *Computational Statistics and Data Analysis* **51**(8), 3666–3678.

Jöreskog, K. (1973), A general method for estimating a linear structural equation system, *in* A. Goldberger & O. Duncan, eds, 'Structural Equation Models in the social sciences', Academic Press, New York, pp. 85–112.

Landaluce, M., Fernández, K. & Modroño, J. (1999), 'Reflexiones sobre el uso

comparativo del análisis factorial múltiple y de la metodología statis para el análisis factorial múltiple', *Methodologica* **7**, 37–65.

Landaluce, M. I. (1995), Estudio de la estructura de gasto medio de las Comunidades Autónomas españolas. Una aplicación del Análisis Factorial Múltiple, PhD thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU).

Lavit, C. (1988), *Analyse Conjointe de Tableaux Quantitatifs*, Masson, Paris.

Le Dien, S. & Pagés, J. (2003), 'Analyse factorielle multiple hierarchique', *Revue de Statistique Appliquée* **51**(4), 83–93.

Le Dien, S. & Pagés, J. (2010), 'Dmfa: Dual multiple factor analysis', *Communications in Statistics-Theory and Methods* **39**(3), 483–492.

Lebart, L. (1975), 'L'orientation du dépouilement de certaines enquêtes par l'analyse des correspondences multiples', *Consommation* **2**, 73–96.

Lebart, L. (1982), Exploratory analysis of large sparse matrices, with application to textual data, *in* 'COMPSTAT', Physica Verlag.

Lebart, L. (1994), Complementary use of correspondence analysis and cluster analysis, *in* M. J. Greenacre & J. Blasius, eds, 'Correspondence Analysis in the Social Sciences'.

Lebart, L. (2006), Validation techniques in multiple correspondence analysis, *in* M. Greenacre & J. Blasius, eds, 'Multiple correspondence analysis and related methods', Chapman-Hall, Boca-Raton, FL.

Lebart, L., Morineau, A. & Piron, M. (2000), *Statistique Exploratoire Multidimensionelle*, 3 edn, Dunod, Paris.

Lebart, L., Morineau, A. & Piron, M. (2006), *Statistique Exploratoire Multidimensionelle*, 4 edn, Dunod, Paris.

Lebart, L., Piron, M. & Steiner, J. (2003), *La Sémiométrie*, Dunod, Paris.

Lebart, L., Salem, A. & Berry, L. (1998), *Exploring Textual Data*, Kluwer Academic Publishers, New York.

Lebart, L., Salem, A. & Bécue-Bertaut, M. (2000), *Análisis estadístico de textos*, Milenio, Lleida.

L'Hermier des Plantes, H. (1976), STATIS, Structuration des tableaux à trois indices de la statistique, PhD thesis, Université de Montpellier.

Lohmöller, J.-B. (1989), *Latent Variable Path Modeling with Partial Least Squares*, Physica-Verlag, Heidelberg.

Maddala, G. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York.

Markus, M. T. (1994), *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*, DWSO Press, Leiden University.

Masson, M. (1974), 'Analyse non linéaire de données', *C. R. Acad. Sc.* **278**.

McFadden, D. (1984), Econometric analysis of qualitative response models, *in* Z. Griliches & M. Intriligator, eds, 'Handbook of Ecnonometrics', Vol. 2, North Holland, Amsterdam.

Michailidis, G. & de Leeuw, J. (1998), 'The gifi system of descriptive multivariate analysis', *Statistical Science* **3**, 307–336.

Morand, E. & Pagès, J. (2007), 'L'analyse factorielle multiple procrustéenne', *Journal de la Société Française de Statistique* **148**(2), 65–97.

Morineau, A. (1984), 'Note sur la caractérisation statistique d'une classe et les valeurs-tests', *Bulletin Technique du Centre de Statistique et d'Informatique Appliquées* **2**(1-2), 20–27.
URL: *http://www.deenov.com/analyse-de-donnees/documents/article-valeur-test.aspx*

Nakache, J.-P. & Confais, J. (2005), *Approche pragmatique de la classification*, Technip, Paris.

Nishisato, S. (1980), *Analysis of Categorical Data*, Univ. of Toronto Press, Toronto.

Pagés, J. (1996), 'Eléments de comparaison entre l'analyse factorielle multiple et la méthode statis', *Revue de Statistique Appliquée* **XLIV**(4), 81–95.

Pagès, J. (2002), 'Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes', *Revue de Statistique Appliquée* **50**(4), 5–37.

Pagès, J. (2004), 'Analyse factorielle de données mixtes', *Revue de Statistique Appliquée* **52**(4), 93–111.

Pagés, J. (2005), 'Analyse factorielle multiple et analyse procustéene', *Revue de Statistique Appliquée* **LIII**(4), 61–86.

Pagés, J. & Camiz, S. (2008), 'Analyse factorielle multiple de données mixtes: application à la comparaison de deux codages', *Revue MODULAD* **38**, 178–183.

Pagés, J. & Tenenhaus, M. (2001), 'Multiple factor analysis combined with PLS path modelling. application to the analysis of relationships between physico-chemical variables, sensory profiles and hedonic judgements', *Chemometrics and Intelligent Laboratory Systems* **58**, 261–273.

Pagés, J. & Tenenhaus, M. (2002), 'Analyse factorielle multiple et approche PLS', *Revue de Statistique Appliquée* **L**(1), 5–33.

Podani, J. (1999), 'Extending Gower's general coefficient of similarity to ordinal characters', *Taxon* **48**(2), 331–340.

Robert, P. & Escoufier, Y. (1976), 'A unifying tool for linear multivariate methods: the RV-coefficient', *Journal of Applied Statistics* **25**(3), 257–265.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M. & Lauro, C. (2005), 'PLS path modeling', *Computational Statistics & Data Analysis* **48**, 159–205.

Tenenhaus, M. & Young, F. W. (1985), 'An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data', *Psychometrika* **50**(1), 91–119.

Ter Braak, C. J. F. (1988), Partial canonical correspondence analysis, *in* H. H. Bock, ed., 'Classification and Related Methods of Data Analysis', Elsevier Science Publishers, pp. 551–558.

Trinchera, L. & Russolillo, G. (2010), 'On the use of Structural Equation Models and PLS Path Modeling to build composite indicators', Working paper n. 30, Dipartimento di Studi sullo Sviluppo Economico, Università degli Studi di Macerata.

Wold, H. (1979), 'Model construction and evaluation when theoretical knowledge is scarce: an example of the use of partial least squares', Cahier 79.06 de Département d'econométrie, Faculté des Sciences Économiques et Sociales, Université de Genève, Genéve.

Wold, H. (1982), Soft modeling: the basic design and some extensions, *in* K. G. Jöreskog & H. Wold, eds, 'Systems Under Indirect Observation', Part II, North Holland, Amsterdam, pp. 1–54.

Wold, H. (1985*a*), Partial least squares, *in* S. Kotz & N. L. Johnson, eds, 'Encyclopedia of Statistical Sciences', Vol. 6, Wiley, N.Y., pp. 581–591.

Wold, H. (1985*b*), Partial least squares, *in* S. Kotz & N. Johnson, eds, 'Encyclopedia of Statistical Sciences', Vol. 6, Wiley, New York, pp. 581–591.

Yule, G. U. (1944), *The Statistical Study of Literacy Vocabulary*, Cambridge University Press.

Zárraga, A. & Goitisolo, B. (2002), 'Méthode factorielle pour l'analyse simultanée de tableaux de contingence', *Revue de Statistique Appliquée* **L**(2), 47–70.

Zárraga, A. & Goitisolo, B. (2006), Simultaneous analysis: A joint study of several contingency tables with different margins, *in* M. Greenacre & J. Bla-

sius, eds, 'Multiple Correspondence Analysis and Related Methods', Chapman & Hall/CRC, Boca Raton, FL, pp. 327–350.

Zárraga, A. & Goitisolo, B. (2009), 'Simultaneous analysis and multiple factor analysis for contingency tables: Two methods for the joint study of contingency tables', *Computational Statistics & Data Analysis* **53**(8), 3171–3182. **URL:** *http://ideas.repec.org/a/eee/csdana/v53y2009i8p3171-3182.html*