ISCTE IUL
Instituto Universitário de Lisboa

Escola de Gestão
Departamento de Métodos Quantitativos

Extensões via *splines* da Análise em Componentes Principais

Nuno Filipe Jorge Lavado

Tese especialmente elaborada para obtenção do grau de

Doutor em Métodos Quantitativos

Especialidade em Estatística e Análise de Dados

Orientadora:
Doutora Teresa Calapez, Professora Auxiliar,
ISCTE-IUL

Abril, 2012

Composição do júri

Doutor Rui Manuel Campilho Pereira de Menezes
Doutor João António Branco
Doutor Jorge Filipe Campinos Landerset Cadima
Doutora Ana Maria Nobre Vilhena Nunes Pires de Melo Parente
Doutora Helena Maria Barroso Carvalho
Doutora Maria Teresa Delgado Calapez

Resumo

Uma nova abordagem para generalizar a Análise em Componentes Principais (ACP) para estruturas não-lineares é proposta nesta tese: *quasi-linear PCA* (qlPCA). Esta inclui transformações *spline* das variáveis originais optimizadas através de um processo de *Mínimos Quadrados Alternados* sobre uma determinada *função perda*. As transformações óptimas são explicitamente conhecidas após a convergência, sendo o sumário do modelo semelhante ao da ACP. Apesar do algoritmo proposto ser dedicado a ambientes de variáveis contínuas com eventual presença de relações não-lineares, a sua inspiração foram os algoritmos que emergiram do *sistema Gifi*, tendo estes sido especialmente concebidos para variáveis categoriais. Deste ponto de vista, pode afirmar-se que esta tese propõe uma solução para o seguinte problema:

> Tendo as variantes da ACP associadas ao *sistema Gifi* sido desenvolvidas para variáveis categoriais, como adaptá-las de modo a serem consideradas como uma abordagem não-linear em contexto de variáveis contínuas?

As variantes associadas ao *sistema Gifi* não são usualmente abordagens a considerar pelos investigadores de áreas do conhecimento que lidem com variáveis contínuas. Nesse sentido, considera-se que a qlPCA representa um contributo relevante, alargando o leque de aplicações do referencial teórico desenvolvido por Gifi.

Palavras-chave: Análise em Componentes Principais, linear, não-linear, *splines*, qlPCA, CATPCA.

Abstract

A new approach to generalize Principal Components Analysis (PCA) in order to handle nonlinear structures is proposed in this thesis: *quasi-linear PCA* (qlPCA). It includes spline transformations of the original variables, using *Alternating Least Squares* fitting of a suitable objective loss function to achieve optimal transformations. Optimal transformations are explicitly known after convergence and qlPCA reports model summary in a linear PCA fashion. Even though the proposed algorithm is designed for continuous variables eventually with nonlinear relationships, it was inspired by algorithms that emerged from the *Gifi system*, whose focus were categorical variables. Thus, this thesis proposes a solution for the following problem:

> Having Gifi's related approaches been developed for categorical variables, how to adapt them in order to be considered a nonlinear option also in the context of continuous variables?

Gifi's related approaches are not usually a valid option for researchers dealing with continuous variables. The proposed approach, qlPCA, can enlarge the scope of applications of Gifi's theoretical framework, being therefore a relevant contribution.

Keywords: Principal Components Analysis, linear, nonlinear, *splines*, qlPCA, CATPCA.

Agradecimentos

A realização duma tese de doutoramento envolve sempre pessoas e instituições com as quais, por diversas razões, contraímos uma dívida de gratidão indelével. As próximas linhas são-lhes dedicadas.

As minhas primeiras palavras são para a Professora Doutora Teresa Calapez, minha orientadora científica. Ao longo dos últimos anos, em todos os pequenos e grandes momentos da minha carreira científica, esteve sempre presente a meu lado, orientando-me cientificamente e motivando-me sempre para ir um pouco mais longe. O seu empenho, compreensão e amizade proporcionaram-me um ambiente onde o diálogo foi sempre uma constante e onde pude sem reservas desenvolver as minhas aptidões. Com a Professora Teresa Calapez aprendi muito e continuarei a aprender com a certeza de que nunca poderei retribuir. O meu muito obrigado.

Agradeço ao Instituto Superior de Engenharia de Coimbra e, em particular, ao Departamento de Física e Matemática o empenho demonstrado no apoio à realização destes trabalhos. Este permitiu que pudesse beneficiar de uma redução de serviço docente, sem a qual não teria conseguido fazer igual. Agradeço ainda todo o apoio financeiro que me tem concedido em diversas ocasiões, para apresentação dos avanços alcançados em congressos nacionais e internacionais.

À minha unidade de investigação, a Unidade de Investigação em Desenvolvimento Empresarial, UNIDE, ISCTE-IUL, agradeço também todo o apoio financeiro que me tem concedido em diversas ocasiões, para apresentação dos avanços alcançados em congressos nacionais e internacionais.

Ao Professor Doutor Jan de Leeuw gostaria de agradecer pela partilha dos resultados da sua investigação e pelo entusiasmo que transmite nas suas publicações, sem dúvida grande fonte de inspiração para os meus trabalhos.

À Professora Doutora Anita van der Kooij gostaria de agradecer pelas conversas e sugestões de que a minha investigação beneficiou e também por me ter recebido em Leiden.

À Professora Doutora Helena Carvalho agradeço a disponibilização da base de dados que usei no capítulo das aplicações.

Gostaria também de agradecer aos colegas do (meu) Departamento de Física e Matemática que sempre aceitaram, compreenderam e acompanharam este meu projecto.

Não me esqueço também dos meus pais e irmão que são sempre implicitamente parte de todos os meus trajectos.

E especialmente à minha esposa, Catarina Almeida, e aos meus dois filhos André e Beatriz que sempre a meu lado, orgulhosos e sorridentes, contribuíram em todo este percurso para colorir muitos momentos desta tese. Sem o seu apoio e paciência tudo teria sido bem mais difícil.

Este momento significa o culminar da minha investigação para doutoramento. Indica também que a persistência e os anos de trabalho deram fruto e que o desânimo de muitos momentos não foi em vão.

Índice

Lista de Figuras

Lista de Tabelas

Introdução

A Análise em Componentes Principais (ACP) tradicional apresenta duas limitações, a saber: foi concebida para analisar variáveis quantitativas e é pouco eficiente na presença de relações não-lineares entre variáveis.

Uma nova abordagem para generalizar a ACP para estruturas não-lineares é proposta nesta tese: *quasi-linear PCA* (qlPCA). Esta inclui transformações *spline* das variáveis originais, tendo o adjectivo *quasi* sido escolhido para sublinhar as vantagens que advêm do uso de *splines* lineares (transformações seccionalmente lineares). Apesar do algoritmo proposto ser dedicado a ambientes de variáveis contínuas com eventual presença de relações não-lineares, a sua génese foram os algoritmos para lidar com variáveis qualitativas.

Quando se codifica uma variável qualitativa, usando uma qualquer codificação numérica e se efectua a ACP tradicional, está-se a admitir que a análise das variáveis nominais e ordinais seja feita como se estas fossem numéricas, usando para o efeito a dita codificação. No entanto, diferentes opções na codificação produzem resultados diferentes e não é claro como estes devem ser interpretados, nem qual o seu significado.

A ideia fundamental das variantes da ACP concebidas para variáveis categoriais é partir dessa codificação arbitrária e, optimizando um determinado critério definido por uma *função perda*, determinar uma *quantificação óptima* para as categorias das variáveis. O critério a optimizar visa maximizar a homogeneidade entre as variáveis transformadas, ou de forma equivalente, minimizar a perda inerente à redução da dimensão. Sublinhe-se que não se trata de atribuir um significado numérico a variáveis por natureza sem esse significado, mas antes tornar a análise independente da escolha inicial da

codificação.

Na base desta tese está o conjunto de abordagens que tiveram a sua génese no referencial teórico desenvolvido por Gifi, vulgarmente designado de *sistema Gifi* (de Leeuw [7]). Na verdade, Gifi foi o heterónimo dum grupo de investigadores de Leiden, desde 1968 até 1991, para publicações conjuntas. Na sua última publicação, Gifi [12] clarifica que os desenvolvimentos originais se devem aos contributos de Bert Bettonvil, Eeke van der Burg, John van de Geer, Willem Heiser, Jan de Leeuw, Jacqueline Meulman, Jan van Rijckevorsel e Ineke Stoop.

Para além dos desenvolvimentos teóricos foi particularmente importante para o algoritmo proposto uma das soluções desenvolvidas por Gifi: a HOMALS (HOMogeneity analysis by Alternating Least Squares), que se mostra ser equivalente a uma Análise de Correspondências Múltiplas (Bekker e de Leeuw [2]). Igualmente importante foi o algoritmo CATPCA (CATegorical Principal Components Analysis, disponível no SPSS desde 1999), que podendo ser considerado uma evolução do *sistema Gifi*, contém como casos particulares a HOMALS e a ACP tradicional, tendo sido introduzidas as transformações *spline* como ferramenta de generalização (Meulman et al. [30]). Sublinhe-se que a CATPCA não se pode considerar da autoria de Gifi, mas de elementos associados durante os anos 1990 ao *Data Theory Scaling System Group* da Universidade de Leiden[1].

No capítulo seguinte apresentam-se alguns conceitos preliminares, nomeadamente resultados sobre transformações *spline*, bem como a formulação da *função perda* de forma suficientemente abrangente por forma a ser considerada o denominador comum da HOMALS, da CATPCA e da qlPCA. O capítulo 2 é dedicado a uma revisão da CATPCA partindo da HOMALS, pela importância que esta teve no algoritmo que se propõe nesta tese e com o intuito comparativo posterior. Mostra-se também no capítulo 2 que, de todas as técnicas provenientes do sistema *Gifi*, a HOMALS é a mais potente devido à flexibilidade permitida para as quantificações, surgindo a CATPCA,

[1] A CATPCA foi desenvolvida por Willem Heiser, Jacqueline Meulman, Gerda van den Berg, Patrick Groenen, Peter Neufeglise e Anita van de Kooij (Meulman e Heiser [29])

histórica e tecnicamente, como uma generalização da HOMALS.

Apesar de já contar com quase 15 anos de existência, a CATPCA continua a ser objecto de investigação, estando associada a várias publicações recentes[2]. Não estando associado à CATPCA, Jan de Leeuw, um dos elementos fundadores do grupo Gifi, e vários seus colaboradores, também têm realizado desenvolvimentos recentes que assentam no *sistema Gifi*, através do envolvimento em projectos de computação estatística associados ao *software* R (de Leeuw e Mair [8], Mair e de Leeuw [28]).

Sendo por definição uma variante da ACP para variáveis categoriais, a CATPCA apresenta-se também como um algoritmo para realizar uma ACP com variáveis das várias escalas de medida e em simultâneo como uma variante não-linear (Linting e Kooij [25]). O termo não-linear justifica-se para todas as abordagens associadas ao *sistema Gifi*, como é o caso da CATPCA, com a existência de uma transformação não-linear das variáveis originais aquando da *quantificação óptima*.

Usualmente designam-se por categoriais as variáveis qualitativas de nível de medida nominal ou ordinal. No entanto, os autores associados ao *sistema Gifi*, estendem o conceito para variáveis quantitativas, interpretando-as como um caso particular com várias categorias associadas a um número elevado de valores discretos.

Jan de Leeuw [6] refere que "[...] all data are categorical, although perhaps some data are more categorical than others." e Linting et al. [26] referem que "[...] even true numeric variables can be viewed as categorical variables with c categories, where c indicates the number of different observed values. Both ratio and interval variables are considered numeric [...]".

A pretensão de estender o *sistema Gifi* para variáveis quantitativas faz todo o sentido pois para obter determinados níveis de eficiência a ACP tradicional espera que subconjuntos de variáveis apresentem correlações lineares relativamente elevadas. Na presença de estruturas não-lineares, apenas numa situação ideal com transformações linearizantes conhecidas esta limitação te-

[2]Em 2012, Linting e Kooij [25], em 2007, Linting et al. [26, 27] e em 2004, Meulman et al. [30].

ria solução trivial. A *quantificação óptima* também pode ser interpretada como uma estimativa para a transformação linearizante desconhecida[3]. Mais, a análise gráfica dos valores originais da variável *versus* a sua *quantificação óptima* sugere o grau de não-linearidade presente no conjunto das variáveis, permitindo assim sinalizar a necessidade desta abordagem (Linting et al. [26], Linting e Kooij [25]).

No entanto, até à introdução de *splines* no seio do *sistema Gifi*, não era possível aos algoritmos existentes lidarem com um número elevado de categorias. Já em 1988, de Leeuw e van Rijckevorsel [9] referem que:

> "Variables with a large number of possible values, or even continuous variables, can be incorporated in theory, but the implementations of the techniques more or less expect a small number of categories. If the number of categories is very large, say close to the number of objects that are classified, then homogeneity analysis as currently implemented does not work very well. It will tend to produce unsatisfactory and highly unstable solutions, in which chance of capitalization is a major source of variation."

Num artigo de 2010, Mair e de Leeuw [28] referindo-se à sua implementação mais recente, sublinham que toda a teoria que associa os *splines* ao *sistema Gifi* está desenvolvida desde 1988 e detalhada em van Rijckevorsel e de Leeuw [35]. No entanto, Mair e de Leeuw [28] apresentam nas conclusões que:

> "Considering each variable as categorical is not very efficient when having many categories, as typically in the numerical case. Therefore, in a future update we will use splines to make it more efficient."

[3]Embora sugestiva, a expressão "estimativa para a transformação linearizante", não costuma ser usada na literatura, talvez pode ser demasiado forte. Por exemplo, Linting et al. [26] referem em alternativa que "Nonlinear PCA can assign values to the categories of such numeric variables that will maximize the association (Pearson correlation) between the quantified variables".

Assim, a única implementação actual dos *splines* no seio do *sistema Gifi* é a CATPCA. Em 2012, Linting e Kooij [25], pela primeira vez numa publicação associada a responsáveis pelo desenvolvimento da CATPCA, subscrevem as motivações apresentadas por Jan de Leeuw, para a introdução de *splines* no seio da CATPCA, referindo que:

> "If a variable has many categories (as, e.g., a continuous variable), and the researcher is interested in nonlinear relationships between that variable and other variables, nominal and ordinal analysis levels might lead to very irregular quantifications (going wildly up and down in the transformation plot) that lack insightfulness and stability. Alternatively, a more restrictive spline ordinal or spline nominal analysis level can be specified."

No entanto, a CATPCA, algoritmo originalmente concebido para variáveis categoriais (ordinais e nominais), necessita, *a priori*, de um processo de discretização quando aplicada a variáveis contínuas, sendo os *splines* aplicados *a posteriori*. Nesta tese defende-se que o processo de discretização é desnecessário num contexto em que todas as variáveis em análise são contínuas, propondo-se a qlPCA como alternativa. No entanto, num contexto de mistura de variáveis de várias naturezas pode não ser fácil dispensar a discretização, pensando-se que terá sido esse o motivo que levou o grupo de desenvolvimento da CATPCA a implementar os *splines* mantendo a discretização. Por este motivo, quando na literatura associada às ciências que lidam com variáveis contínuas aparecem referências à ACP não-linear (ACPNL) raramente é à CATPCA que se estão a referir mas sim à utilização de:

- redes neuronais, abordagem apresentada em 1991 por Kramer [16];
- *principal curves and manifolds*, abordagem apresentada em 1984 por Hastie [13];
- *kernel Principal Components Analysis* (KPCA), abordagem apresentada em 1998 por Schoolkopf et al. [32].

Assim, esta tese propõe uma solução para o seguinte problema:

> Tendo estas variantes da ACP sido desenvolvidas para variáveis categoriais, como adaptá-las de modo a serem consideradas como uma abordagem não-linear em contexto de variáveis contínuas?

Sublinhe-se que a qlPCA admite que todas as variáveis são contínuas, não sendo por isso uma alternativa à CATPCA mas antes uma solução para realizar uma ACP em estruturas não-lineares associadas exclusivamente a variáveis contínuas. Detalhes sobre a teoria da qlPCA e sobre a sua implementação serão apresentados no capítulo 3. Apresentam-se ainda dois exemplos de aplicação, um no final do capítulo 3 com base em dados simulados e outro com dados reais no capítulo 4. Conforme referido, a única implementação actual dos *splines* no seio do *sistema Gifi* é a CATPCA. Dadas as limitações referidas, considera-se que a qlPCA representa um contributo relevante, no seio do conjunto de abordagens que tiveram a sua génese no referencial teórico desenvolvido por Gifi.

A qlPCA está publicada em dois artigos recentes (Lavado e Calapez [22, 23]), tendo sido publicados ainda 3 trabalhos preliminares nos últimos anos (Lavado e Calapez [19, 20, 21]) bem como proferidas várias comunicações nacionais e internacionais em torno desta temática.

Capítulo 1

Preliminares

Este capítulo apresenta os conceitos preliminares para esta tese. Começa-se por fazer uma breve revisão das transformações *spline*, com particular destaque para a base de *I-splines*. Após introduzir na segunda secção o conceito de *homogeneidade* e de *quantificação óptima* através da definição de *função perda*, mostra-se na terceira secção deste capítulo que a ACP tradicional e a Análise de Correspondências Múltiplas (ACM) são casos particulares dum problema de optimização de uma determinada *função perda*. Mostra-se ainda que a introdução de transformações *spline* no seio da *função perda* permite um enquadramento daquelas técnicas como casos particulares e extremais em termos de flexibilidade, abrindo portas para uma vasta gama de variantes não-lineares. Parte desta secção está publicada nas Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística em Lavado e Calapez [20].

1.1 Transformações *spline*

Ainda sem o uso da palavra *spline,* a teoria das funções *spline*, remonta a Runge[1] (1901), tendo a terminologia "*função spline*" sido introduzida por

[1]Citado por Schumaker [33].

Schoenberg[2] em 1946, aquando da resolução de alguns problemas de ajustamento de dados. O termo *spline* foi escolhido por Schoenberg, devido à existência de um mecanismo mecânico com esse nome, usado por arquitectos (em especial arquitectos navais) para desenhar curvas suaves passando por pontos dados[3]. As primeiras monografias sobre o tema, que são as referências bibliográficas de eleição, surgiram no final dos anos 1960 e durante as décadas de 1970 e 1980. Os autores das variantes não-lineares da ACP, que usam funções *spline*, não hesitam em indicar a monografia de De Boor [5] e a de Schumaker [33] como as referências fundamentais. O livro de De Boor [5] contém, para além do desenvolvimento teórico, os primeiros algoritmos para lidar computacionalmente com funções *spline*. Estes dois autores foram as principais referências usadas.

Ainda que informalmente, fixe-se desde já que as funções *spline* são funções seccionalmente polinomiais com certas regularidades. Note-se que também se podem definir funções *spline* sobre funções trigonométricas ou sobre funções exponenciais, entre outras. Não serão, no entanto, feitas quaisquer considerações sobre funções *spline* não polinomiais pois não são necessárias para os objectivos deste capítulo.

Considere-se o intervalo $[a, b]$ e seja

$$\Delta = \{\varepsilon_i\}_{i=0}^{r+1} \text{ com } a = \varepsilon_0 < \varepsilon_1 < \ldots < \varepsilon_r < \varepsilon_{r+1} = b \tag{1.1}$$

uma colecção de pontos que definem uma sua partição em $r+1$ subintervalos

$$I_i = [\varepsilon_i, \varepsilon_{i+1}[\,,\ i = 0, 1, \ldots, r-1 \text{ e } I_r = [\varepsilon_r, \varepsilon_{r+1}]\,. \tag{1.2}$$

Os r pontos interiores de Δ designam-se por pontos de junção.

[2]Nos artigos "Contributions to the problem of approximation of equidistant data by analytic functions, Part A: On the problem of smoothing of graduation, a first class of analytic approximation formulae" e "Contributions to the problem of approximation of equidistant data by analytic functions, Part B: On the problem of osculatory interpolation, a second class of analytic approximation formulae", ambos publicados na *Quartely Applied Mathematics*, 4, páginas 45-99 e 112-141, respectivamente.

[3]Para mais pormenores sobre a história do desenvolvimento da teoria das funções *spline*, ver Schumaker [33].

Usa-se a notação $f_i = f|_{I_i}$, com f uma função definida em $[a, b]$, para designar a restrição de f ao intervalo I_i. Usa-se a notação $f_i^{(l)}$ para a derivada de ordem l da função f_i.

Uma função f, definida em $[a, b]$, é um *spline* de grau v (ordem $v + 1$), com pontos de junção $\varepsilon_1, \ldots, \varepsilon_r$ se, e só se, para todo o $i = 0, \ldots, r$

$$f_i \text{ é um polinómio de grau } v \tag{1.3}$$

e, para todo o $i = 0, \ldots, r - 1$,

$$\lim_{x \to \varepsilon_{i+1}^-} f_i^{(l)}(x) = \lim_{x \to \varepsilon_{i+1}^+} f_{i+1}^{(l)}(x),\ l = 0, 1, \ldots, v - 1. \tag{1.4}$$

A condição (1.3) significa que f tem que ser um polinómio de grau v em cada subintervalo, apesar de poder ser um polinómio diferente para cada subintervalo. As funções que verificam apenas esta condição são chamadas polinómios segmentados de grau v ou funções seccionalmente polinomiais de grau v.

A condição (1.4) assegura que os diferentes polinómios se ligam de forma mais, ou menos, suave nos pontos de junção. Esta suavidade faz com que na representação gráfica de um *spline* os pontos de junção sejam imperceptíveis. Esta condição é a exigência mais comum quando se recorre a funções *spline*. No entanto, a literatura apresenta uma definição mais geral, permitindo que a suavidade do *spline* não seja a mesma em todos os pontos de junção. A condição (1.4) é a usada nas variantes não-lineares da ACP em análise.

Um caso de particular importância é o *spline cúbico*, ou seja, $v = 3$ na definição anterior. Como os splines cúbicos têm derivadas contínuas até à segunda ordem, são particularmente interessantes na modelação de fenómenos físicos, em que a segunda derivada representa a aceleração.

Pode-se demonstrar (De Boor [5], Schumaker [33]) que o conjunto dos *splines* de grau v, com a partição definida por Δ, é um espaço linear de funções, de dimensão $w = v + 1 + r$, ou seja, de dimensão igual à ordem do *spline* somada ao número de nós interiores.

Em 1966, Curry e Schoenberg (citados por De Boor [5]) demonstraram que existe uma base, computacionalmente interessante, para o espaço linear

dos *splines* polinomiais. As funções desta base foram designadas por *B-splines*. Para esse efeito, a partição Δ e a informação sobre a suavidade em cada um dos seus pontos é incorporada numa *sequência de nós* não decrescente $t = \{t_1, t_2, \ldots, t_{2v+r+2}\}$, onde:

1. $t_1 \leq \ldots \leq t_{2v+r+2}$;
2. $t_1 = \cdots = t_{v+1} = a$;
3. $t_{v+r+2} = \cdots = t_{2v+r+2} = b$;
4. $t_{v+2}, \ldots, t_{v+r+1}$ são os r nós interiores.

Assim, quanto menos suave for o ponto de Δ mais nós lhe estão associados, sendo esta multiplicidade de nós necessária para a recursividade. Consequentemente, segundo a condição (1.4), apenas os pontos fronteiros de Δ, $\varepsilon_0 = a$ e $\varepsilon_{r+1} = b$, vão estar associados (cada um) a $v+1$ nós, pois o *spline* não está sujeito às exigências de regularidade impostas aos pontos de junção interiores. Logo, dados os $r+2$ pontos de Δ são necessários $r + 2(v+1) = r + 2v + 2 = 2v + r + 2$ nós.

Dada a sequência de nós t, define-se para todo $q = 1, 2, \ldots, w$ a função *B-spline* de grau v pela seguinte relação recursiva

$$B_q^{[1]}(x) = \begin{cases} 1\ , & t_q \leq x < t_{q+1} \\ 0\ , & \text{caso contrário} \end{cases},$$

$$B_q^{[v+1]}(x) = \frac{x - t_q}{t_{q+v} - t_q} B_q^{[v]}(x) + \frac{t_{q+v+1} - x}{t_{q+v+1} - t_{q+1}} B_{q+1}^{[v]}(x),$$

onde

$$\frac{x - t_q}{t_{q+v} - t_q} B_q^{[v]}(x) \quad \text{e} \quad \frac{t_{q+v+1} - x}{t_{q+v+1} - t_{q+1}} B_{q+1}^{[v]}(x)$$

são iguais a zero quando os denominadores são nulos.

Uma outra base particularmente interessante é a base dos *M-spline*,

$$M_q^{[v+1]} = \frac{v+1}{t_{q+v+1} - t_q} B_q^{[v+1]} \quad , \quad q = 1, \ldots, w. \tag{1.5}$$

Pode ser demonstrado (Schumaker [33]) que $M_q^{[v+1]}$ é positiva e inferior a 1 no intervalo $]t_q, t_{q+v+1}[$, sendo zero caso contrário, e também que

$$\int_{-\infty}^{\infty} M_q^{[v+1]}(x)\ dx = 1.$$

logo, $M_q^{[v+1]}$ tem as características de uma função densidade probabilidade.

Transformações monótonas podem ser obtidas usando uma base de funções de *spline* monótonas juntamente com coeficientes não negativos. Nesse sentido, Winsberg e Ramsay [36] aproveitando as propriedades das funções densidade probabilidade *M-spline*, propuseram a base de *I-splines* (*integrated splines*), as correspondentes funções distribuição da base de *M-splines*.

Dada a sequência de nós $\{t\}$ os *I-spline de ordem* $v+2$ são definidos, para todos os $q = 1, 2, \ldots, w$ por

$$I_q^{[v+2]}(x) = \int_{-\infty}^{x} M_q^{[v+1]}(u)\ du. \tag{1.6}$$

Como cada *M-spline* é um polinómio segmentado de grau v, o *I-spline* associado é um polinómio segmentado de grau $v+1$, ou ordem $v+2$. Assim, o espaço associado tem dimensão $w+1$, sendo w a dimensionalidade da base dos *M-splines*. No entanto, por construção, existem apenas w *I-splines* independentes, por isso só é possível obter o subespaço gerado por estes. A partir de agora w refere-se à dimensão deste subespaço, sendo $w = v+r$ para um *spline* de grau v com r nós interiores.

Pela definição (1.6), conclui-se que $I_q^{[v+2]}$ é não constante no intervalo $]t_q, t_{q+v+1}[$, sendo zero para valores inferiores a t_q, e 1 para valores superiores a t_{q+v+1}. A definição dos *I-splines* através de uma relação recursiva fornece uma conveniente abordagem computacional.

Como exemplo ilustrativo, considere-se uma variável contínua x de mínimo m_1, mediana m_2 e máximo m_3 e o espaço dos *splines* lineares com um nó interior na mediana. Os elementos da base são os seguintes:

$$I_1(x) = \begin{cases} 0, & x < m_1 \\ \frac{x-m_1}{m_2-m_1}, & m_1 \leq x < m_2 \\ 1 & x \geq m_2 \end{cases}$$

$$I_2(x) = \begin{cases} 0, & x < m_2 \\ \frac{x-m_2}{m_3-m_2}, & m_2 \leq x < m_3 \\ 1 & x \geq m_3 \end{cases}$$

Para obter por recorrência esta base de *I-splines* é necessário em primeiro lugar calcular as funções da base de *M-splines*. Como se pretende obter *I--splines* de grau 1, as funções *M-spline* que lhes dão origem têm grau zero. A base dos *M-splines* que gera todo o espaço de *splines* de grau $v = 0$ com $r = 1$ nó interior tem dimensão $w = v + 1 + r = 2$. Por isso, o conjunto das funções da base de *I-splines* terá dois elementos. Como o espaço das funções *spline* de grau 1 com um nó interior é tridimensional, o conjunto referido pode apenas gerar um seu subespaço.

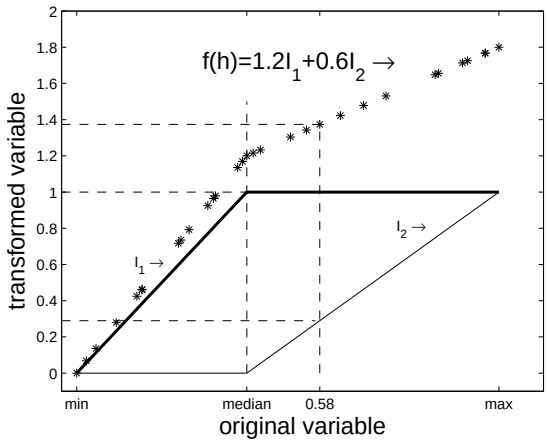

Figura 1.1: *Spline* de grau um com um nó interior na mediana.

A figura 1.1 mostra a família de *I-splines* de grau 1 definida em $[0, 1]$ com um nó interior na mediana. I_1 e I_2 são as funções da base. As estrelas representam as imagens de x através do *spline* obtido como combinação linear de I_1 e I_2 com os coeficientes 1.2 e 0.6, respectivamente. Cada *I-spline* é seccionalmente linear e não constante num intervalo. A figura também

mostra um exemplo de imagens obtidas em cada uma das três funções (duas funções da base e um *spline*) para um valor de x superior à mediana ($x = 0.58$, $I_1(0.58) = 1$, $I_2(0.58) = 0.29$ e $f(0.58) = 1.37$).

Um polinómio de grau v é determinado por $v + 1$ pontos. Por isso, cada segmento do *spline* na figura 1.1 pode ser definido usando os pontos associados ao mínimo/mediana e mediana/máximo. No entanto, o algoritmo da qlPCA irá optimizar a escolha dos coeficientes da combinação linear de *I-splines* através de uma regressão multivariada, sendo I_1 e I_2 as variáveis predictoras e por isso envolvendo todos os dados e não apenas conjuntos de dois pontos.

1.2 Função perda

O conceito de *homogeneidade* é central no *sistema Gifi* . Em termos genéricos, k variáveis serão homogéneas se forem semelhantes, ou seja, se medirem a mesma propriedade ou propriedades. Para o concretizar, torna-se necessário definir o que se entende por semelhança entre variáveis. Tal passará necessariamente pela definição prévia duma forma de comparação/medida das variáveis.

Suponha-se que existe uma variável $\mathbf{x}$ que mede aproximadamente a mesma propriedade que um conjunto de variáveis, ou seja, funciona como um índice dessa propriedade (por exemplo, o Índice de Preços no Consumidor). A ideia subjacente à medição da semelhança/diferença entre as variáveis é quantificar a perda de informação inerente à substituição de um conjunto de variáveis por um índice. Quando a perda de informação é mínima diz-se que as variáveis iniciais são homogéneas podendo ser substituídas pelo índice.

A questão central que permite generalizar a ACP tradicional é saber até que ponto será vantajoso transformar as variáveis iniciais, considerando determinadas classes de transformações admissíveis, antes de quantificar a perda de informação, com o objectivo de maximizar a semelhança entre elas e consequentemente minimizar a perda.

As diferentes técnicas que dão resposta ao problema anterior diferem na

forma de medição da semelhança/diferença entre as variáveis e nas classes de transformações admitidas.

A ACP tradicional tem como objectivo reter o máximo de informação das variáveis originais nas componentes principais retidas. As componentes principais, multiplicadas por um conjunto de pesos óptimos designados por *loadings*, devem aproximar tanto quanto possível os dados originais. Usualmente na ACP tradicional, os *loadings* e as componentes principais são obtidos através de uma decomposição em valores singulares da matriz de dados estandardizada, ou através de uma decomposição em valores e vectores próprios da matriz de correlação. No entanto, os mesmos resultados podem ser obtidos através de um processo iterativo no qual uma *função perda* é minimizada. A perda a ser minimizada é a perda de informação inerente à representação das m variáveis por um número reduzido de componentes principais. Esta é quantificada pela diferença entre as componentes principais, ponderadas pelos respectivos *loadings*, e as variáveis originais.

Seja $\mathbf{H}$ a matriz dos dados centrada do tipo $n \times m$, formada por colunas $\mathbf{h}_j$ $(j = 1, \ldots, m)$, que reúnem os valores observados para cada variável em análise. Seja $\mathbf{X}$, do tipo $n \times p$, a matriz das p componentes principais retidas e $\mathbf{A}$, do tipo $p \times m$, a matriz dos *loadings*, sendo a j-ésima coluna designada por $\mathbf{a}_j$.

Nas variantes da ACP associadas ao *sistema Gifi*, o processo iterativo para minimização da perda permite adicionalmente transformações não-lineares das variáveis originais. Estas variantes não-lineares são por isso uma generalização da ACP, ao permitirem no mesmo modelo, a substituição da matriz $\mathbf{H}$ dos dados observados pela matriz $\mathbf{F}$, do tipo $n \times m$, contendo as variáveis transformadas $\mathbf{f}_j = \phi_j(\mathbf{h}_j)$ cujos valores são designados quantificações.

A função perda usada na ACP tradicional para minimizar a diferença entre os dados originais e as componentes principais pode nesta generalização ser escrita como $\sigma(\mathbf{X}, \mathbf{F}) = n^{-1} \sum_j \sum_i (\sum_s x_{is} a_{sj} - f_{ij})^2$, ou em notação matricial como

$$\sigma(\mathbf{X}, \mathbf{F}) = n^{-1} \sum_j tr[(\mathbf{X}\mathbf{a}_j - \mathbf{f}_j)'(\mathbf{X}\mathbf{a}_j - \mathbf{f}_j)]. \tag{1.7}$$

Pode demonstrar-se (Gifi [12]) que a função perda (1.7) é equivalente a

$$\sigma(\mathbf{X}, \mathbf{F}) = n^{-1} \sum_j tr[(\mathbf{X} - \mathbf{f}_j \mathbf{a}_j^{'})^{'} (\mathbf{X} - \mathbf{f}_j \mathbf{a}_j^{'})], \tag{1.8}$$

sendo esta formulação a mais usual no *sistema Gifi*. A função perda (1.8) está especificada para a situação mais simples, sem valores omissos ou a possibilidade de diferentes pesos para os indivíduos. Estes casos podem ser incorporados na função perda (1.8), no entanto, estão fora do âmbito desta tese.

No próximo capítulo será abordada a possibilidade de atribuir p quantificações diferentes, uma por cada dimensão retida, a cada valor original. É o que se costuma designar por tratamento *Multiple* (de Leeuw e van Rijckevorsel [9], Gifi [12], Meulman et al. [30]). Por forma a permitir esta possibilidade, a função perda (1.8) é generalizada para a função $\sigma : \mathrm{M}_{n\times p} \times \mathrm{M}_{pm\times n} \rightarrow \mathbb{R}$ tal que

$$\sigma\left(\mathbf{X}, \mathbf{M}\right) = n^{-1} \sum_j tr\left[\left(\mathbf{X} - \mathbf{M}_j\right)^{'} \left(\mathbf{X} - \mathbf{M}_j\right)\right], \tag{1.9}$$

sujeita à restrição $\mathbf{X}^{'}\mathbf{X} = n\mathbf{I}$, onde

- $\mathbf{M}_j = \left[\begin{array}{ccc} \mathbf{m}_{j1} & \ldots & \mathbf{m}_{jp} \end{array}\right]$ é a matriz $n \times p$ contendo as p (diferentes) imagens do mesmo vector $\mathbf{h}_j$;
- $\mathbf{m}_{jt}$ é o vector das imagens de $\mathbf{h}_j$, obtidas através da transformação ϕ_{jt} da variável j para a dimensão t, $t = 1, \ldots, p$;
- $\mathbf{M} = \left[\begin{array}{ccc} \mathbf{M}_1 & \ldots & \mathbf{M}_m \end{array}\right]^{'}$ é uma matriz $pm \times n$.

Sublinhe-se que a matriz $\mathbf{M}_j$ contém as imagens dum mesmo vector $\mathbf{h}_j$, sujeito a p transformações diferentes. Quando se pretende impor às quantificações obtidas pela minimização da função perda, restrições de ordem, de distância e de disposição segundo certas regras, exige-se durante o processo de optimização que a matriz $\mathbf{M}_j$ tenha característica unitária, ou seja,

$\mathbf{M}_j = \mathbf{f}_j \mathbf{a}_j'$ tal como na função (1.8). É o que se costuma designar por tratamento *Single* (de Leeuw e van Rijckevorsel [9], Gifi [12], Meulman et al. [30]).

A função perda (1.9) é suficientemente abrangente para incluir como casos particulares a função perda da HOMALS, da CATPCA e da qlPCA. Nesse sentido, a função perda (1.9) será designada por *função perda comum*. Como se mostra no capítulo seguinte a HOMALS usa o tratamento *Multiple* para todas as variáveis e a CATPCA permite um tratamento diferenciado por variável permitindo ambos os tratamentos. No capítulo 3 mostra-se que a qlPCA usa o tratamento *Single* para todas as variáveis.

1.3 Um enquadramento via *splines* das variantes da ACP

Considere-se na função perda (1.8) que o vector $\mathbf{f}_j$ das imagens de $\mathbf{h}_j$ está associado à seguinte transformação:

$$\mathbf{f}_j = \phi_j(\mathbf{h}_j) = \sum_{i=1}^{w} \alpha_i I_i^{[v]}(\mathbf{h}_j), \tag{1.10}$$

ou seja, ϕ_j é uma função *spline* de grau v, com r nós interiores, gerado por w *I-splines*.

A transformação (1.10) contempla os seguintes casos particulares:

1. *Transformações lineares*
   Obtêm-se considerando *splines* de grau 1, sem nós interiores, o que resulta, como pretendido, num vulgar polinómio de grau 1.

2. *Transformações seccionalmente constantes*
   Obtêm-se considerando *splines* de grau 0, com qualquer número de nós interiores, o que resulta, como pretendido, numa função em patamares. No caso extremo sem nós interiores obtém-se uma função constante, sem utilidade neste contexto. No extremo oposto, com número máximo de nós interiores, obtém-se uma função em que cada objecto está associado a um patamar.

Estes dois casos particulares podem ser considerados extremais no que diz respeito à quantidade de nós interiores.

Considere-se agora que a classe de transformações admissíveis é a das funções lineares para todas as variáveis. Neste caso, pretende-se determinar, para cada variável, um escalar óptimo y_j tal que a *quantificação óptima* $\mathbf{f}_j = \phi_j(\mathbf{h}_j) = y_j\mathbf{h}_j$ minimize a função perda. Pode demonstrar-se (Bekker [2], Escofier e Pagès [10], Lavado [19]), que nestas circunstâncias a minimização da função perda é equivalente à solução da ACP tradicional, sendo assim mais uma fundamentação para designar tal técnica de linear. Sabe-se ainda que, neste contexto, a *quantificação óptima* $\mathbf{f}_j$ para a variável j é a estandardização dessa variável, ou seja, y_j é o inverso do desvio-padrão (Meulman et al. [30]). Sendo as transformações lineares funções *spline* de grau 1, sem nós interiores, pode-se afirmar que a ACP linear pode ser vista como caso particular do *sistema Gifi* via funções *spline*.

Considere-se agora que a matriz dos dados $\mathbf{H}_{n\times m}$ é constituída por n observações em m variáveis qualitativas nominais, onde a j-ésima variável tem k_j categorias ($j = 1, \ldots, m$). Estas variáveis, aquando da sua introdução na matriz dos dados, são alvo dum processo de quantificação *a priori*, chamado *codificação*, que associa, por tradição, os primeiros números inteiros positivos a cada uma das categorias de cada variável. Usando este tipo de codificação, o vector de observações terá sempre elementos que variam entre 1 e k_j, percorrendo apenas números naturais.

A utilização de uma ACP linear sobre variáveis nominais limita a flexibilidade da optimização da homogeneidade pelo uso de transformações que mantêm a proporcionalidade entre as distâncias associadas à quantificação atribuída *a priori*. No entanto, essa quantificação é perfeitamente arbitrária, pelo que, é legítimo flexibilizar a optimização de modo a maximizar a homogeneidade entre as categorias das variáveis em análise. No caso de variáveis nominais, verifica-se um relacionamento entre as *quantificações óptimas* e os *object scores* que remetem para o *princípio baricêntrico* (Bekker e de Leeuw [2], Tenenhaus e Young [34]), ou seja, como se mostra no próximo capítulo, a quantificação de uma categoria será proporcional à média dos *scores* dos objectos que a ela estão associados, e o *score* de um objecto será proporcional

à média das quantificações das categorias que o caracterizam.

Na Análise de Correspondências Múltiplas (ACM, ver por exemplo Carvalho, [3]) tradicional é efectuada uma decomposição adequada da matriz disjuntiva completa, matriz binária do tipo $n \times \sum_{j=1}^{m} k_j$, resultante da justaposição de m matrizes $\mathbf{G}_j$ do tipo $n \times k_j$, onde $g_j(i,s) = 1$ sse o i-ésimo indivíduo optou pela s-ésima categoria da j-ésima variável. Ou seja, na matriz disjuntiva completa, cada coluna corresponde a uma categoria de uma variável. Fazendo o paralelismo com o explanado no caso da ACP linear, é como se, agora, as variáveis indicatrizes das categorias desempenhassem o papel das variáveis. Pode por essa via demonstrar-se (Bekker e de Leeuw [2], Escofier e Pagès [10], Lavado [19]) que o recurso a transformações seccionalmente constantes com número máximo de nós interiores para cada variável, tem como consequência a equivalência entre os resultados provenientes duma ACM e a minimização da função perda (1.9) no seio dessa classe de transformações. Sendo essas transformações funções *spline* de grau 0, com o número máximo de nós interiores, pode afirmar-se que a ACM pode ser vista como um caso particular do *sistema Gifi* via funções *spline*.

Os dois casos particulares anteriores (ACP e ACM tradicionais) contemplam valores extremos para o número de nós interiores e os polinómios segmentados que constituem esses *splines* têm grau mínimo. As transformações lineares são as mais rígidas, pois só permitem transformar a variável/vector num vector colinear, ou seja, a distância entre os valores da variável transformada é proporcional à distância entre os valores originais. As transformações seccionalmente constantes são as mais livres, pois permitem actuar em subconjuntos de componentes do vector/variável sem qualquer restrição.

Tanto a ACP linear como a ACM consideram a mesma classe de transformações para todas as variáveis. Para aumentar a flexibilidade desta abordagem é desejável que a classe de transformações admissíveis, isto é a classe de *splines*, seja escolhida variável a variável, o que se torna possível devido à separabilidade em j da função perda.

As ideias fundamentais das variantes não-lineares da ACP são:

- enfraquecer a rigidez das transformações lineares; e

- restringir a liberdade das transformações seccionalmente constantes.

Sublinhe-se que as situações intermédias podem ser obtidas pela utilização de outros tipos de *splines* fazendo variar quer o número de nós interiores quer o grau do *spline*. Esta abordagem permite um compromisso entre a flexibilização das transformações admissíveis tendo em vista a minimização da perda e a restrição das mesmas devido à informação que se dispõe *a priori* sobre a variável original e que se pretende manter. Por exemplo:

- Habitualmente, se a escala de medida da variável é ordinal, pretendem-se impor restrições na ordem dos valores da variável transformada, i.e., permite-se a alteração não proporcional dos valores da variável original mas não a alteração da ordem entre eles. Isto significa que a transformação *spline* a optimizar deve ser uma função monótona crescente.

- Se a variável é contínua normalmente pretendem-se restrições ao nível da proporcionalidade da distância entre os valores da variável que podem ser ditadas pela escolha dos parâmetros do *spline*.

- Se existirem razões para suspeitar da existência de relações não-lineares entre as variáveis, pode impor-se outro tipo de restrições à variável transformada, através da parametrização adequada do *spline*.

Para ilustrar as ideias anteriores considerem-se os seguintes exemplos:

1. Lavado e Calapez [23], recorrendo a dados simulados que serão discutidos em mais detalhe no terceiro capítulo, consideram doze variáveis, sendo dez construídas através de funções monótonas, não-lineares, das restantes duas variáveis. Trata-se de uma situação em que a aplicação de transformações logarítmicas em todas as variáveis, resulta numa matriz de característica dois. Consequentemente, espera-se que uma ACP não-linear sobre a matriz inicial, bem como uma ACP linear sobre a matriz transformada, apresente um ajustamento quase perfeito com apenas duas dimensões e que as transformações *spline* óptimas revelem um comportamento aproximadamente logarítmico. Concluiu-se que todas as vantagens que decorreriam da realização duma ACP linear

sobre as variáveis linearizadas via aplicação de logaritmos, são também obtidas via aplicação directa da ACP não-linear sobre a matriz inicial.

2. Num estudo sobre segurança rodoviária (Gifi, [11]), observam-se as idades dos inquiridos arredondadas para o número inteiro de anos mais próximo. Assim, a variável "idade do condutor" é definida pelo investigador como sendo de natureza quantitativa (discreta). É provável que esta variável esteja relacionada com a segurança rodoviária de forma não-linear: directamente proporcional até aos 25 anos e depois sem influência relevante ou, noutro cenário, directamente proporcional até aos 25 anos, sem influência relevante dos 25 aos 65 e inversamente proporcional depois dos 65 anos. Em qualquer dos casos, a transformação linear limitaria em demasia a qualidade do ajustamento. O investigador pode por isso optar por usar uma transformação não-linear. Transformações *spline* com e sem a exigência de monotonia, de grau 1 com 2 nós interiores, são as opções indicadas para o primeiro e segundo cenário, respectivamente. Se o investigador tivesse considerado a variável "classe etária do condutor", transformações seccionalmente constantes (*spline* de grau 0) poderiam ser interessantes para esta variável.

3. Num estudo de opinião sobre o aborto (adaptado de Gifi [11]) observam-se as preferências políticas dos inquiridos. Os partidos políticos com representação parlamentar são codificados com os primeiros 6 números inteiros positivos, usando ainda o número 7 para outras opções. Assim, a variável "preferência política" é definida pelo investigador como sendo de natureza qualitativa nominal. Uma questão que pode provocar incerteza no investigador é se deve considerar todos os partidos individualmente, ou agrupar alguns. Em vez de decidir com base no seu conhecimento *a priori*, pode fazê-lo depois da análise: se algumas categorias obtêm a mesma quantificação, ou uma muito semelhante, então poderão ser agrupadas sem afectar os resultados finais. Se a codificação inicial tiver explícita uma ordenação por quadrante político (p.ex. usando os códigos 1 a 2 para partidos de esquerda, 3 para partidos de centro e 4 a 6 para partidos de direita) o investigador pode es-

tar interessado numa transformação que respeite a ordem inicialmente escolhida para os partidos, mas que sugira a melhor distância entre eles, usando para esse efeito transformações monótonas seccionalmente constantes, tendo em vista a manutenção da ordem inicial.

Com a rapidez dos meios computacionais actuais, uma boa sugestão para obter a mais fiel redução de dimensionalidade é experimentar várias possibilidades em termos de transformações das variáveis iniciais.

Capítulo 2

CATPCA - uma breve revisão

O objectivo deste capítulo é fazer uma breve revisão da CATPCA (CATegorical Principal Components Analysis) pela importância que esta teve para o algoritmo proposto nesta tese. A CATPCA é uma variante não-linear da Análise em Componentes Principais, especialmente concebida para variáveis nominais e ordinais, que está preparada para estruturas não-lineares. Em primeiro lugar será apresentada a HOMALS (HOMogeneity analysis by Alternating Least Squares), que opera com tratamento *Multiple* para todas as variáveis e apenas permite transformações seccionalmente constantes, surgindo a CATPCA como forma de generalização, permitindo diferentes tratamentos por variável e alargando a classe de transformações admissíveis.

Considere-se que a matriz dos dados $\mathbf{H}_{n\times m}$ é constituída por m variáveis categoriais (variáveis qualitativas nominais ou ordinais), onde a variável $\mathbf{h}_j$ tem k_j categorias (valores possíveis), com $j = 1, \ldots, m$.

As variáveis categoriais, aquando da sua introdução na matriz dos dados, são alvo dum processo de quantificação *a priori,* usualmente designada *codificação,* que associa, por tradição, os primeiros números inteiros positivos a cada uma das categorias de cada variável. Por exemplo, para a variável "cor dos olhos", cujas categorias seriam, "castanhos", "azuis", "verdes" e "outras", uma possível codificação seria: "castanhos código 1", "azuis código

2", "verdes código 3", "outras código 4". Para certo tipo de análises exploratórias esta quantificação é suficiente, ou seja, perguntas do tipo "que percentagem de indivíduos têm os olhos azuis?" são exactamente equivalentes a perguntar à base de dados "que percentagem de 2's há na variável "cor dos olhos"?". Numa análise como a ACP pode-se levantar uma série de questões quanto à sua implementação sobre uma matriz de dados obtida via codificação. Basta lembrar o papel de cada variável como vector de $\mathbb{R}^n$. Face a quantificação sugerida, cada elemento de um desses vectores será sempre um número natural entre 1 e k_j. A aplicação duma ACP a uma matriz de dados desta natureza é algo substancialmente restritivo e desprovido de sentido. Assim, pretende-se determinar qual seria a quantificação óptima, para os fins duma ACP, a aplicar a cada categoria de cada variável, ou seja, às $\sum_{j=1}^{m} k_j \, (= \sum k_j)$ categorias.

A procura desta quantificação óptima será realizada via maximização da homogeneidade, ou analogamente, via minimização duma função perda. Para implementar esta ideia será necessário definirem-se algumas matrizes auxiliares.

Chama-se *matriz indicatriz* associada à variável $\mathbf{h}_j$, $j = 1, \ldots, m$, e designa-se por $\mathbf{G}_j$, à matriz do tipo $n \times k_j$, com entradas $G_j(i,t) = 1$, $i = 1, \ldots, n$, $t = 1, \ldots, k_j$ se o indivíduo i pertence à categoria t e $G_j(i,t) = 0$ se pertence a outra categoria.

Exemplo:

Se a variável $\mathbf{h}_1$ for $\mathbf{h}_1 = \begin{bmatrix} 1 & 4 & 5 & 3 \end{bmatrix}'$ então a matriz indicatriz que lhe está associada será:

$$\mathbf{G}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

A matriz indicatriz contém exactamente a mesma informação que a variável categorial que lhe está associada e está, de certa forma, dissociada da codificação tradicional. Por definição as categorias são mutuamente exclusivas e exaustivas, logo:

i) Cada linha de $\mathbf{G}_j$ é constituída por um elemento "1" e $(k_j - 1)$ elementos "0";

ii) Das k_j colunas de $\mathbf{G}_j$ uma delas é redundante pois fica totalmente determinada pelas restantes $(k_j - 1)$;

iii) As colunas de $\mathbf{G}_j$ são ortogonais entre si, logo $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$ é uma matriz diagonal cujos elementos principais são a frequência de cada categoria.

Chama-se *super-matriz indicatriz* associada à matriz dos dados $\mathbf{H}$, e designa-se por $\mathbf{G}$, à matriz do tipo $n \times \sum k_j$ formada pela justaposição das matrizes $\mathbf{G}_j$:

$$\mathbf{G} = \left[\begin{array}{cccc} \mathbf{G}_1 & \mathbf{G}_2 & \cdots & \mathbf{G}_m \end{array} \right].$$

A super-matriz indicatriz contém exactamente a mesma informação que a matriz $\mathbf{H}$. O somatório dos elementos de cada linha de $\mathbf{G}$ é igual a m. O somatório dos elementos de cada coluna de $\mathbf{G}$ indica a frequência de cada uma das $\sum k_j$ categorias.

2.1 HOMALS - a base da CATPCA

Na HOMALS a classe de transformações admissíveis é a das transformações seccionalmente constantes e todas as variáveis têm tratamento *Multiple*. Assim, cada variável está associada a p quantificações distintas, onde p é o número de componentes principais retidas.

As k_j categorias da variável $\mathbf{h}_j$ são representadas por k_j valores numéricos em p vectores distintos $\mathbf{y}_{jt}$, $t = 1, \ldots, p$, sendo a quantificação da variável $\mathbf{h}_j$ associada à t-ésima componente principal dada por $\mathbf{f}_{jt} = \phi_{jt}(\mathbf{h}_j) = \mathbf{G}_j\mathbf{y}_{jt}$.

Seja $\mathbf{Y}_j = [\mathbf{y}_{j1} \ldots \mathbf{y}_{jp}]$ a matriz do tipo $k_j \times p$ que representa as k_j categorias da variável $\mathbf{h}_j$ nas p dimensões retidas. A quantificação múltipla da variável $\mathbf{h}_j$ é dada por

$$\mathbf{M}_j = \mathbf{G}_j\mathbf{Y}_j.$$

Seja $\mathbf{X}_{n\times p}$ a matriz dos *object scores* e $\mathbf{Y}_j$'s do tipo $k_j \times p$, $j = 1, \ldots, m$ as matrizes das quantificações das categorias. As coordenadas do objecto i no espaço de dimensão p, correspondem aos elementos da linha i de $\mathbf{X}$,

$i = 1, \ldots, n$. As coordenadas da categoria c no espaço correspondem aos elementos da linha c de $\mathbf{Y}_j$, $c = 1, \ldots, k_j$. Ou seja, os *object scores* são os representantes dos objectos observados e as quantificações das categorias as representantes das categorias iniciais. Assim, a função perda comum (1.9), é na HOMALS escrita com o formato:

$$\sigma(\mathbf{X}, \mathbf{Y}) = m^{-1} \sum_j tr\left[(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)'(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)\right], \tag{2.1}$$

onde $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \ldots & \mathbf{Y}_m \end{bmatrix}'$.

Está demonstrado (Bekker e de Leeuw [2], Gifi [12], Lavado [19]) que cada solução $(\mathbf{x}, \mathbf{y})$ resultante da minimização da função perda para variáveis categoriais sujeita a $\mathbf{X'X} = \mathbf{I}$ satisfaz as seguintes relações de proporcionalidade:

$$\begin{aligned} \text{a) } \mathbf{x} &\propto \mathbf{Gy}/m \\ \text{b) } \mathbf{y} &\propto \mathbf{D}^{-1}\mathbf{G'x} \end{aligned}$$

onde, $\mathbf{D}$ é a diagonal de $\mathbf{G'G}$.

Mais, a condição[1] $\mathbf{X}'\mathbf{X} = \mathbf{I}$ implica que a constante de proporcionalidade de b) seja unitária, ou seja, $\mathbf{y} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{x}$.

A relação $\mathbf{y} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{x}$ significa que a quantificação de uma categoria é igual à média dos *object scores* que nela se inserem. Do ponto de vista geométrico pode-se então afirmar que a quantificação de uma categoria é o centro de gravidade, ou baricentro, dos *object scores* que nela se inserem. Da mesma forma, no caso de dimensão p, diz-se que a quantificação de uma categoria é o centróide dos *object scores* que nela se inserem.

De forma análoga a relação $\mathbf{x} \propto \mathbf{Gy}/m$ significa que os *object scores* são proporcionais à média das quantificações das categorias a que cada objecto pertence.

No contexto da HOMALS, o processo de *Mínimos Quadrados Alternados* consubstancia-se no *Princípio das médias recíprocas* para determinar $\mathbf{x}$ e $\mathbf{y}$, pois devido às relações anteriores procede, de forma simplificada, segundo os passos seguintes:

[1]Note-se que se a normalização tivesse sido $\mathbf{y'Dy} = 1$ teríamos outras constantes de proporcionalidade (ver por exemplo página 11 de Bekker e de Leeuw [2]).

(1) *Inicialização*:

a) $\mathbf{x}$ é construído de forma aleatória;

b) $\mathbf{x}$ é normalizado[2] e centrado e designado por $\widetilde{\mathbf{x}}$;

c) $\mathbf{y}$ é determinado pela média dos *object scores* que se inserem em cada categoria e designado por $\widetilde{\mathbf{y}}$.

(2) *Actualização dos object scores:*

$\widetilde{\mathbf{x}}$ é actualizado através da média das quantificações das categorias a que cada objecto pertence[3] e designa-se por $\mathbf{x}^+$.

(3) *Normalização*: $\mathbf{x}^+$ passa a ter norma unitária.

(4) *Actualização das quantificações das categorias:*

$\widetilde{\mathbf{y}}$ é actualizado repetindo (1)-c) com $\mathbf{x}^+$ e é designado por $\mathbf{y}^+$.

(5) *Teste de convergência:*

Seja ε infinitesimal fixado à partida, se $\sigma(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) - \sigma(\mathbf{x}^+, \mathbf{y}^+) \leq \varepsilon$ então $(\mathbf{x}^+, \mathbf{y}^+)$ é a solução. Caso contrário repetem-se os passos (2) a (4) partindo de $(\mathbf{x}^+, \mathbf{y}^+)$.

Substituíndo $\mathbf{x}$ por $\mathbf{X}$ e $\mathbf{y}$ por $\mathbf{Y}$, e o passo (3) pela ortonormalização do conjunto das p colunas de $\mathbf{X}$ obtém-se a versão de dimensão p do algoritmo anterior.

A determinação da solução para a optimização da função perda é obtida meramente por cálculo de médias. Daí a designação de *Princípio das médias recíprocas*.

[2]Usualmente recorre-se à normalização unitária associada ao pressuposto das variáveis iniciais serem previamente centradas e divididas por $\sqrt{n}$. Neste caso concreto implica a variância unitária dos *object scores*. O algoritmo da HOMALS obtém a variância unitária dos *object scores* normalizando $\mathbf{x}$ para a soma dos quadrados igual a n, pois não usa o pressuposto anterior. A primeira via tem vantagens para a apresentação teórica a segunda tem vantagens ao nível da implementção prática do algoritmo.

[3]Ou seja, usa-se a relação $\mathbf{x} \propto \mathbf{G}\mathbf{y}/m$ com constante de proporcionalidade igual a 1.

Em termos geométricos, a minimização da função perda pode ser interpretada como a construção de uma representação gráfica, no plano ou no espaço, na qual os objectos e as categorias estejam posicionadas de forma a encontrar padrões relevantes, ao mesmo tempo que se retém informação suficiente de modo a produzir uma representação aceitável da realidade m-dimensional.

Cada representação gráfica relacionará os n *object scores* com as k_j categorias de uma das m variáveis. Será portanto uma nuvem de pontos (pontos--objecto e pontos-categoria) e segmentos de recta que ligam cada *object score* à categoria a que pertence. Uma das características desejáveis, para se conseguir reter mais informação das representações produzidas, é os comprimentos dos segmentos de recta serem mínimos. Pretende-se que os objectos estejam perto das categorias em que se inserem e que as categorias estejam perto dos objectos que lhes pertencem. Assim, o objectivo é construir uma representação gráfica que minimize a soma de quadrados dos comprimentos dos segmentos de recta, ou, de forma equivalente minimize a função perda da HOMALS. Associando a esta interpretação geométrica o *Princípio das médias recíprocas*, pode-se afirmar que:

1. Se há apenas um objecto que pertence a uma determinada categoria então o ponto-objecto e o ponto-categoria coincidem.

2. A distância dum ponto-categoria à origem é inversamente proporcional à frequência marginal dessa categoria.

3. Pontos-objecto com perfis raros distam mais da origem que pontos--objecto com perfis semelhantes ao "perfil médio".

A minimização da função perda da HOMALS vai assim melhorar a visualização da nuvem de segmentos de recta. Simultaneamente deriva posições para os objectos e para as categorias que facilitam a descoberta de padrões relevantes. Assim, os objectos com o mesmo perfil de resposta recebem *object scores* idênticos. Em geral, a distância entre dois pontos-objecto da representação gráfica está relacionada com a semelhança entre os seus perfis, facilitando a descoberta de grupos homogéneos de objectos.

Exemplo 2.1.1

Para ilustrar algumas destas ideias, recorre-se ao estudo[4] sobre sentimentos relacionados com a identidade nacional, envolvendo 30894 inquiridos de 24 países. Seleccionaram-se aleatoriamente 20 inquiridos da Áustria e realizou-se uma HOMALS[5] sobre as variáveis relacionadas com o grau de proximidade do respondente em relação à sua vizinhança, à sua cidade, ao seu distrito, ao seu País e ao seu Continente. Esta variáveis são ordinais, apresentando os seguintes níveis: 1 - "Very close"; 2 - "Close"; 3 - "Not very close"; 4 - "Not close at all". Em relação à variável sobre o grau de proximidade do respondente em relação à sua vizinhança, a distribuição da amostra dos 20 inquiridos seleccionados é dada pela tabela 2.1.

Tabela 2.1: Distribuição de frequências da variável sobre o grau de proximidade do respondente em relação à sua vizinhança

Categoria	Freq. absoluta
Very close	2
Close	10
Not very close	7
Not close at all	1
Total	20

Na figura 2.1 pode observar-se a representação dos 20 inquiridos (objectos) e das 4 categorias da variável nas duas primeiras dimensões da HOMALS. Tendo sido retidas duas dimensões, cada categoria está associada a duas quantificações óptimas que coincidem com as coordenadas do ponto-categoria que representa a categoria na figura 2.1. Pode observar-se que a ordem inicial das categorias, "Very close" - "Close" - "Not very close" - "Not close at all", não foi respeitada nas quantificações óptimas. A ordem da quantificação óptima associada à primeira dimensão resultou numa inversão da ordem passando para: "Not close at all" - "Not very close" - "Close"

[4]International Social Survey Programme 1995: National Identity I (ISSP 1995 [14]), base de dados pública e disponível *online* em doi:10.4232/1.2880

[5]A HOMALS foi realizada no SPSS versão 19. Nesta versão a HOMALS tem a designação de *Multiple Correspondence Analysis*.

- "Very close". A ordem da quantificação óptima associada à segunda dimensão alterou a ordem para: "Close" - "Not close at all" - "Not very close" - "Very close". Para além da ordem dos códigos iniciais foi também alterada a distância relativa entre categorias. Na codificação inicial a distância unitária entre categorias é desprovida de qualquer sentido, nas quantificações óptimas a distância relativa entre categorias tem uma interpretação clara - é a que minimiza a função perda.

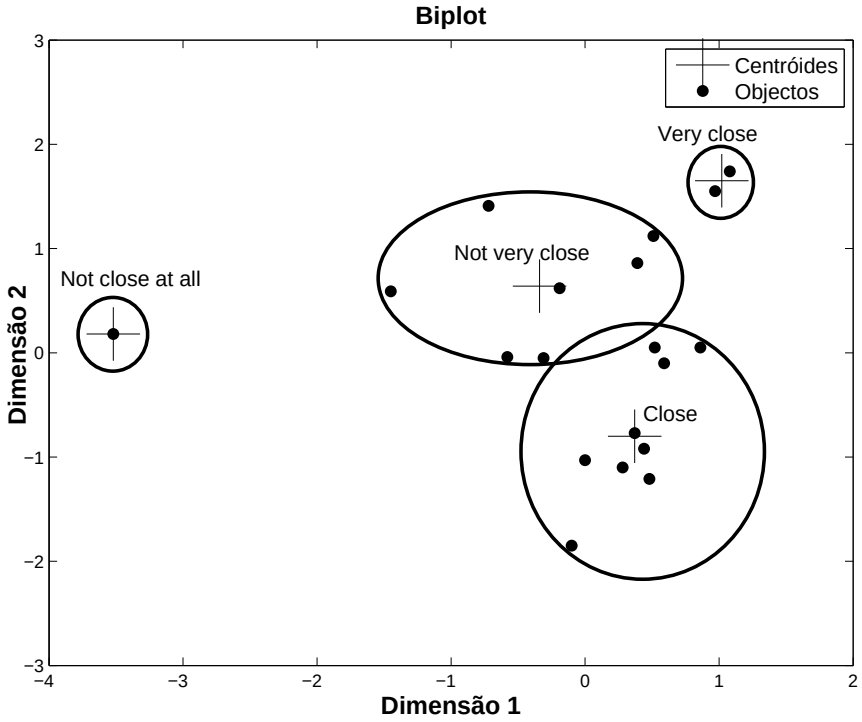

Figura 2.1: *Biplot* para a HOMALS.

No interior de cada elipse da figura 2.1 estão os pontos-objecto que correspondem aos inquiridos com resposta na mesma categoria. Como se pode observar cada ponto-categoria é o centróide de uma dessas nuvens de pontos. Na categoria 'Not close at all", com frequência unitária, o ponto-objecto e o ponto-categoria coincidem, sendo esta também a que mais dista da origem. Sublinhe-se que na figura 2.1 optou-se por marcar apenas os pontos-categoria

associados à variável relacionada com o grau de proximidade do respondente em relação à sua vizinhança. No entanto, poder-se-ia marcar em simultâneo os pontos-categoria das restantes 4 variáveis usadas na análise. O posicionamento dos pontos-objecto seria exactamente o mesmo e observar-se-ia que cada ponto-categoria seria centróide da nuvem de pontos associadas aos inquiridos que responderam nessa categoria.

Comentários finais

De todas as técnicas provenientes do *sistema Gifi*, a HOMALS é a mais potente devido à flexibilidade permitida para as quantificações. A minimização da função perda da HOMALS, estando associada à classe das transformações seccionalmente constantes, trabalha todas as variáveis tendo apenas em consideração quais os objectos que estão em cada categoria. Como se mostrou no exemplo anterior, até mesmo a ordem das categorias de uma variável pode ser alterada nas quantificações óptimas. Assim, a HOMALS está especialmente indicada para a análise de dados provenientes de variáveis de natureza nominal. Pode também ser usada para casos em que, existindo também variáveis de natureza ordinal, o investigador decida pela eventual flexibilização da ordem inicial tendo em vista a maior qualidade do ajustamento do modelo.

A CATPCA surge, histórica e tecnicamente, como uma generalização da HOMALS por forma a permitir alargar a classe de transformações admissíveis e, dessa forma, restringir a flexibilidade permitida para a quantificações. Como se mostra na secção seguinte, a definição das classes de transformações admissíveis é, na CATPCA, definida variável a variável, e deve resultar da dialéctica entre a qualidade do ajustamento e o respeito pela natureza das variáveis originais.

2.2 Da HOMALS à CATPCA

A CATPCA é uma generalização da HOMALS, apresentada na secção anterior, sendo apropriada quando se pretende reduzir a dimensionalidade com tratamento diferenciado das variáveis. A classe das transformações admissíveis é definida variável a variável e abarca várias opções, que poderão ir desde o menos restrito caso onde qualquer atribuição é válida (como acontece na HOMALS) até ao nível mais restrito de manutenção de proporcionalidade de distâncias (como acontece na ACP). Tal como se encontra neste momento implementada a CATPCA, estas classes de transformações são referenciadas como *optimal scaling levels* e o investigador tem a liberdade de escolher em

cada caso o nível de quantificação que considerar mais adequado ao problema em análise, independentemente da natureza da variável subjacente.

Optimal Scaling Levels

As variáveis tratadas como na HOMALS, ou seja, tendo apenas em consideração a manutenção dos mesmos objectos nas mesmas categorias, permitindo eventuais alterações de ordem, distâncias relativas e quantificações múltiplas, dizem-se associadas ao *optimal scaling level* **Multiple Nominal**. Um primeiro passo para a introdução de restrições sobre o comportamento das transformações admissíveis consiste em exigir que as quantificações das categorias de uma variável estejam numa recta que passa pela origem do espaço de dimensão p. Esta condição é traduzida para a matriz $\mathbf{Y}_j$ exigindo que a sua característica seja 1, ou seja, que as p colunas sejam proporcionais entre si:

$$\mathbf{Y}_j = \mathbf{y}_j \mathbf{a}_j^{'}, \text{ para } j \in \{1, 2, \ldots, m\}\,, \tag{2.2}$$

onde $\mathbf{y}_j$ é um vector coluna de dimensão k_j e $\mathbf{a}_j^{'}$ é um vector linha de dimensão p. Desta relação retém-se o vector $\mathbf{y}_j$ para a quantificação (única) das categorias da variável j. Note-se que o vector $\mathbf{y}_j$ depende da dimensão p a reter, pois como se mostra na secção sobre a optimização da função perda da CATPCA, este está obviamente dependente de $\mathbf{a}_j^{'}$ que é um vector de dimensão p. Esta é uma restrição sobre a característica da matriz $\mathbf{Y}_j$ (*rank restriction*) que passa a ser unitária e, no contexto da CATPCA, não sendo incluídas quaisquer outras restrições adicionais, corresponde ao tratamento nominal de quantificação singular (*optimal scaling level* **Nominal**). Se para além da restrição (2.2) for ainda imposto que:

1. $\mathbf{y}_j$ seja constituído por k_j elementos não decrescentes, designa-se por restrição monótona e o *optimal scaling level* associado por **Ordinal**;

2. $\mathbf{y}_j$ seja constituído por k_j elementos não decrescentes, cuja distância é proporcional à distância inicial entre a codificação das categorias, designa-se por restrição linear e o *optimal scaling level* associado por **Numerical**. Na ACP Linear todas as variáveis são tratadas desta forma.

Para além dos tratamentos anteriores, estão também disponíveis tratamentos associados a transformações *spline*. Sejam s_j e t_j, respectivamente, o grau do *spline* e o número de nós interiores escolhidos pelo utilizador para a variável j. Seja $\mathbf{S}_j$ a matriz do tipo $k_j \times (s_j + t_j)$ que contém as imagens das k_j categorias segundo cada uma das $(s_j + t_j)$ funções da base de *I-splines*. Seja $\mathbf{b}_j$ o vector com os coeficientes do *spline*. Se para além da restrição (2.2) for ainda imposto que:

1. $\mathbf{y}_j$ esteja no contradomínio de um *spline*, ou seja, $\mathbf{y}_j = \mathbf{S}_j\mathbf{b}_j$, designa-se por restrição *spline* nominal (*optimal scaling level* **Spline Nominal**);

2. $\mathbf{y}_j$ esteja no contradomínio de um *spline* monótono não decrescente, ou seja, com $\mathbf{y}_j = \mathbf{S}_j\mathbf{b}_j$ e todos os componentes de $\mathbf{b}_j$ não negativos (cf.ver referência para a secção dos splines), designa-se por restrição *spline* monótono (*optimal scaling level* **Spline Ordinal**).

Exemplo 2.2.1

Para exemplificar o possível impacto da escolha de um particular nível de quantificação para uma variável, recorre-se novamente ao estudo sobre sentimentos relacionados com a identidade nacional. Escolheram-se para análise 11 variáveis, 5 sobre grau de proximidade, 5 sobre desejo de mudança e ainda a variável "idade". As 5 variáveis sobre o grau de proximidade são ordinais, apresentando os seguintes níveis: 1 - "Very close"; 2 - "Close"; 3 - "Not very close"; 4 - "Not close at all". As 5 variáveis sobre o desejo de mudança também são ordinais, apresentando os seguintes níveis: 1 - "Very willing"; 2 - "Fairly willing"; 3 - "Not willing and not unwilling"; 4 - "Fairly unwilling"; 5 - "Very unwilling". A variável "idade" foi registada tendo em conta a idade em anos dos inquiridos no momento do inquérito.

Seleccionaram-se aleatoriamente 20 inquiridos da Áustria. Pretende-se analisar o efeito de diferentes opções para o *optimal scaling level* da variável "idade", mantendo o tratamento *Ordinal* para as restantes variáveis. Os cinco gráficos da figura 2.2 apresentam as transformações óptimas para a

variável "idade", revelando que diferentes opções para o *optimal scaling level* podem conduzir a transformações bastante distintas. O gráfico do tratamento *Nominal* apresenta um comportamento pouco suave, traduzindo a escassez de restrições deste *optimal scaling level*. Nos gráficos dos tratamentos *Ordinal* e *Spline Ordinal*, linear com dois nós interiores, pode observar-se o efeito da restrição da manutenção da ordem inicial traduzido pelas funções crescentes que estes apresentam. No gráfico associado ao *optimal scaling level Numerical* pode observar-se uma transformação linear, mantendo a ordem inicial e transformando de forma proporcional as distâncias relativas iniciais. O gráfico associado ao nível *Spline Nominal*, linear com dois nós interiores, apresenta uma função crescente seguida de uma função decrescente.

A decisão pelo *optimal scaling level* mais adequado deve ter em conta não apenas a natureza da variável, como ainda a flexibilidade que o investigador está disposto a admitir ao transformá-la e a qualidade do ajustamento da variável ao modelo. Para cada caso, a medida de ajustamento ao modelo, da variável "idade" transformada é dada pela soma dos quadrados das correlações entre a variável transformada e as componentes principais não-lineares retidas (foram retidas duas componentes). Esta medida é usualmente designada por *VAF por variável transformada*, ou ainda *comunalidade* (Linting e Kooij [25]), indicando a proporção da variância da variável transformada que é explicada pelas componentes principais não-lineares retidas. A VAF da variável "idade" transformada atinge o valor máximo de 0.924 quando esta é tratada como *Nominal* e o valor mínimo de 0.156 quando tratada como *Numerical*. Ambos os tratamentos parecem desadequados, pelo excesso e defeito de flexibilidade, respectivamente. O tratamento *Ordinal* está associado a uma VAF de 0.406 e o tratamento *Spline Ordinal* a uma VAF de 0.262. Apesar de ambos manterem a ordenação inicial, é claro, neste exemplo, a perda inerente à redução da flexibilidade inerente ao *spline*. Quando tratada ao nível *Spline Nominal* a VAF da variável "idade" transformada tem o valor de 0.576. Este último tratamento, que apresenta um acréscimo de cerca de 40% da variância explicada da variável transformada em relação ao tratamento *Numerical*, parece ser o mais adequado pois permite liberdade suficiente para descrever a eventual relação não-linear entre a

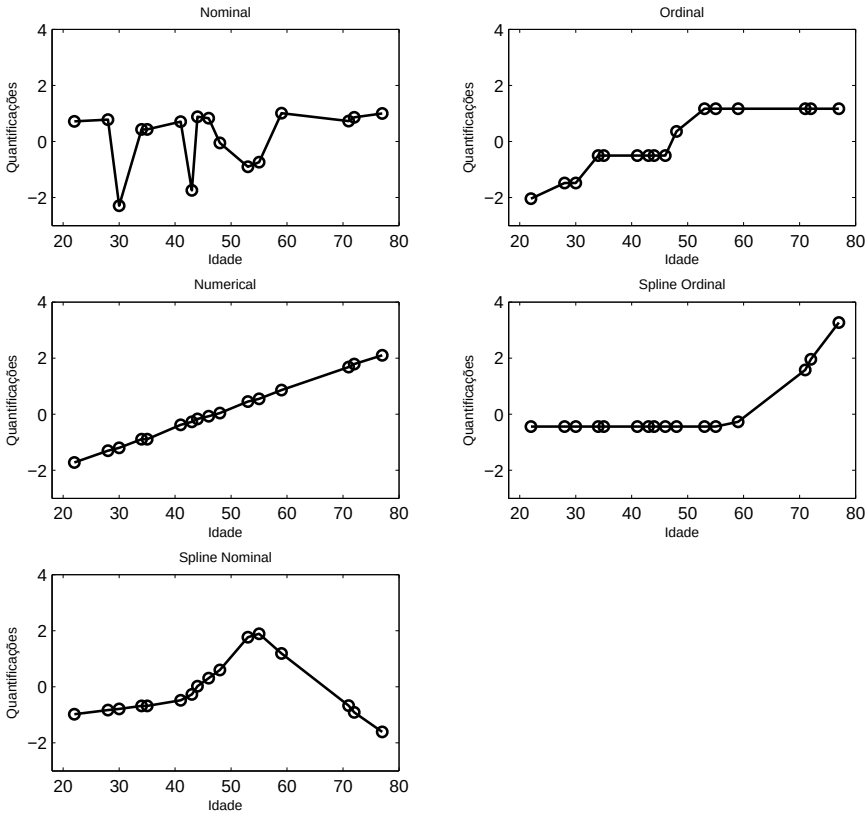

Figura 2.2: Gráficos das transformações da variável "idade" para diferentes *optimal scaling levels*.

variável "idade" e as restantes 10 variáveis. Em simultâneo, a flexibilidade do tratamento *Spline Nominal* não é excessiva, permitindo ao investigador interpretar facilmente o comportamento desta variável.

Exemplo 2.2.2

Para exemplificar que a *quantificação óptima* é um conceito relativo, considere-se a variável de caracterização "estado civil" que foi usada no estudo referido no exemplo anterior com cinco categorias e com a seguinte codificação: 1 - casado/vivendo como casado, 2 - viúvo, 3 - divorciado, 4 - separado e 5 - solteiro. Determinar a *quantificação óptima* para a variável

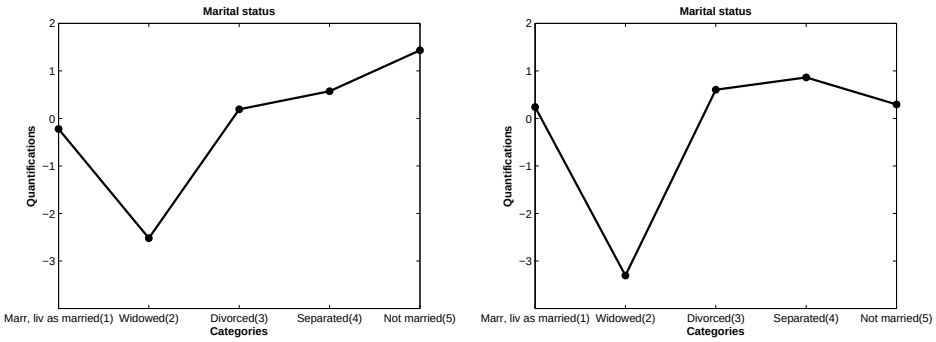

Figura 2.3: Quantificações óptimas ao nível nominal para a variável "estado civil", no gráfico da esquerda no seio de 11 variáveis e no da direita no seio de 6 variáveis, usando a mesma amostra.

nominal "estado civil" é admitir que não há nenhuma razão devidamente fundamentada para escolher determinada codificação e partindo dessa flexibilidade permitir uma optimização da escolha tendo em vista a redução da dimensão, ou de forma equivalente, a maximização da homogeneidade entre as variáveis em análise. As *quantificações óptimas* obtidas usando a CATPCA podem observar-se na figura 2.3 em dois cenários obtidos com a mesma amostra: (a) no seio das 11 variáveis já definidas (5 sobre grau de proximidade, 5 sobre desejo de mudança e "estado civil") e (b) no seio de apenas 6 dessas variáveis (5 relativas ao grau de proximidade e "estado civil"). No eixo das abcissas encontram-se as categorias da variável pela ordem associada à codificação inicial, no eixo das ordenadas a *quantificação óptima* para cada categoria. Como se pode observar na figura 2.3, ambas as transformações são não-lineares, o que se traduz numa alteração de forma não proporcional das distâncias iniciais entre categorias. As transformações não são monótonas, o que se traduz numa alteração da ordem inicial das categorias passando agora a categoria "viúvo" a estar associada à quantificação de valor mínimo. Este cenário é o mais flexível, no sentido em que não existem restrições devido à natureza da variável.

Como se pode observar na figura 2.3, a optimização relativa à mesma amostra produziu resultados diferentes consoante se incluíam na análise 11 ou

apenas 6 variáveis. A *quantificação óptima* é um conceito relativo, com outra amostra ou com outra selecção de variáveis seria obtida outra *quantificação*.

2.2.1 Função perda da CATPCA e sua optimização

A função perda da CATPCA é dada por:

$$\sigma\left(\mathbf{X};\underline{\mathbf{Y}}\right)=n^{-1}\sum_{j=1}^{m}c_j^{-1}tr\left[\left(\mathbf{X}-\mathbf{G}_j\underline{\mathbf{Y}_j}\right)'\left(\mathbf{X}-\mathbf{G}_j\underline{\mathbf{Y}_j}\right)\right] \tag{2.3}$$

onde o sublinhado indica a eventual existência das restrições enunciadas e $c_j = p$ se a variável j é tratada ao nível *Multiple Nominal* e $c_j = 1$, caso contrário[6].

A estrutura da função perda da CATPCA é igual à da HOMALS, a menos de uma constante parcial e das restrições. O algoritmo de minimização da função anterior também tem uma estrutura semelhante ao da HOMALS, mas vai ter um sub-algoritmo adicional em que se alterna entre minimizar $\mathbf{y}_j$ (para as restrições pretendidas) para $\mathbf{a}_j'$ fixo e minimizar $\mathbf{a}_j'$ para $\mathbf{y}_j$ fixo.

Na secção anterior mostrou-se que minimizar a função perda da HOMALS é equivalente, em termos geométricos, a determinar a localização dos *object scores* (pontos-objecto, representantes dos objectos observados) e das quantificações das categorias (pontos-categoria, representantes das categorias iniciais) de cada variável, por forma a minimizar a soma dos quadrados dos comprimentos dos segmentos de recta que unem cada ponto-objecto aos pontos-categoria a que esse indivíduo pertence. Assim, cada variável a tratar como *Multiple Nominal* tem apenas uma componente de perda, a inerente à substituição da variável pelas componentes principais.

A principal ideia para implementar as restrições na minimização da função perda da CATPCA, é a partição, para as variáveis a restringir, da componente de perda em duas ou mais componentes. Ou seja, quantificar a perda

[6] A função perda da CATPCA tem ainda uma matriz destinada a controlar o tratamento a dar aos valores omissos e outra que permite ao utilizador atribuir pesos diferentes aos indivíduos. No enquadramento desta tese, optou-se por considerar que não existem valores omissos e que todos os indivíduos têm peso unitário, ou seja, ambas as matrizes anteriores são a matriz identidade.

adicional inerente ao desvio das quantificações óptimas das categorias para a localização exigida por cada restrição.

Para variáveis a tratar ao nível *Nominal* a perda terá duas componentes: a habitual e outra correspondente ao desvio dos centróides face a uma recta que passa pela origem. Quanto às variáveis a tratar ao nível *Ordinal* a perda terá três componentes: as duas referidas e a inerente à manutenção da ordem inicial. Já quanto àquelas a tratar ao nível *Numerical* a perda terá quatro componentes: as três referidas e a perda inerente à deslocação dos pontos-categoria por forma a estarem igualmente espaçados. Para variáveis a tratar aos níveis *Spline Nominal* ou *Spline Ordinal* a perda terá três componentes: a habitual, a da passagem para a recta e a perda inerente à disposição dos pontos-categoria na recta conforme ditado pelo contradomínio dos *splines*.

Para minimizar a função perda da CATPCA, a parte relevante de (2.3) separa-se em duas componentes de perda:

$$\sum_{j=1}^{m} tr\left[(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j)'(\mathbf{X}-\mathbf{G}_j\mathbf{Y}_j)\right]+\sum_{j=1}^{m} tr\left[\left(\underline{\mathbf{Y}_j}-\mathbf{Y}_j\right)'\mathbf{D}_j\left(\underline{\mathbf{Y}_j}-\mathbf{Y}_j\right)\right]. \tag{2.4}$$

A primeira componente corresponde à perda habitual da HOMALS e designa-se por *Multiple Loss*. A segunda componente, designada por *Single Loss*, pode ter dois significados, se:

1. $\underline{\mathbf{Y}_j}=\mathbf{Y}_j$ não há contribuição para a perda da parcela *Single Loss*, ou seja, as quantificações das categorias continuaram no centróide, o que acontece para as variáveis com o tratamento *Multiple Nominal*;

2. $\underline{\mathbf{Y}_j}=\mathbf{y}_j\mathbf{a}_j'$ corresponde ao somatório dos quadrados dos desvios dos centróides para a recta ponderados pela respectiva frequência marginal de cada categoria.

Assim, minimizar a função perda da CATPCA é equivalente a minimizar cada uma das componentes. No contexto da CATPCA, o processo de *Mínimos Quadrados Alternados* consubstancia-se no *Princípio das médias recíprocas* com ciclos adicionais para implementar os diferentes *optimal scaling levels.*

O algoritmo completo da CATPCA para determinar $\mathbf{X}$ e $\underline{\mathbf{Y}}$ pode-se resumir nos passos seguintes:

(1) *Inicialização (sai* $\mathbf{X}^+$ *e* $\mathbf{a}_j'^+$*)*:

(a) $\mathbf{X}$ é construído de forma aleatória;

(b) $\mathbf{X}$ é normalizado, ficando centrado e satisfazendo $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, sendo designado por $\mathbf{X}^+$;

(c) Seja $\mathbf{x}_s^+$ a s-ésima coluna de $\mathbf{X}^+$ e $\mathbf{a}_j'^+ = \begin{bmatrix} a_{j1}^+ & \dots & a_{js}^+ & \dots & a_{jp}^+ \end{bmatrix}$, o vector a optimizar para usar na relação (2.2). O vector $\mathbf{a}_j'^+$ é dado por:

$$a_{js}^+ = corr\left(\mathbf{x}_s^+, \mathbf{h}_j\right) \tag{2.5}$$

onde $\mathbf{h}_j$ é o vector/variável inicial estandardizado.

(2) *Actualização das quantificações das categorias (entra* $\mathbf{H}$, $\mathbf{X}^+$ *e* $\mathbf{a}_j'^+$ *e sai* $\underline{\mathbf{Y}_j}^+$*)*:

(a) actualização sem restrições, $\mathbf{Y}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}^+$, para todo o j;

(b) actualização com restrições [7]; *se a variável* j tem *optimal scaling level*:

i. *Multiple Nominal*, então $\underline{\mathbf{Y}_j}^+ = \mathbf{Y}_j$;

ii. *Single*, então[8] $\underline{\mathbf{Y}_j}^+ = \mathbf{y}_j^+\mathbf{a}_j'^+$.

(3) *Actualização dos object scores (entra* $\underline{\mathbf{Y}_j}^+$ *e sai* $\mathbf{X}^+$*):*

(a) actualização: $\mathbf{X} = \sum_{j=1}^{m} \mathbf{G}_j\underline{\mathbf{Y}_j}^+$;

(b) igual a (1) - b).

[7] ciclo pelas variáveis $j = 1, ..., m$.

[8] Para não sobrecarregar a apresentação do algoritmo completo, apresentam-se os ciclos referentes à computação de $\mathbf{y}_j^+$ e $\mathbf{a}_j'^+$ para cada *optimal scaling level* no fim da apresentação global do algoritmo. Esse ciclo inicia-se com $\mathbf{a}_j'^+$ fixo dado em (1)-(c).

(4) *Teste de convergência.*

Seja ε infinitesimal fixado à partida e $(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_1$ e $(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_2$ o resultado da aplicação consecutiva dos passos (2) e (3).
Se $\sigma(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_1 - \sigma(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_2 \leq \varepsilon$ então $(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_2$ é a solução, senão repetem-se os passos (2) e (3) partindo de $(\mathbf{X}^+, \underline{\mathbf{Y}_j}^+) = (\mathbf{X}^+, \underline{\mathbf{Y}_j}^+)_2$.

A ideia do sub-algoritmo da CATPCA, para determinar $\mathbf{y}_j^+$ e $\mathbf{a}_j^{'+}$ a usar no passo (2)-b)-ii), é usar $\mathbf{Y}_j$ proveniente de (2)-a) e determinar $\mathbf{y}_j^+$ (com as restrições pretendidas) e $\mathbf{a}_j^{'+}$ que minimizem:

$$tr\left[\left(\mathbf{y}_j^+\mathbf{a}_j'^+ - \mathbf{Y}_j\right)'\mathbf{D}_j\left(\mathbf{y}_j^+\mathbf{a}_j'^+ - \mathbf{Y}_j\right)\right]. \tag{2.6}$$

Este pode-se resumir nos passos seguintes[9]:

(1) *Inicialização (entra* $\mathbf{a}_j'^+$ *e* $\mathbf{Y}_j$; sai $\widetilde{\mathbf{y}}_j$):

(a) $\mathbf{a}_j'^+$ inicializa-se usando (1)-c);

(b) $\mathbf{y}_j^+$ inicializa-se, designando-se por $\widetilde{\mathbf{y}}_j$, sendo determinado pela relação $\widetilde{\mathbf{y}}_j = \mathbf{Y}_j\mathbf{a}_j^+$

(2) *Actualização*[10] *da quantificação da variável* j *(entra* $\widetilde{\mathbf{y}}_j$ *e sai* $\mathbf{y}_j^+$). Se a variável j tem *optimal scaling level*:

(a) *Nominal*, então $\mathbf{y}_j^+ = \widetilde{\mathbf{y}}_j$;

(b) *Ordinal:* atribui-se a $\mathbf{y}_j^+$ o resultado proveniente do processo de regressão monótona ponderada[11] sobre $\widetilde{\mathbf{y}}_j$, com pesos dados pela frequência marginal de cada categoria;

[9] ciclo pelas variáveis $j = 1, ..., m$.

[10] Trata-se de encontrar $\mathbf{y}_j^+$ que minimize a soma dos quadrados dos desvios de $\widetilde{\mathbf{y}_j}$ para a posição desejada nas várias restrições.

[11] O estudo dos pormenores deste processo estão fora do âmbito desta tese. Ver literatura de apoio ao SPSS para referências adicionais.

(c) *Numerical*: atribui-se a $\mathbf{y}_j^+$ o resultado proveniente do processo de regressão linear ponderada sobre $\widetilde{\mathbf{y}}_j$, com pesos dados pela frequência marginal de cada categoria[12];

(d) *Spline Nominal e Spline Ordinal*: faz-se $\mathbf{y}_j^+ = \mathbf{S}_j\mathbf{b}_j$ resultantes de determinar a imagem de $\widetilde{\mathbf{y}}_j$ segundo as $(s_j + t_j)$ funções da base de *I-splines* e armazenar essa informação na matriz $\mathbf{S}_j$. Considerar o problema de determinar os (s_j+t_j) coeficientes a aplicar à base de *I-splines* por forma a minimizar a soma dos quadrados dos desvios entre os elementos no contradomínio do *spline* assim obtido e $\widetilde{\mathbf{y}}_j$. Armazenar em $\mathbf{b}_j$ os ditos coeficientes obtidos por um processo de regressão linear ponderada com pesos dados pela frequência marginal de cada categoria.

(3) *Normalização de* $\mathbf{y}_j^+$ *por forma a ter variância unitária.*

(4) *Actualização do vector* $\mathbf{a}_j'^+$ *(entra* $\mathbf{y}_j^+$ *e sai* $\mathbf{a}_j'^+$*)*:

$$\mathbf{a}_j'^+ = n^{-1}\mathbf{Y}_j'\mathbf{D}_j\mathbf{y}_j^+.$$

(5) *Introduzir os vectores* $\mathbf{y}_j^+$ *e* $\mathbf{a}_j'^+$ *resultantes dos passos anteriores no passo (2)- b)- ii) do algoritmo completo.*

2.2.2 Variáveis contínuas na CATPCA

Note-se que no algoritmo da CATPCA, a minimização de $\mathbf{y}_j^+$ nos *optimal scaling levels Nominal, Ordinal* e $\mathbf{Y}_j$ no *Multiple Nominal* assenta nas categorias. Já a minimização de $\mathbf{y}_j^+$ nos *optimal scaling levels Spline Nominal* e *Spline Ordinal* assenta nos coeficientes dos *I-Splines,* que regra geral não passarão de quatro. Esta estratégia permite incorporar no algoritmo da CATPCA, variáveis contínuas, nas quais o número de categorias é sensivelmente o mesmo que o número de objectos.

[12]Note-se que apesar de formalmente ser esta a apresentação do algoritmo, a referida regressão não é necessária, pois a quantificação óptima das variáveis tratadas como numéricas é simplesmente obtida pela estandardização dessas variáveis.

Um processo de discretização é usado para transformar as variáveis iniciais em variáveis categoriais, podendo também ser usado para recodificar variáveis categoriais.

O programa CATPCA possui um subcomando, denominado *Discretization,* para esse efeito. Este considera que uma variável é categorial se os valores observados forem números inteiros positivos, sendo disponibilizados quatro métodos para implementar o processo, a saber: *unspecified*, *grouping*, *ranking* e *multiplying*.

Unspecified

Esta é a opção por omissão.

As variáveis são convertidas em variáveis categoriais pelo método *grouping* com sete categorias segundo uma distribuição aproximadamente normal (ver *grouping*).

As variáveis categoriais, com todos os valores observados inteiros positivos, permanecem inalteradas. Sublinhe-se que se existirem valores inteiros menores que 1 estes são considerados pelo CATPCA, através do subcomando *missing* na opção por omissão, como valores omissos sendo excluídos da análise. Uma forma de evitar este efeito é recodificar as variáveis categoriais usando o método *grouping* ou *ranking*.

Grouping

Este método de discretização tem duas modalidades:

1. *Number of categories*: o utilizador determina o número de categorias que pretende e escolhe entre distribuição normal ou uniforme. Por omissão, quando se opta pelo método *grouping* é esta a modalidade de discretização, sendo o número de categorias igual a sete e a distribuição normal.

2. *Equal intervals:* o utilizador com esta modalidade está a optar por intervalos/categorias com a mesma amplitude, sendo necessário escolher a amplitude dos intervalos (nenhum valor é sugerido). O número de categorias resultante depende obviamente da amplitude especificada.

Ranking

Este método atribui o valor um ao menor valor observado na variável e assim sucessivamente. O número de categorias é igual ao número de valores distintos na variável.

Multiplying

Este método constrói as categorias da seguinte forma:

1. estandardiza a variável;
2. multiplica os valores estandardizados por dez;
3. arredonda os valores resultantes de 2. para o valor inteiro mais próximo;
4. adiciona uma constante aos valores resultantes de 3. por forma a que o valor mínimo na variável seja um.

O arredondamento pode resultar no agrupamento na mesma categoria de valores inicialmente distintos, desde que estes estejam suficientemente perto. Assim, o número de categorias é igual ou inferior ao número de valores distintos na variável. Sublinhe-se que o utilizador pode efectuar ele próprio a discretização, fora da CATPCA, aproveitando ao máximo o conhecimento que tem da variável.

Exemplo 2.2.3

No estudo da reacção de seres humanos a determinada vacina é considerada a variável idade, em anos, tendo os dados sido recolhidos em diferentes fontes. Para indivíduos com idade superior a 12 anos, os investigadores registam apenas o número inteiro de anos mais próximo; com idades entre 6 e 12 anos consideram ainda meios anos; com idades entre 3 e 6 anos registam com precisão trimestral; com idades entre 1 e 3 anos registam com precisão mensal; para indivíduos com idade inferior a um ano registam com rigor semanal.

Com as observações da variável idade segundo os preceitos anteriores, esta será considerada, pelo CATPCA, uma variável não categorial e está por isso sujeita a discretização.

Variável original	*Multiplying*	*Ranking*
0.019230769	1	1
0.057692308	2	2
0.076923077	2	3
0.173076923	2	4
0.211538462	2	5
0.461538462	2	6
0.480769231	3	7
0.5	3	8
0.576923077	3	9
0.596153846	3	10
0.634615385	3	11
0.846153846	3	12
0.865384615	3	13
1.083333333	4	14
1.166666667	4	15
1.25	4	16
1.333333333	5	17
1.416666667	5	18
1.5	5	19
1.583333333	5	20
2.083333333	6	21
2.166666667	7	22
3.25	9	23
3.5	10	24
3.75	10	25
7	18	26
7.5	19	27
8	20	28
8.5	22	29
13	32	30
14	35	31
15	37	32

Tabela 2.2: Discretização da variável pelos métodos *Multiplying* e *Ranking*.

O investigador decidiu, com base no conhecimento que tem sobre a influência da variável idade no fenómeno em estudo, que esta pode eventual-

mente ser transformada por motivos de optimização. No entanto, pretende uma transformação suave, facilmente interpretável e que a distância entre valores iniciais seja de alguma forma proporcional aos valores iniciais. Se a opção for usar a CATPCA, o método de discretização para esta variável deve ser escolhido de forma a minimizar a perda de informação relativamente aos valores originais.

Conforme se pode observar na tabela 2.2 o método *multiplying*, apesar de respeitar a informação sobre a distância e a ordem entre os valores observados, não distingue alguns dos valores medidos com precisão superior à semi-anual. O método *ranking* apenas tem em conta a ordem dos valores observados, não distingue se entre indivíduos consecutivos a diferença de idades é uma semana ou um mês.

2.3 Conclusão

A definição das classes de transformações admitidas para cada uma das variáveis a incluir na função perda deve resultar da dialéctica entre a qualidade do ajustamento e o respeito pela natureza das variáveis originais.

Se a variável inicial for de natureza qualitativa nominal, ter-se-á apenas em consideração quais os objectos que estão em cada categoria e nada é assumido sobre a distância ou ordem entre as categorias. Assim, a transformação usada na HOMALS é a ideal para essa variável, pois obtém-se o melhor ajustamento e o máximo respeito pela informação inicial das variáveis.

Se a variável inicial for de natureza qualitativa ordinal, para além de se considerar quais os objectos que estão em cada categoria, assumir-se-á ainda que a ordem entre as categorias é relevante. Assim, a transformação usada na HOMALS carece de aceitação prévia pelos investigadores, pois esta destrói a informação sobre ordem aquando da determinação da quantificação óptima. A transformação linear é admissível para essa variável, pois mantém a informação sobre ordem, mas impõe a restrição, desnecessária, de apenas alterar a distância entre as categorias de forma proporcional. Assim, a transformação linear não é a ideal para essa variável, pois limita em demasia a qualidade do ajustamento.

Se a variável inicial for de natureza quantitativa, interessa considerar quais os objectos que estão em cada categoria e a ordem e a distância entre elas. A transformação linear preserva ao máximo a distância entre categorias, pois apenas admite que a mudança seja feita de forma proporcional para todas.

O recurso a *splines* está particularmente indicado para variáveis contínuas ou categoriais com várias categorias. Sendo uma forma de desrespeito pela natureza das variáveis iniciais, directamente proporcional ao número de nós interiores e com relação exponencial em relação à sua ordem, fornece no entanto um meio termo, que carecendo de validação dos investigadores permite aumentar a qualidade do ajustamento.

Na presença de variáveis contínuas, a CATPCA surge como a única solução quando existem também variáveis categoriais para análise. No en-

tanto, o investigador deve ter presente que a CATPCA opera sobre a matriz resultante do processo de discretização e que este pode representar uma perda da informação contida na variável original.

No caso particular em que todas as variáveis são contínuas não faz sentido usar a CATPCA devido ao processo de discretização. No próximo capítulo propõe-se um algoritmo inspirado no sistema *Gifi* e na CATPCA que pretende ser uma solução para cenários de variáveis contínuas.

Capítulo 3

qlPCA - *quasi-linear* Principal Components Analysis

Uma nova abordagem para generalizar a Análise em Componentes Principais para estruturas não-lineares foi recentemente proposta por Lavado e Calapez [23]: *quasi-linear PCA* (qlPCA). Esta inclui transformações *spline* das variáveis originais, tendo o adjectivo *quasi* sido escolhido para sublinhar as vantagens que advêm do uso de *splines* lineares (transformações seccionalmente lineares). A minimização de uma função perda através de um processo de mínimos quadrados alternados, conforme definido por Gifi [12], permite obter as transformações *spline* óptimas e as componentes principais não-lineares. As transformações óptimas são explicitamente conhecidas após a convergência. Na primeira secção deste capítulo é apresentado o algoritmo da qlPCA e suas principais propriedades. Parte desta secção está publicada no *IAENG International Journal of Applied Mathematics* em Lavado e Calapez [23]. Na segunda secção deste capítulo é apresentada a implementação do algoritmo da qlPCA em MatLab. Na última secção é apresentado um exemplo com base em dados simulados.

A Análise em Componentes Principais (ACP), sendo provavelmente a técnica descritiva mais comum para reduzir a dimensão de estruturas line-

ares, tem sido alvo de várias tentativas de generalização para estruturas não-lineares. O conceito fundamental da ACP é projectar os dados originais, que incluem ruído e variáveis redundantes, num espaço latente de dimensão inferior, com o objectivo de revelar a verdadeira dimensionalidade dos dados.

Todos os métodos descritivos para redução da dimensão partilham a mesma premissa básica e objectivos gerais. Os dados originais podem ser considerados como uma colecção de n pontos num espaço superior m-dimensional, correspondendo os pontos à amostra de indivíduos e as dimensões às variáveis medidas. Pretende-se determinar uma aproximação num espaço inferior p--dimensional, tal que os pontos sejam projectados de forma a reter o máximo de informação do espaço original. Ao reduzir a dimensão, torna-se posteriormente possível interpretar e conduzir outras análises sobre um número reduzido de componentes ao invés dum número elevado de variáveis. Diferentes interpretações do objectivo "tal que os pontos sejam projectados de forma a reter o máximo de informação do espaço original" conduziram a diferentes técnicas multivariadas para redução da dimensão.

Uma técnica pode ser considerada *linear* quando o conjunto de coordenadas de dimensão superior é substituído por outro de dimensão reduzida, de tal forma que existe uma transformação *linear* entre as coordenadas. Todas as generalizações da ACP para estruturas não-lineares, as geralmente designadas por ACP não-linear (ACPNL), partilham a mesma premissa e objectivos gerais enunciados, mas abordam o problema da não-linearidade relaxando as restrições lineares entre espaços.

Uma primeira tentativa para generalizar a ACP foi apresentada por Gnanadesikan e Wilk nos anos 1960. A ideia era estender o espaço m-dimensional acrescentando variáveis definidas por funções não-lineares das variáveis originais (termos quadráticos e de ordem superior) e posteriormente realizar uma ACP sobre este espaço (Krzanowski e Marriott [18]). A chave para esta abordagem residia na escolha da dimensionalidade apropriada do espaço estendido, bem como nas relações não-lineares entre as variáveis originais necessárias para descrever o sistema. Esta limitação foi removida nos anos 1990 por Schoolkopf et al. [32] usando uma função do espaço m-dimensional original num espaço de dimensão superior arbitrária (conhecido como *fea-*

ture space na comunidade de aprendizagem automática) que, de forma "automática", leva a cabo a correspondência não-linear. Esta correspondência é realizada de forma implícita usando funções *kernel* não necessitando por isso de ser especificadas. Esta abordagem, conhecida como ACP *kernel*, aplica posteriormente ACP sobre o *feature space*. Recentemente, Kruger et al. [17] apresentaram uma revisão do trabalho existente sobre ACPNL referindo que este se pode dividir em:

- *principal curves and manifolds*, abordagem apresentada em 1984 por Hastie [13];
- redes neuronais, abordagem apresentada em 1991 por Kramer [16];
- *kernel Principal Components Analysis* (KPCA), abordagem apresentada em 1998 por Schoolkopf et al. [32];
- técnicas que usam a combinação destas abordagens.

A lista das técnicas de ACPNL mais conhecidas entre investigadores que lidam com variáveis contínuas, não inclui a CATPCA (CATegorical Principal Components Analysis, técnica abordada no capítulo anterior), técnica especialmente apropriada para realizar uma ACP sobre variáveis nominais e ordinais (Meulman et al. [30]). Neste capítulo apresenta-se a qlPCA, abordagem que emergiu do *sistema Gifi*, tendo sido especificamente concebida para variáveis contínuas. Considera-se que a qlPCA representa um contributo relevante, para ambientes de variáveis contínuas, sendo talvez um primeiro passo para colocar uma técnica derivada do *sistema Gifi*, numa tal lista.

3.1 Fundamentação teórica

Nesta secção é apresentada a fundamentação da qlPCA. De seguida é introduzido o conceito de matriz pseudo-indicatriz, essencial para a introdução de *splines* na qlPCA, sendo que na inicialização da qlPCA cada variável é associada a uma matriz pseudo-indicatriz. Segue-se a apresentação do algoritmo da qlPCA e suas principais propriedades: sumário do modelo; escolha

do número de componentes a reter; projecção de novas observações no espaço das componentes principais não-lineares; reconstrução dos valores originais através das projecções. Termina-se com uma proposta de extensão do conceito de *loadings*, que se propõe designar *piecewise loadings*, definidos como correlações "segmentadas" entre as componentes principais não-lineares e as variáveis originais.

3.1.1 Matriz pseudo-indicatriz

Na tabela 3.1 é apresentada a discretização $\mathbf{hd}_j$, segundo a opção *ranking* (ver capítulo anterior), duma variável $\mathbf{h}_j$, seguida da matriz indicatriz $\mathbf{G}_j$ que está associada a essa discretização.

Variável original $\mathbf{h}_j$	Variável discretizada $\mathbf{hd}_j$
1.75	4
1.76	5
1.8	6
1.73	2
1.75	4
1.74	3
1.81	7
1.72	1

Tabela 3.1: Discretização da variável.

$$\mathbf{G}_j = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Cada coluna da matriz indicatriz pode ser interpretada como o conjunto das imagens resultantes da aplicação a cada indivíduo de uma das funções da base do espaço vectorial de *splines* de grau zero, com tantos pontos de junção quantos os "espaços" entre categorias. Assim, no exemplo anterior trata-se do espaço vectorial de grau zero com seis pontos de junção, também designados nós interiores. Conforme referido no capítulo 1, para gerar *splines* de grau zero (ordem um), com seis pontos de junção, é necessária uma base de dimensão sete. Sete são precisamente o número de colunas da matriz indicatriz em causa, sendo cada coluna obtida através duma base de *B-splines*. Sublinhe-se que as bases mais usadas, *B-spline* e *I-spline*, estão descritas para qualquer ordem e número de pontos de junção.

O princípio básico da matriz pseudo-indicatriz é o mesmo que o das matrizes indicatriz: cada coluna pode ser interpretada como o conjunto das imagens resultantes da aplicação a cada indivíduo de uma das funções da base de determinado espaço vectorial de *splines*. A generalização reside no grau pretendido, no número de pontos de junção e na sua localização.

Para a variável considerada na tabela 3.1, a matriz pseudo-indicatriz associada a *B-splines* de grau 1 com dois nós interiores (localizados no percentil 33 e 66) é:

$$\mathbf{G}_j^{\triangle} = \begin{bmatrix} 0 & 0.48 & 0.52 & 0 \\ 0 & 0 & 0.96 & 0.04 \\ 0 & 0 & 0.26 & 0.74 \\ 0.43 & 0.57 & 0 & 0 \\ 0 & 0.48 & 0.52 & 0 \\ 0.05 & 0.95 & 0 & 0 \\ 0 & 0 & 0.09 & 0.91 \\ 0.81 & 0.19 & 0 & 0 \end{bmatrix}$$

Note-se que a soma em linha continua a ser sempre unitária. Para a mesma variável, a matriz pseudo-indicatriz associada a *I-splines* de grau dois com dois nós interiores (localizados no percentil 33 e 66) é:

$$\mathbf{G}_j^{\triangle} = \begin{bmatrix} 0.31 & 0.02 & 0 & 0 \\ 0.75 & 0.15 & 0 & 0 \\ 0.97 & 0.42 & 0 & 0 \\ 1 & 0.78 & 0.02 & 0 \\ 1 & 0.78 & 0.02 & 0 \\ 1 & 0.98 & 0.17 & 0 \\ 1 & 1 & 0.97 & 0.6 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Como se pode observar para a matriz anterior a soma em linha já não é unitária. Note-se que a matriz pseudo-indicatriz da variável j é do tipo $n \times w$, onde n é o número de indivíduos e $w = v + r$ é a dimensão do subespaço dos *splines* de grau v com r nós interiores gerado por *I-splines*, sendo as colunas os vectores das imagens da variável j em cada uma das funções da base de *I-splines*.

3.1.2 Função objectivo

No primeiro capítulo foi definida a função perda comum (1.9) como $\sigma : \mathrm{M}_{n\times p} \times \mathrm{M}_{pm\times n} \to \mathbb{R}$ tal que

$$\sigma(\mathbf{X}, \mathbf{M}) = n^{-1} \sum_j tr\left[(\mathbf{X} - \mathbf{M}_j)' (\mathbf{X} - \mathbf{M}_j) \right],$$

sujeita à restrição $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, onde

- $\mathbf{M}_j = \begin{bmatrix} \mathbf{m}_{j1} & \dots & \mathbf{m}_{jp} \end{bmatrix}$ é a matriz $n \times p$ contendo as p (diferentes) imagens do mesmo vector $\mathbf{h}_j$;
- $\mathbf{m}_{jt}$ é o vector das imagens de $\mathbf{h}_j$, obtidas através da transformação ϕ_{jt} da variável j para a dimensão t, $t = 1, \dots, p$;
- $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \dots & \mathbf{M}_m \end{bmatrix}'$ é uma matriz $pm \times n$.

O algoritmo da qlPCA usa $\mathbf{M}_j$ com característica unitária para todas as variáveis, ou seja, é usado exclusivamente o tratamento *Single* (conforme definição no capítulo anterior). A garantia da característica unitária de $\mathbf{M}_j$ é dada por construção. Após a convergência, as variáveis estão associadas a uma única transformação e portanto a um único vector de imagens $\mathbf{f}_j$. A matriz $\mathbf{M}_j$ é definida com colunas colineares ao vector $\mathbf{f}_j$, sendo por isso de característica unitária:

$$\mathbf{M}_j = \mathbf{f}_j \mathbf{a}_j^{'} = [\mathbf{f}_j a_{1j} \dots \mathbf{f}_j a_{pj}].$$

Sendo todas as variáveis tratadas como *Single*, a função perda comum pode ser escrita de forma mais simples como[1]

$$\sigma(\mathbf{X}, \mathbf{F}) = n^{-1} \sum_j (\mathbf{X}\mathbf{a}_j - \mathbf{f}_j)^{'} (\mathbf{X}\mathbf{a}_j - \mathbf{f}_j), \tag{3.1}$$

sendo esta a função objectivo da qlPCA. Assim, a perda a ser minimizada é a perda de informação inerente à representação das m variáveis transformadas por um número reduzido de componentes principais não-lineares. Esta é quantificada pela diferença entre as componentes principais não-lineares, ponderadas pelos respectivos *loadings*, e as variáveis transformadas.

Considere-se ainda que na função (3.1), o vector $\mathbf{f}_j$ das imagens de $\mathbf{h}_j$ está associado a uma transformação *spline* de grau v, com r nós interiores e por isso gerada por w *I-splines*, com $w = v+r$. Assim, a classe de transformações da qlPCA fica completamente determinada fixando o grau v e o número de nós interiores r de cada transformação ϕ_j.

Sublinhe-se que o problema de ajustamento é linear nos coeficientes da base de *I-splines*, mas altamente não-linear no número de nós e no seu posicionamento (Winsberg e Ramsay [36]), sendo por isso aconselhável que estes sejam parâmetros do modelo e não objectos de optimização. Sugere-se que o investigador experimente diferentes escolhas para estes parâmetros e observe o efeito produzido no modelo, nomeadamente em termos de percentagem de variância explicada (ver parágrafo A da secção 3.1.4). No en-

[1] Ver mais detalhes na secção 1.2.

tanto, tal como em qualquer modelo estatístico, a ACP não-linear através de *splines* é susceptível de sobre-parametrização. De modo a prevenir a sobre-parametrização deve existir um número razoável de dados na vizinhança de qualquer nó interior (Winsberg e Ramsay [36]).

Tendo fixado a classe de transformações, o objectivo é determinar os *object scores*, $\mathbf{X}$, e as transformações ϕ_j para cada variável de modo a minimizar a função (3.1), sujeita à restrição $\mathbf{X}'\mathbf{X} = n\mathbf{I}$.

Seja $\mathbf{f}_j$ o vector das imagens de $\mathbf{h}_j$ através de um *spline* ϕ_j de grau v com r nós interiores, gerado por $w = v + r$ *I-splines*

$$\mathbf{f}_j = \phi_j(\mathbf{h}_j) = \sum_{i=1}^{w} \alpha_{ji} I_{ji}^{[v]}(\mathbf{h}_j). \tag{3.2}$$

Tem-se que,

$$\mathbf{f}_j = \mathbf{G}_j^{\triangle}\mathbf{y}_j, \tag{3.3}$$

onde: $\mathbf{G}_j^{\triangle}$ é a matriz pseudo-indicatriz da variável j, de ordem $n \times w$ cujas colunas são os vectores das imagens da variável j por cada uma das funções *I-splines* da base; $\mathbf{y}_j$ é o w-vector cujos elementos são os coeficientes da combinação linear $\mathbf{y}_j = [\alpha_1 \alpha_2 \ldots \alpha_w]'$.

Usando a equação (3.3) é possível redefinir o processo de quantificação óptima como sendo a fase que, dados os *object scores*, optimiza os vectores $\mathbf{y}_j$ por forma a minimizar a função perda, ou de forma análoga, como a fase que determina a combinação linear óptima de cada base de *splines*, dados os *object scores*.

3.1.3 Algoritmo: pseudo-código

Seja $\mathbf{H}$ a matriz $n \times m$ dos dados *estandardizados*, p o número de componentes principais a reter, r o número de nós interiores e v o grau do *spline*. O algoritmo da qlPCA usa um processo de *Mínimos Quadrados Alternados* (MQA) para minimizar (3.1). O algoritmo consiste em duas fases de forma iterativa e alternada até ser obtida a convergência: a fase de actualização

das quantificações e a fase da estimação dos *object scores*. O diagrama do algoritmo da qlPCA pode ser observado na figura 3.1.

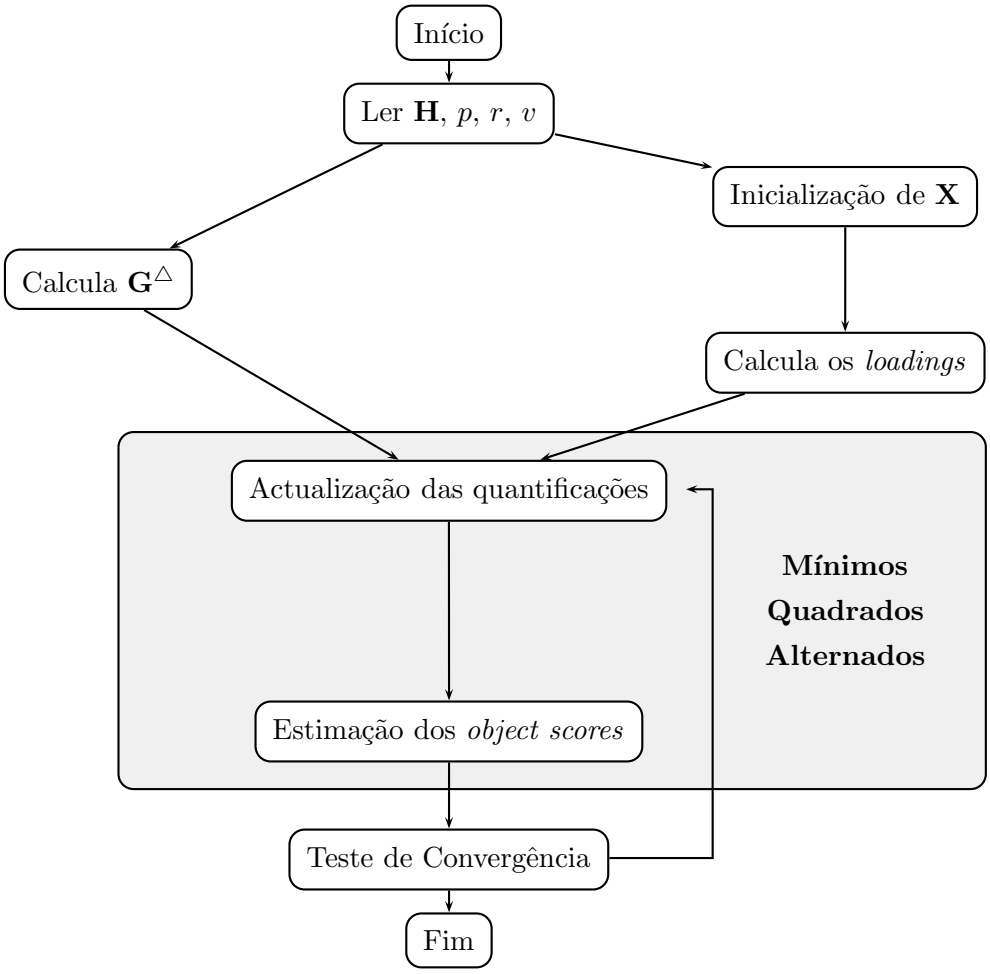

Figura 3.1: Diagrama do algoritmo da qlPCA.

Sejam **H**, p, r e v os *inputs* do algoritmo. A configuração inicial é estabelecida como se segue.

I. Inicialização

1: $\mathbf{Z}_{n\times p} \leftarrow linearPCA(\mathbf{H})$

2: $[K, \Lambda^{1/2}, W] \leftarrow svd(\mathbf{Z})$

3: $\mathbf{X} \leftarrow \sqrt{n}\mathbf{K}\mathbf{W}'$

4: $\mathbf{a}_j \leftarrow \frac{1}{n}\mathbf{X}'\mathbf{h}_j, j = 1, \ldots, m$

5: $[\mathbf{G}_1^{\triangle} \ldots \mathbf{G}_m^{\triangle}] \leftarrow createIspline(H, v+1, r)$

A primeira operação incluída na inicialização consiste em efectuar uma ACP linear sobre $\mathbf{H}$, sendo retidas as p componentes principais na matriz dos *object scores* $\mathbf{Z}$. Esta matriz é ortonormalizada (passos 2 e 3, sendo *svd* a decomposição em valores singulares) de tal forma que $\mathbf{X}'\mathbf{X} = n\mathbf{I}$. No passo 4, $\mathbf{a}_j = [a_{1j} \ldots a_{sj} \ldots a_{pj}]'$ para cada j, sendo a_{sj} o coeficiente de correlação de Pearson entre a variável j e a s-ésima componente principal, também conhecido como *component loading*. O passo 5 calcula a matriz do tipo $n \times mw$, $\mathbf{G}^{\triangle} = [\mathbf{G}_1^{\triangle} \ldots \mathbf{G}_m^{\triangle}]$, através do processo recursivo descrito na secção 1.1 e justapondo as matrizes de codificação difusa. A implementação deste último processo será detalhada na próxima secção.

II. Mínimos Quadrados Alternados: fase de actualização das quantificações (ciclo em j, corre as variáveis):

1: $\tilde{\mathbf{y}}_j \leftarrow \mathbf{X}\mathbf{a}_j$

2: $\mathbf{y}_j \leftarrow \frac{1}{n}((\mathbf{G}_j^{\triangle})'\mathbf{G}_j^{\triangle})^{-1}(\mathbf{G}_j^{\triangle})'\tilde{\mathbf{y}}_j$

3: $\mathbf{f}_j = f_j(\mathbf{h}_j) \leftarrow \mathbf{G}_j^{\triangle}\mathbf{y}_j$

4: $\mathbf{f}_j \leftarrow \mathbf{f}_j - \overline{f}_j\mathbf{u}$

5: $\mathbf{f}_j \leftarrow \sqrt{n}\frac{\mathbf{f}_j}{\|\mathbf{f}_j\|}$

6: $\mathbf{a}_j \leftarrow \frac{1}{n}\mathbf{X}'\mathbf{f}_j$

Nesta fase actualizam-se as quantificações das variáveis, $\mathbf{f}_j$ (passos 1 a 5) e os vectores $\mathbf{a}_j$ (passo 6). Para minimizar a soma dos quadrados dos resíduos entre $\mathbf{X}\mathbf{a}_j$ e a variável transformada j, recorre-se à estimação de mínimos quadrados (passo 2). Note-se que no passo 2 é realizada uma regressão linear

múltipla, sendo $\mathbf{X}\mathbf{a}_j$ a variável dependente e as colunas de $\mathbf{G}_j^{\triangle}$ as variáveis independentes. Este passo actualiza $\mathbf{y}_j$, um vector cujas entradas são os coeficientes da combinação linear para a base de *I-splines*, que por sua vez vão actualizar a quantificação de cada uma das variáveis $\mathbf{f}_j$. Cada variável transformada é centrada (passo 4) e normalizada para $\sqrt{n}$ (passo 5), de forma a que a sua variância seja unitária. No passo 6, $\mathbf{a}_j$ é actualizado, sendo agora a_{sj}, o coeficiente de correlação de Pearson entre a variável transformada j e a s-ésima componente principal não-linear, tendo por isso a designação de *loading* não-linear.

III. Mínimos Quadrados Alternados: fase de estimação dos objects scores

1: $\mathbf{Z}_{n\times p} \leftarrow linearPCA(\mathbf{F})$

2: $[\mathbf{K}, \mathbf{\Lambda}^{1/2}, \mathbf{W}] \leftarrow svd(\mathbf{Z})$

3: $\mathbf{X} \leftarrow \sqrt{n}\mathbf{K}\mathbf{W}'$

Na fase de estimação dos *object scores* é actualizada a matriz $\mathbf{X}$. O passo 1 efectua uma ACP linear sobre $\mathbf{F}$, retendo as p componentes principais na matriz dos *object scores* $\mathbf{Z}$. Esta matriz é ortonormalizada (passos 2 e 3, sendo *svd* a decomposição em valores singulares) de tal forma que $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ atribui a $\mathbf{Z}$ o somatório de $\mathbf{F}_j$ em j, sendo por isso $\mathbf{Z}$ uma matriz centrada. A matriz $\mathbf{Z}$ é ortonormalizada (passos 2 e 3) tal que $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, ficando os *object scores* com variância unitária.

IV. Teste de convergência

O teste de convergência é realizado testando as diferenças entre dois sucessivos valores da função perda (3.1) contra 0.1×10^{-5}. O algoritmo volta ao passo II.1 se a convergência falha ou pára se o número máximo de iterações for atingido. Atingida a convergência, a qlPCA produz os seguintes *outputs*:

- $\mathbf{X}$ - matriz $n \times p$ contendo os *objects scores* não-lineares;
- $\mathbf{F}$ - matriz $n \times m$ contendo as variáveis transformadas de forma óptima;
- $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_m]$ - matriz $p \times m$ com os *loadings* não-lineares;

- **y** - mw-vector com os coeficientes óptimos associados aos m *I-splines* da base.

3.1.4 Propriedades da qlPCA

Aplicar a qlPCA envolve tentar diferentes opções para os parâmetros de entrada: grau do *spline* e número de nós interiores. Os *splines* lineares têm várias vantagens, mas o utilizador pode experimentar um ajustamento com *splines* de ordem superior, estando esta possibilidade implementada na versão actual da qlPCA. A implementação actual da qlPCA não permite ao utilizador a escolha da localização dos nós: estes são sempre colocados em percentis igualmente espaçados e por isso cada sub-intervalo contém aproximadamente o mesmo número de observações. Já foi mencionado que para prevenir a sobre-parametrização uma quantidade razoável de observações deve estar na vizinhança de qualquer nó interior (Winsberg e Ramsay [36]). A causa da sobre-parametrização está relacionada com o passo II.2 do algoritmo da qlPCA, no qual é realizada uma regressão multivariada com $w = v + r$ (soma do grau do *spline* com o número de nós interiores) variáveis preditoras. Com a opção apresentada para o posicionamento dos nós, prevenir o sobre-parametrização é uma questão do número de observações disponíveis, n, versus w. Na literatura, não existe consenso relativamente ao número mínimo de observações disponíveis por cada preditor na regressão linear multivariada. É comummente referido que o valor mínimo aceitável são 5 observações por preditor, mas também que é mais confortável se estiverem disponíveis pelo menos 10 observações por preditor.

O algoritmo da qlPCA irá tirar partido de *splines* lineares, sem limitações no que diz respeito ao número de nós interiores. Esta variante não-linear da ACP é uma generalização imediata da ACP no que diz respeito às medidas de desempenho e de interpretação. Nesta secção, são introduzidos resultados sobre o sumário do modelo, projecção de novas observações no espaço das componentes principais não-lineares e reconstrução das variáveis originais. Um novo conceito é ainda definido nesta secção - *piecewise loadings* - como as correlações "segmentadas" entre as componentes principais não-lineares e

as variáveis originais.

Dada a matriz $\mathbf{H}$, do tipo $n \times m$, a sua representação no espaço das componentes principais não-lineares pode ser encontrada minimizando a função perda (3.1) usando o algoritmo descrito na secção anterior. Dados os parâmetros da qlPCA (p - número de componentes principais retidas, r - número de nós interiores e v o grau do *spline* e por consequência $w = v + r$ a dimensão do espaço linear de funções *spline*), depois de terminada a optimização, as seguintes matrizes e vectores, definidos na secção anterior, são conhecidos: $\mathbf{X}$,$\mathbf{F}$,$\mathbf{A}$ e $\mathbf{y}$.

A. Sumário do modelo

Tal como na ACP, é interessante saber a capacidade do modelo para explicar a variância total da matriz de dados original. No entanto, nas variantes não-lineares da ACP, esta medida refere-se à capacidade do modelo para explicar a variância total da matriz transformada óptima.

Define-se a *Variance Accounted For* (VAF) *por dimensão* como

$$VAF_s = \sum_{j=1}^{m} a_{sj}^2, s = 1, \ldots, p$$

onde a_{sj} é o coeficiente de correlação de Pearson entre a variável j e a s-ésima componente principal não-linear, o *loading* não-linear. Por consequência, a percentagem de variância explicada é $VAF_s \times 100/m$.

A percentagem de variância explicada pelo modelo é a medida mais usual da qualidade do ajustamento, sendo definida como o somatório das VAF pelas várias dimensões retidas. No entanto, esta medida não pode ser usada para comparar o ajustamento de uma ACP com o de uma variante não-linear, pois como foi referido tratam-se de variâncias explicadas de matrizes diferentes. Apesar deste facto, é comum encontrar-se na literatura esta comparação (p.ex. Lavado e Calapez [22, 23], Meulman et al. [30], Linting et al. [27], Linting e Kooij [25]). Sabe-se porém que as variantes não-lineares da ACP ficam, em geral, em desvantagem em relação à ACP nesta comparação. Como refere Lebart [24] a propósito da *Multiple Correspondence Analysis*

(MCA), "percentages of variance are misleading measures of information". Asan e Greenacre [1], subscrevem esta observação a propósito também da *fuzzy MCA*, justificando que esta desvantagem se deve ao facto das variantes não-lineares terem mais variância para explicar do que a ACP tradicional[2]. Asan e Greenacre [1] referem ainda que a lógica desta medida no seio das variantes não-lineares é incorrecta, pelo que deveriam ser propostas outras alternativas. A revisão da literatura efectuada até ao início de 2012, mostra que esta é ainda uma questão em aberto. Defende-se que a mesma conclusão se aplica à qlPCA, pelo que comparações entre as variantes não-lineares e entre estas e a ACP deve ser efectuada com precaução.

Seja $\mathbf{R}$ a matriz de correlações de $\mathbf{F}$. Cada variável transformada está centrada e normalizada para $\sqrt{n}$, assim, $\mathbf{R} = n^{-1}\mathbf{F}'\mathbf{F}$. À semelhança da ACP, pode ser demonstrado que os primeiros p valores próprios de $\mathbf{R}$ são iguais a $VAF_s,\ s = 1, \ldots, p$, e que o s-ésimo elemento da diagonal de $\mathbf{\Lambda}^{1/2}$ (passo III.2) é igual a $\sqrt{n}VAF_s$. Assim, a definição da VAF introduz o conceito de valor próprio na qlPCA, sendo uma generalização directa da ACP sobre a capacidade do modelo não-linear explicar a variância total da matriz dos dados transformados de forma óptima. Por este motivo, pode-se também definir a percentagem de VAF por dimensão como o seu valor próprio dividido pelo número de variáveis.

Outra medida importante de ajustamento é a *VAF por variável transformada*, também designada por *comunalidade*, definida como o somatório do quadrado dos *loadings* dessa variável nas componentes retidas. Define-se ainda a *VAF por dimensão por variável* como o quadrado do *loading* dessa variável, nessa dimensão. Quando o objectivo é a selecção das variáveis mais relevantes para o modelo, é usual seleccionarem-se as variáveis que apresentam VAF por variável transformada (*comunalidade*) superior ou igual a 0.25 (Linting e Kooij [25]). Ou seja, pelo menos 25% da variância da variável

[2]Asan e Greenacre [1], referem que: "It is well-known that regular MCA gives very pessimistic estimates of explained variance [...] and the same is true for fuzzy MCA [...] this is because fuzzy MCA embeds the data in a much higher-dimensional space [...] compared to PCA [...], so the chances of good reconstruction of the data in a two-dimensional solution are better for PCA."

transformada é explicada pelas componentes principais não-lineares retidas. Linting e Kooij [25], usam esta abordagem na análise preliminar dos dados, excluindo da CATPCA final as variáveis que numa CATPCA preliminar apresentam valores abaixo dos 25%[3].

Linting e Kooij [25], apresentam uma forma de inspecção da não-linearidade multivariada nos dados através das *VAFs por variável transformada* e dos gráficos das transformações (valores da variável original versus valores da variável transformada). As autoras referem que uma estrutura é linear se a maior parte dos gráficos das transformações apresentar funções "aproximadamente" lineares e se as variáveis associadas a transformações não-lineares apresentarem uma *VAF da variável transformada* relativamente baixa e concluem que, nestas situações as variantes não-lineares são desnecessárias, ou seja, a estrutura de dados é linear. No entanto, mesmo nestas situações, a confirmação da não necessidade de uma variante não-linear constitui em si mesmo um importante motivo para aplicar as variantes não-lineares numa fase inicial. Recorrendo às conclusões de Linting e Kooij [25], pode reafirmar-se que as comparações entre as variantes não-lineares e entre estas e a ACP deve ser efectuada com precaução, ou seja, para além da comparação das VAFs dos modelos deve-se inspeccionar os gráficos das transformações e as *VAFs por variável transformada*.

B. Escolha do número apropriado de componentes a reter

Tal como na ACP, o investigador tem que decidir qual o número adequado de componentes principais a reter na solução. Na ACP essa decisão é tomada perante a solução com todas as possíveis componentes principais. Na qlPCA, tal como na CATPCA, o número de componentes a reter, p, é um *input* do algoritmo, pelo que a solução não apresenta todas as componentes principais possíveis. No entanto, a escolha do número apropriado de componentes a reter é fundamentada nos mesmos critérios, usando para o efeito os valores

[3]Linting e Kooij[25] apresentam uma aplicação em que passa de 66 variáveis iniciais para 42 recorrendo à análise das VAF por variável transformada numa CATPCA preliminar.

próprios da matriz de correlação das variáveis transformadas.

Um dos critérios mais conhecidos é reter todas as componentes associadas a valores próprios superiores à unidade. No entanto, na literatura existe o consenso que este critério está entre os métodos com menos precisão (Linting et al. [26], Jackson [15]).

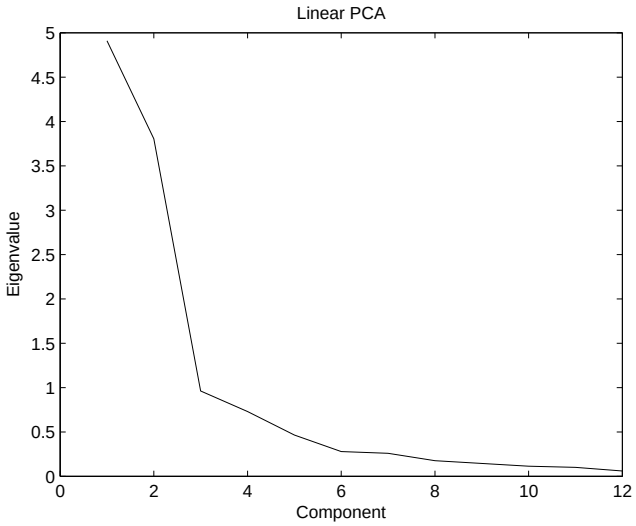

Figura 3.2: Exemplo de *Scree plot* para uma ACP linear. No eixo das ordenadas estão os valores próprios da matriz de correlações das variáveis originais.

Critérios alternativos incluem o *scree plot, parallel analysis, Velicers partial correlation technique, cross-validation* entre outros (Jackson [15]). O teste do *scree plot* envolve examinar o gráfico que cruza a identificação da componente principal retida (no eixo das abcissas) com o seu valor próprio ou percentagem de variância explicada por essa componente. Pretende-se determinar o ponto de quebra, ou "cotovelo", ou seja, os pontos onde a derivada apresenta uma quebra abrupta (ver figura 3.2).

Existe alguma discussão na literatura sobre se a componente onde ocorre o ponto de quebra deve ser incluído na solução (Linting et al. [26]). Uma das razões para não incluir essa componente é a sua fraca contribuição para a percentagem de variância explicada pelo modelo. Assim, o número de pontos

com valores próprios acima do ponto de "quebra", ou seja, não incluindo o ponto onde a quebra ocorre, é considerado o número de componentes a reter. No entanto, nem sempre é claro qual o ponto de "quebra". A figura 3.2 não apresenta uma "quebra" clara, pois o declive apresenta uma quebra abrupta duas vezes, na terceira e sexta componentes. Assim, este *scree plot* sugere a retenção de duas ou cinco componentes principais.

Como o algoritmo da qlPCA minimiza a função perda para um dado número p de componentes principais a reter, a solução com $p+1$ componentes retidas não contém necessariamente a solução com p componentes. Esta propriedade designa-se por "soluções não encaixadas"[4]. Ou seja, os primeiros p valores próprios obtidos da matriz de correlações das variáveis transformadas de forma óptima pela qlPCA com *input* p são, em geral, diferentes dos primeiros p valores próprios obtidos da matriz de correlações das variáveis transformadas de forma óptima pela qlPCA com *input* $p+1$. Por isso, os *scree plots* são distintos para diferentes dimensionalidades, pelo que se devem comparar os *scree plots* das soluções p, $p-1$ e $p+1$ (Linting et al. [26]). A "quebra" no *scree plot* associado à solução da qlPCA de *input* p deve ser consistente com as "quebras" nos *scree plots* associado às soluções $p-1$ e $p+1$ para que o número adequado de componentes a reter seja p (Linting et al. [26]).

C. Projecção de novas observações no espaço das componentes principais não-lineares

Todas as propriedades apresentadas até aqui são válidas para *splines* de qualquer grau. As propriedades apresentadas de seguida são apenas válidas para *splines* lineares e foram elas a motivação para a designação *quasi*-linear PCA. Seja $\mathbf{h}'_{nova}$ o m-vector linha com as novas observações nas m variáveis, obtidas em condições similares às anteriores. O problema da projecção de novas observações no espaço das componentes principais não-lineares é determinar o p-vector linha, $\mathbf{x}'_{novo}$, correspondente aos *object scores* não-lineares. Se

[4]do inglês *not nested*.

a matriz dos dados original não está estandardizada, os vectores das médias e variâncias devem ser registados e o algoritmo da qlPCA aplicado na matriz estandardizada. Uma correcção em $\mathbf{h}'_{nova}$ é aplicada usando esse vectores.

Sublinhe-se que o passo III.3 do algoritmo, pela decomposição em valores singulares $\mathbf{Z} = \mathbf{K}\mathbf{\Lambda}^{1/2}\mathbf{W}'$, pode ser reescrito como $\mathbf{X} \leftarrow \sqrt{n}\mathbf{Z}\mathbf{W}\mathbf{\Lambda}^{-1/2}\mathbf{W}'$. Seja $\mathbf{A}$ a matriz dos *loadings* não-lineares, tem-se que (passo III.1),

$$\mathbf{Z} = \mathbf{F}\mathbf{A}', \tag{3.4}$$

logo

$$\mathbf{X} = \sqrt{n}\mathbf{F}\mathbf{A}'\mathbf{W}\mathbf{\Lambda}^{-1/2}\mathbf{W}', \tag{3.5}$$

onde $\mathbf{W}$ e $\mathbf{\Lambda}^{-1/2}$ derivam da decomposição em valores singulares do passo III.2.

A representação das novas observações no espaço das componentes não-lineares é obtida em dois passos.

Primeiro, os valores transformados de $\mathbf{h}'_{nova}$ são calculados (ciclo pelas variáveis j, $j = 1, \ldots, m$):

if $min(\mathbf{h}_j) \leq h_{nova,j} \leq max(\mathbf{h}_j)$ **then**
  $f_{nova,j} \leftarrow interp(\mathbf{h}_j, \mathbf{f}_j, h_{nova,j})$,
else
  $f_{nova,j} \leftarrow extrap(\mathbf{h}_j, \mathbf{f}_j, h_{nova,j})$,
end if

Para valores dentro da amplitude de cada variável, é realizada uma interpolação linear para determinar $f_{nova,j}$, a imagem da função *spline* óptima subjacente f_j para o valor $h_{nova,j}$, a nova observação na variável j. Sublinhe-se que após a convergência, o *spline* linear óptimo com r nós interiores fica completamente definido com $r + 2$ pontos: r associados aos nós interiores e dois associados ao máximo e mínimo de cada variável. Assim, a interpolação fica bem definida com esses $r + 2$ pontos para determinar o valor exacto do *spline*. Para valores fora da amplitude da variável original é realizada uma extrapolação linear.

Segundo, os *object scores* não-lineares para as novas observações são obtidos usando a equação (3.5) com a matriz transformada óptima do tipo $1 \times m$, $\mathbf{F}_{nova} = [f_{nova,1} \dots f_{nova,m}]$.

D. Reconstrução

O problema da reconstrução dos dados originais refere-se a encontrar uma estimativa $\hat{\mathbf{H}}$ de $\mathbf{H}$ usando os *object scores* não-lineares. Este processo inclui dois passos: dos *object scores* não-lineares $\mathbf{X}$ para uma aproximação $\hat{\mathbf{F}}$ da matriz dos dados transformados; de $\hat{\mathbf{F}}$ para uma estimativa $\hat{\mathbf{H}}$ dos dados originais.

Quando $p < m$, as componentes principais não-lineares $\mathbf{X}\mathbf{a}_j$ são uma aproximação da variável transformada $\mathbf{f}_j$ (algoritmo da qlPCA, passo II.2), pelo que $\mathbf{XA}$ é uma estimativa $\hat{\mathbf{F}}$ de $\mathbf{F}$. Logo

$$\mathbf{F} = \mathbf{XA} + (\mathbf{F} - \hat{\mathbf{F}}) \tag{3.6}$$

onde o primeiro termo do membro da direita da equação representa a contribuição devido às componentes principais não-lineares retidas e o segundo termo representa a parte não explicada pelo modelo da qlPCA - o resíduo.

Tendo usado a qlPCA com *splines* lineares em cada coluna, $\mathbf{XA}$ é por isso uma aproximação das imagens do *spline* linear óptimo. A interpolação linear inversa dá de forma exacta a função inversa de uma função seccionalmente linear. Por isso, usando interpolação linear inversa, com os valores de $\mathbf{X}\mathbf{a}_j$ para valores na amplitude do j-ésimo *spline* e extrapolação linear inversa para valores fora da amplitude, obtém-se $\hat{\mathbf{H}}$.

E. Piecewise Loadings

Nesta secção, por motivos de apresentação dos resultados, seja $\mathbf{x} \leftarrow \mathbf{x}_s$ a s-ésima componente principal, $\mathbf{f} \leftarrow \mathbf{f}_j$ a j-ésima variável transformada e $\mathbf{h} \leftarrow \mathbf{h}_j$ a j-ésima variável original. No seio da qlPCA o termo *loading* refere-se ao coeficiente de correlação de Pearson entre $\mathbf{x}$ e $\mathbf{f}$. Nesta secção, $\mathbf{x}_i$, $\mathbf{f}_i$

e $\mathbf{h}_i$ designam os vectores truncados obtidos de $\mathbf{x}$, $\mathbf{f}$ e $\mathbf{h}$, respectivamente, com elementos do i-ésimo segmento desses vectores, $i = 1, \ldots, r+1$, definido por dois nós consecutivos. A ideia dos *piecewise loadings* é obter correlações segmentadas entre os vectores truncados $\mathbf{x}_i$ e $\mathbf{h}_i$ através de $\mathbf{f}_i$. Assim, os *piecewise loadings* podem revelar o comportamento por segmentos entre as componentes principais não-lineares e as variáveis originais.

Considere-se o seguinte resultado auxiliar: sejam $\mathbf{y}$ e $\mathbf{z}$ dois vectores, não necessariamente centrados ou estandardizados, com a mesma dimensão e g uma função linear, sendo $\mathbf{g} = g(\mathbf{z}) = m\mathbf{z} + b\mathbf{u}$ o vector das imagens de $\mathbf{z}$ segundo g, onde $\mathbf{u}$ é um vector de uns. Pode ser demonstrado que:

$$corr(\mathbf{y}, g(\mathbf{z})) = \begin{cases} corr(\mathbf{y}, \mathbf{z}), & m > 0 \\ \\ -corr(\mathbf{y}, \mathbf{z}), & m < 0 \end{cases} \tag{3.7}$$

Depois da convergência da qlPCA, $\mathbf{x}$ e $\mathbf{f}$ estão disponíveis, assim é possível calcular $corr(\mathbf{x}_i, \mathbf{f}_i)$ para cada segmento i. Estando a ser usados *splines* lineares, em cada segmento $\mathbf{f}_i$ e $\mathbf{h}_i$ estão relacionados por $\mathbf{f}_i = m_i\mathbf{h}_i + b_i\mathbf{u}$ para um m_i e b_i conhecidos. Assim, pelo resultado anterior:

$$corr(\mathbf{x}_i, \mathbf{h}_i) = \begin{cases} corr(\mathbf{x}_i, \mathbf{f}_i), & m_i > 0 \\ \\ -corr(\mathbf{x}_i, \mathbf{f}_i), & m_i < 0 \end{cases} \tag{3.8}$$

Sendo os *splines* obtidos por combinação linear da base de *I-splines*, e sendo as funções da base monótonas crescentes,

3.2 Implementação em MatLab

O algoritmo da qlPCA foi implementado usando a versão 2009b do MatLab. O algoritmo permite a utilização de *splines* de qualquer grau e com qualquer número de nós interiores. No entanto, ainda não permite que essa escolha seja efectuada variável a variável, ou seja, é usado o mesmo tipo de *splines* para todas as variáveis. Na figura 3.3 pode observar-se o esquema da implementação em MatLab. Nas secções seguintes apresenta-se o código MatLab

comentado, segundo a sequência do pseudo-código apresentado na secção 3.1.3.

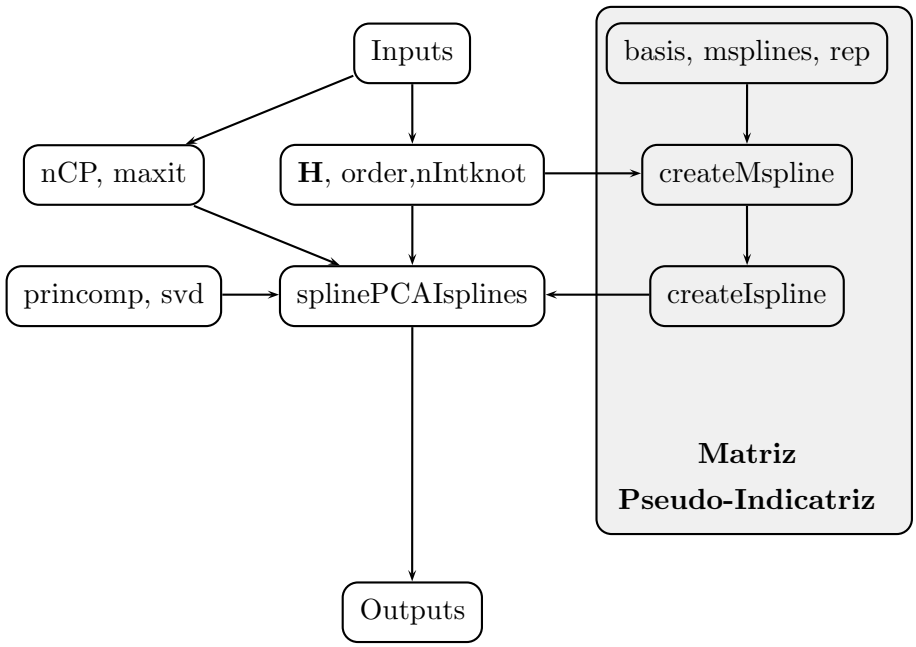

Figura 3.3: Esquema da implementação da qlPCA.

3.2.1 Inicialização: passos I.1 a I.5

A inicialização, passos I.1 a I.5 do pseudo-código, é implementada conforme o código que se apresenta de seguida.

Código MatLab da função principal da qlPCA: parte 1 - Inicialização.

```
1  function [F,X,A,eigval,knots,coef]=splinePCA_Isplines(H)
2  %Input: H - data matrix n, m
3  %Output:
4  %F - matriz dos dados transformadositeration history(total)
5  %X - objects scores no lineares, matrix n por p=nCP
6  %A - matriz dos loadings
7  %eigval - valores prprios
8  %knots - seq. de ns
9  %coef - coefs das cl da base de splines
10 %Sub functions
11 %   [G,knots]=create_Ispline(H,order,n_intknot)
12 %   MatLab functions: pcacov, svd
13 %28/03/2012 - Nuno Lavado .: nlavado@isec.pt
```

```
m=size(H,2);
n=size(H,1);
nCP=input('Dimensions in solution:');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%Inicializa o
H=zscore(H);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%constroi a matriz difusa para H
%Ordem dos Isplines (coincide com o grau para esta base)
order=input('Spline degree:');
n_intknot=input('Interior knots:');
[G,knots]=create_Ispline(H,order,n_intknot);
dimensao=order+n_intknot;
G=zscore(G);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%SCORES aleatrios
%X=rand(n,nCP);
X=G(:,1:nCP);
max_it = 1000;
it=1;
tol=1;
coef=zeros(m*dimensao,1);
F=zeros(n,m);
eigval=zeros(nCP,1);
A=corr(H,X);
communality=sum(A.^2,2);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

splinePCA_Isplines1.m

A função `create_Ispline` determina a super-matriz pseudo-indicatriz, $\mathbf{G}^{\triangle} = [\mathbf{G}_1^{\triangle} \dots \mathbf{G}_m^{\triangle}]$, conforme o passo I.5 do pseudo-código, através do processo recursivo descrito no primeiro capítulo, tendo sido implementada com o código que se segue.

Código MatLab da função para construir a base de I-splines.

```
function [G,knots]=create_Ispline(H,order,n_intknot)
%subfunction: create_mspline
n=size(H,1);
m=size(H,2);
dimensao=order+n_intknot;
for j=1:m%corre as m variaveis de H
    h=H(:,j);
    Gaux=zeros(n,dimensao);
    %constroi a matriz pseudoindicatriz associada a var j
    x=h;
```

```
11      [s,knots2]=isplines(x,order, n_intknot);
12      Gaux=s;
13      knots(:,j)=knots2;
14      if j==1 G=Gaux; else G=[G Gaux];
15      end
16  end
17  end
18  function [s,knots2]=isplines(x, order, n_intknot)
19  dim=order+n_intknot;
20  [G2,knots2]=create_mspline(x,order+1,n_intknot);
21  n=length(x);
22  I=zeros(n,dim);
23  k=order;
24  for i=1:dim
25      for t=1:n
26          if x(t)==knots2(end)
27              I(t,i)=1;
28          else
29          j=sum(knots2<=x(t));
30              if i>j
31              I(t,i)=0;
32              else
33                  if j-k<=i<=j
34                  m=i+1;
35                  I(t,i)=((knots2(m+k+1)-knots2(m))/(k+1))*G2(t,m);
36                      for m=i+2:j
37                      I(t,i)=I(t,i)+((knots2(m+k+1)-knots2(m))/(k+1))*G2(t,m);
38                      end
39                  else
40                      if i<j-k
41                      I(t,i)=1;
42                      end
43                  end
44              end
45          end
46      end
47  end
48  s=I;
49  end
```

create_Ispline.m

A função `create_Ispline` invoca a sub-função `create_mspline`, cujo código se apresenta de seguida.

Código MatLab da função para construir a base de M-splines.

```
1  function [G,knots]=create_mspline(H,order,n_intknot)
```

```
%subfunctions: msplines, basis, rep
n=size(H,1);
m=size(H,2);
%constroi os percentis associados aos nos interiores
p=zeros(1,n_intknot);
p(1)=100/(n_intknot+1);
for i=2:n_intknot
    p(i)=p(1)+(i-1)*p(1);
end
p=round(p);
dimensao=order+n_intknot;
for j=1:m%corre as m variaveis de H
    h=H(:,j);
    %sequencia de nos extraida de h segundo a seq de percentis
    inner_knots=prctile(h,p);
    boundary_knots=[min(h) max(h)];
    Gaux=zeros(n,dimensao);
    boundary_knots=sort(boundary_knots);
    knotsj=[rep(boundary_knots(1), order) sort(inner_knots)...
        rep(boundary_knots(2), order)];
    for i=1:n%constroi a matriz pseudoindicatriz associada a var j
        x=h(i);
        s=msplines(x,order,inner_knots,boundary_knots,knotsj);
        Gaux(i,:)=s;
    end
    knots(:,j)=knotsj;
    if j==1 G=Gaux; else G=[G Gaux];
    end
end
```

create_mspline.m

As três sub-funções auxiliares, `msplines`, `basis` e `rep`, têm o seguinte código.

Código MatLab das sub-funções para construir a base de M-splines.

```
function s=msplines(x,order, inner_knots, boundary_knots,knots)
%subfunctions: basis, rep
np=order+size(inner_knots,2);
s=rep(0, np);
for i=1:np
        s(i)=basis(x, order, i, knots);
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function y= basis(x, order, i, knots)
    if order==1
        if((x<knots(i+1))&&(x>=knots(i))) y=1/(knots(i+1)-knots(i));
```

```
        else y=0;
        end
    else
        if((knots(i+order)-knots(i))==0)
            y=0;
        else
        y=(order*((x-knots(i))*basis(x, (order-1), i, knots)...
            +(knots(i+order)-x)*basis(x,(order-1),(i+1),knots)))/...
            ((order-1)*(knots(i+order)-knots(i)));
        end
    end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function a=rep(x,times)
a=x;
for i=2:times
    a=horzcat(a,x);
end
```

sub_msplines.m

3.2.2 *Mínimos Quadrados Alternados*: passos II.1 a II.6

Nesta fase é actualizada a quantificação de cada variável $\mathbf{f}_j$.

Código MatLab da função principal da qlPCA: parte 2 - Actualização das quantificações.

```

while tol>0.1*10^(-5) &&  it <=max_it;
    jw=1;
    jb=dimensao;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Actualiza o das quantifica es
    for j=1:m
       y=X*A(j,:)';
       coef(jw:jb,1)=(1/n)*(G(:,jw:jb)'*G(:,jw:jb))^(-1)*G(:,jw:jb)'*y;
       F(:,j)=G(:,jw:jb)*coef(jw:jb,1);
       F(:,j)=F(:,j)-mean(F(:,j));
       jw=jw+dimensao;
       jb=jb+dimensao;
    end
    ssq=diag(sum(F.^2,1));
    F = F*ssq^(-1/2).*sqrt(n);%F centrado e norma sqrt(n)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

splinePCA_Isplines2.m

3.2.3 *Mínimos Quadrados Alternados*: passos III.1 a III.3

Código MatLab da função principal da qlPCA: parte 3 - Estimação dos objects scores.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Actualiza o dos scores
    evprev=eigval;
    [eigvec,eigval] = pcacov(F'*F);
    eigval=eigval(1:nCP,1)./n;
    A=eigvec(:,1:nCP)*diag(eigval)^0.5;
    %A matriz dos loadings m por nCP
    X=F*A;
    [K,lambda,W] = svd(X,0);
 X=sqrt(n)*K*W';
%Exemplo com uma CP
%   eigval=eigval(1,1)/n;
%   a=eigvec(:,1).*eigval^0.5;
%   X=sum(F*diag(a),2);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

splinePCA_Isplines3.m

3.2.4 Teste de convergência

O teste de convergência é realizado testando as diferenças entre dois sucessivos valores da função perda contra 0.1×10^{-5}. O algoritmo volta a II se a convergência falha, ou pára se o número máximo de iterações for atingido.

3.3 Exemplo

Nesta secção discute-se um exemplo em que uma estrutura de dados não-linear conhecida é simulada. Uma versão deste exemplo, chamado *problema*

do cilindro, foi usada por Winsberg e Ramsay [36], Gifi [12] e outros autores para testar as suas variantes não-lineares da ACP.

Considerem-se doze variáveis, sendo dez delas definidas como funções não-lineares das restantes duas, conforme se pode observar na tabela 3.2.

Tabela 3.2: Problema do cilindro.

Variável	Fórmula
1. Altura	a
2. Área da base	b
3. Perímetro da base	$2\sqrt{b\pi}$
4. Área lateral	$2a\sqrt{b\pi}$
5. Volume	ab
6. Momento de inércia	$ab^2/2\pi$
7. Razão de esbelteza	$a/\sqrt{2b\pi}$
8. Ângulo diagonal-base	$\arctan[a\sqrt{\pi}/(2\sqrt{b})]$
9. Ângulo diagonal-lateral	$\operatorname{arccot}[a\sqrt{\pi}/(2\sqrt{b})]$
10. Resistência eléctrica	a/b
11. Condutância	b/a
12. Deformabilidade de torsão	$2a\pi/b^2$

Trata-se de uma situação em que existe um conjunto de doze transformações linearizantes (transformações logarítmicas para todas as variáveis, excepto para as variáveis de ângulo a que se aplicam as funções compostas logaritmo após tangente e logaritmo após cotangente, respectivamente), que resultariam numa matriz de dados transformados de estrutura bi-dimensional definida através do logaritmo da altura e do logaritmo da área da base. Consequentemente, espera-se que uma ACP não-linear sobre a matriz inicial, bem como uma ACP tradicional sobre a matriz transformada, apresente um ajustamento quase perfeito com apenas duas dimensões e que as transformações *spline* óptimas revelem um comportamento aproximadamente logarítmico.

Foram efectuadas três simulações, tendo por ponto de partida 51 cilindros diferentes, gerados através de números pseudo-aleatórios para a área da base

e altura. Para cada um destes 51 cilindros foram também geradas as restantes dez variáveis da tabela anterior. Pretende-se testar também os algoritmos na presença de dados com algum nível de ruído. Assim, cada variável foi em primeiro lugar transformada pela transformação linearizante apropriada, sendo depois o ruído adicionado e finalmente os valores com ruído foram inversamente transformados. Foram considerados três cenários: um sem ruído e dois com ruído proveniente de distribuições Normais. Os desvios-padrão considerados correspondem a 10% e a 25% do desvio-padrão de cada variável original.

Tendo em conta a sugestão apresentada no parágrafo A da secção 3.1.4 para prevenir a sobre-parametrização e considerando que existem 51 indivíduos, optou-se por realizar duas qlPCA, ambas usando *splines* lineares para cada variável: uma com dois nós interiores, aproximadamente nos percentis 33 e 67, e outra com três nós interiores localizados aproximadamente nos quartis.

Na tabela seguinte apresentam-se os resultados para cada cenário, obtidos por cada técnica de redução da dimensão: qlPCA2, que designa a qlPCA com dois nós interiores; qlPCA3, que designa a qlPCA com três nós interiores; ACP, que designa a ACP tradicional. O resultado está expresso em termos de percentagem da variância explicada por duas dimensões.

Tabela 3.3: Variância explicada por duas componentes principais (%).

Algoritmo	Sem ruído	Ruído a 10%	Ruído a 25%
qlPCA2	97.77	96.90	92.59
qlPCA3	98.43	97.68	94.41
ACP	82.15	80.18	72.62

Em todos os cenários, a qlPCA apresentou uma performance muito melhor que a ACP. Devido à estrutura teórica imposta aos dados, é razoável assumir que estes têm uma estrutura bi-dimensional. De modo a testar o pior cenário, dados com ruído a 25%, foi conduzida uma análise com base em *scree plots* para validar estes pressupostos com base na ACP e na qlPCA.

Para o *scree plot* da ACP, apresentado na figura 3.4, foram usados os valores próprios da matriz de correlações das variáveis originais.

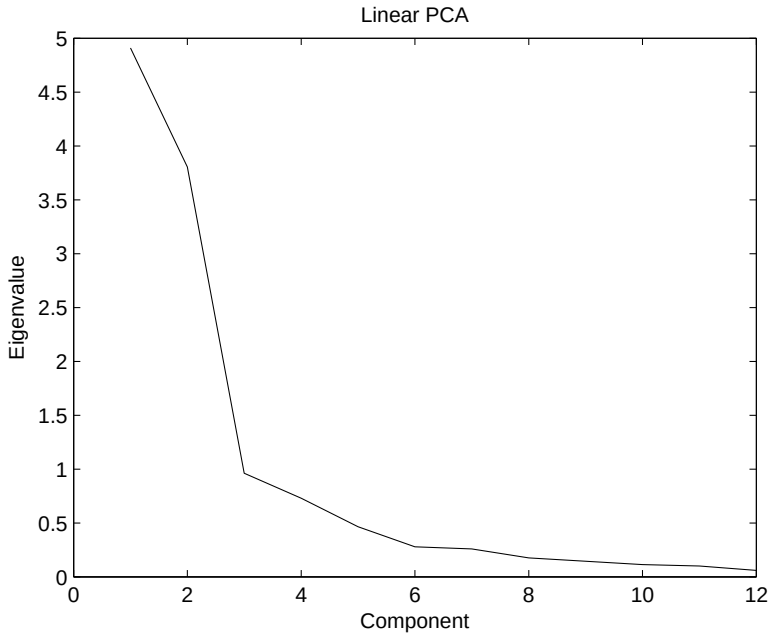

Figura 3.4: *Scree plot* para a ACP sobre os dados do cilindro com ruído a 25%.

Este gráfico não mostra apenas uma "quebra", pois o declive muda de forma abrupta não apenas uma vez mas duas, na terceira e na sexta componentes. Assim, com base neste *scree plot*, é defensável para a ACP tanto a retenção de duas como de cinco componentes principais na solução a considerar.

Para a construção do *scree plot* associado à qlPCA, são usados os valores próprios da matriz de correlações das variáveis transformadas. A figura 3.5 mostra os *scree plots* para a qlPCA com duas componentes retidas sobre os dados do cilindro com ruído a 25%.

Na figura 3.5, associada à solução bi-dimensional, observa-se que a "quebra" está localizada na terceira componente. Como as soluções da qlPCA não são "encaixadas" (ver parágrafo B da secção 3.1.4), um *scree plot* para

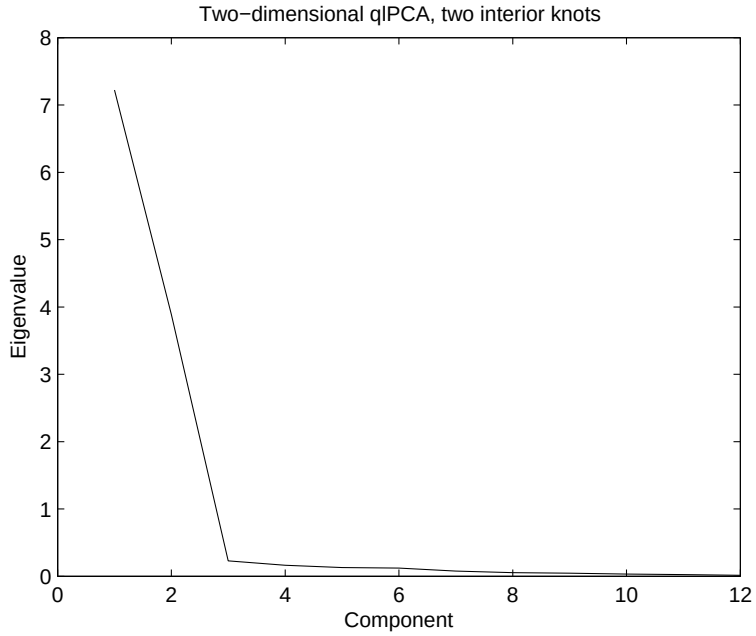

Figura 3.5: *Scree plot* para a qlPCA com dois nós interiores sobre os dados do cilindro com ruído a 25% com duas componentes retidas.

a solução com três componentes principais retidas, pode ser diferente do *scree plot* com duas. Na solução tri-dimensional, apresentada na figura 3.6, observa-se que a "quebra" está novamente localizada na terceira componente. Observada esta consistência, os gráficos sugerem que duas componentes são as apropriadas, como seria de esperar. Os gráficos para a qlPCA com três nós interiores são semelhantes.

A figura 3.7 mostra as transformações *spline* óptimas para as variáveis 6 e 12, associadas à qlPCA bi-dimensional com três nós interiores sobre os dados do cilindro sem ruído. A transformação da variável 6, do tipo logarítmica como esperado, foi escolhida por ser similar a todas as outras transformações com excepção de transformação da variável 12. Uma análise da variável 12 mostrou que os nós interiores foram colocados nos valores 3.85, 10.32 e 25.38 (os quartis da variável) e que o seu máximo era 6498.7. Este é um exemplo no qual uma localização alternativa para os nós poderia ter melhorado a

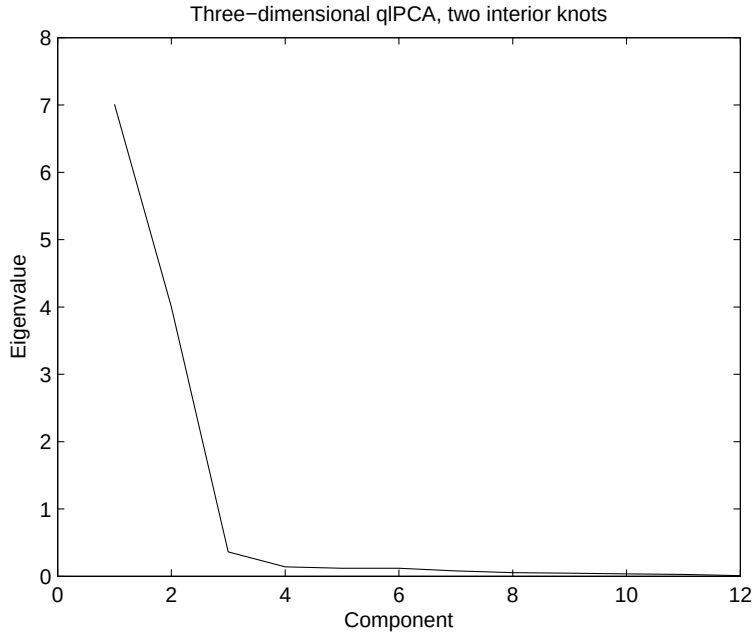

Figura 3.6: *Scree plot* para a qlPCA com dois nós interiores sobre os dados do cilindro com ruído a 25% com três componentes retidas.

transformação da variável.

Este exemplo revelou-se muito interessante para ilustrar o potencial da qlPCA. Por um lado, nas simulações realizadas, a qlPCA revelou um desempenho notável na captação da estrutura não-linear em comparação com a ACP. No pior cenário, com ruído a 25%, registou-se a maior diferença entre a ACP e a qlPCA, tendo a utilização de 3 nós interiores permitido um incremento de cerca de 22% em termos de variância explicada por duas componentes principais. Por outro lado, foi muito interessante verificar que, neste exemplo, a qlPCA apresentou transformações *spline* óptimas que se aproximavam das transformações linearizantes conhecidas.

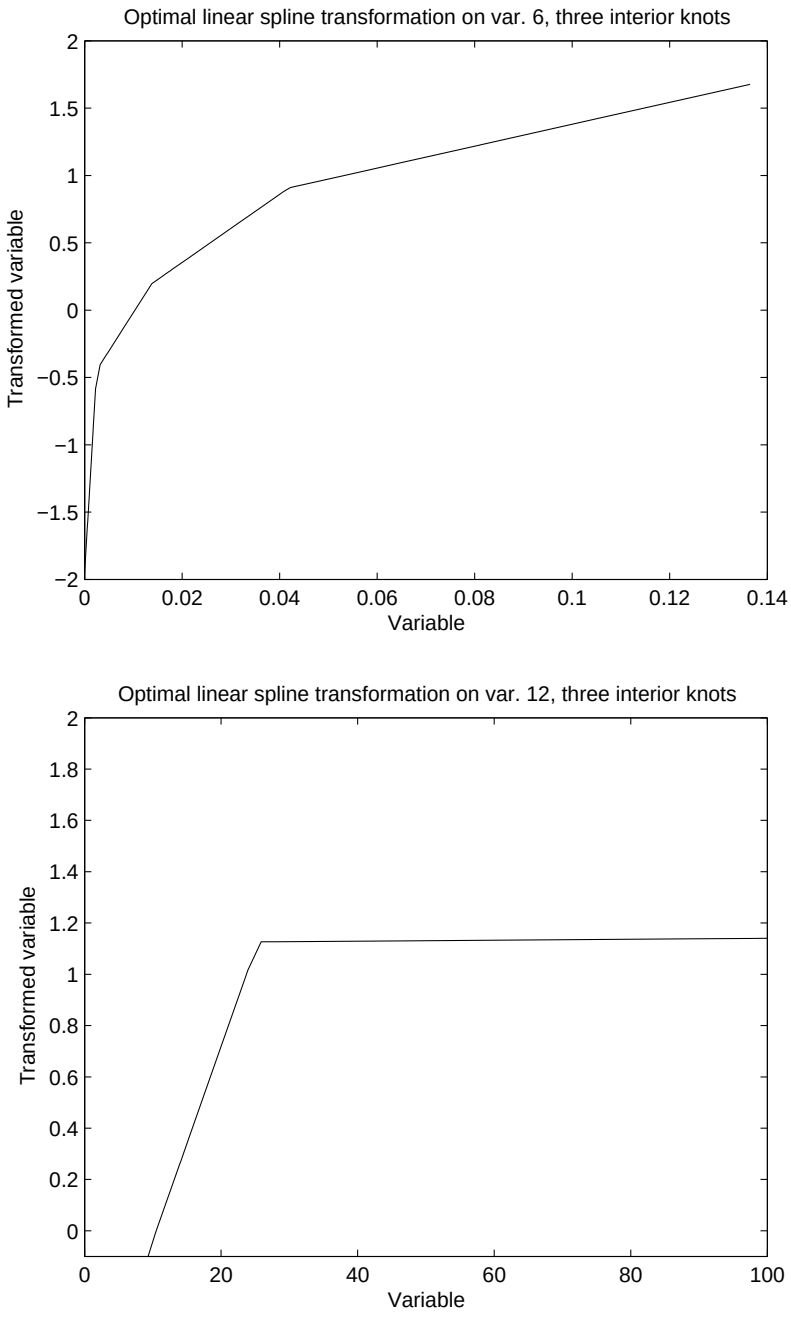

Figura 3.7: Transformações *spline* óptimas para as variáveis 6 e 12.

Capítulo 4

Aplicações

O objectivo deste capítulo é a apresentação da qlPCA através de uma aplicação ilustrativa, tentando reforçar os pontos fortes e fracos deste método. É feita uma análise sobre dados reais, relativos aos 27 países da União Europeia encarados do ponto de vista das diferenças de género na Ciência e Tecnologia. Na primeira secção apresentam-se as variáveis em análise e a respectiva base de dados, tendo por base o estudo de Oliveira e Carvalho [31]. Sendo todas as variáveis presentes neste estudo de natureza contínua, torna-se possível uma análise comparativa entre a ACP e a qlPCA. Na segunda secção são apresentados os resultados da análise destes dados com as duas abordagens referidas. Na terceira secção é feita a discussão dos resultados, apresentando algumas considerações acerca das diferenças/semelhanças na utilização e aplicabilidade das duas abordagens para o mesmo conjunto de dados.

4.1 Ciência e Tecnologia: diferenças de género

Num estudo de 2009, Oliveira e Carvalho [31] exploraram a heterogeneidade no espaço Europeu no que diz respeito aos níveis de desenvolvimento da Ciência e a Tecnologia (C&T) e aos padrões de discriminação por género.

As autoras recorreram a dados do Eurostat e do relatório *She figures 2006 report: Women and science statistics and indicators* [4], tendo seleccionado um conjunto de variáveis devidamente fundamentadas. Os dados referem-se a 2003 e 2004 para a população activa, com idades entre os 15 e os 64 anos, tendo estes sido gentilmente cedidos pelas autoras para uso nesta tese. As variáveis seleccionadas pelas autoras para a análise foram as seguintes:

- *Investigadores* - proporção de investigadores por mil indivíduos na população activa (2003)[1];
- *Cientistas e Engenheiros* - proporção de cientistas e engenheiros na totalidade da população activa (2004);
- *Investimento* - proporção de investimento em C&T por investigador em PPS[2];
- *H-M Fundos* - diferença na taxa de acesso a fundos para investigação entre homens e mulheres (2004);
- *M top* - proporção de mulheres no topo da carreira, no universo das mulheres docentes no ensino superior (2004);
- *H top* - proporção de homens no topo da carreira, no universo dos homens docentes no ensino superior (2004);
- *Glass Ceiling* - *Glass Ceiling Index* (2004)[3];
- *H-M PhD* - diferença percentual de doutorados (ISCED6) entre homens e mulheres (2003);
- *H-M investigadores* - diferença percentual de investigadores entre homens e mulheres (2003);

[1] A população activa inclui empregados e desempregados.

[2] PPS - Purchasing Power Standards.

[3] *The Glass Ceiling Index (GCI) is a ratio between the proportion of women in grade A+B+C and the proportion of women in grade A. The GCI is an indicator that measures the relative chance for women compared to men of reaching a top position.* [4].

- *H-M docentes* - diferença percentual de pessoal docente do ensino superior entre homens e mulheres (2004);
- *H-M CC* - diferença percentual nas comissões científicas entre homens e mulheres (2004).

Alguns dos 27 países apresentavam valores omissos em uma ou mais variáveis, tendo-se optado por excluir esses países da análise[4]. O conjunto de dados dos restantes 15 países a analisar[5] encontra-se no apêndice A.1. Todas as variáveis são de natureza quantitativa contínua, apresentando-se na tabela 4.1 algumas estatísticas de resumo.

Tabela 4.1: Estatísticas de resumo.

Variável	Mínimo	Máximo	Média	D.Padrão	Mediana
Investigadores	4.5	18.87	8.58	4.26	6.37
Cient. e Eng.	2.5	7.7	4.83	1.73	4.50
Investimento	14.12	169.72	75.36	51.03	71.04
H-M Fundos	-4.75	6.1	1.2	3.75	1.09
M top	1.8	16.5	5.82	3.91	4.2
H top	9.8	38.3	19.4	7.32	17.8
Glass Ceiling	1.7	3.2	2.27	0.48	2.2
H-M PhD	-34.38	29.5	6.15	19.51	14.15
H-M investig.	-6.15	65.65	31.97	19.7	31.22
H-M docentes	-15.46	41.69	23.53	16.76	30.11
H-M CC	4.92	85.61	56.19	24.53	59.46

A proporção de investigadores por mil indivíduos na população activa (2003) apresenta o valor mínimo de 4.5 na Itália e o valor máximo de 18.87 na

[4]Os países excluídos foram: Áustria, Bulgária, Chipre, França, Grécia, Irlanda, Luxemburgo, Malta, Portugal, Roménia, Espanha e Reino Unido.

[5]Os dados originais são apresentados na fonte com duas casas decimais, no entanto, como se pode observar no apêndice A.1, algumas observações apresentam três casas decimais. A introdução de diferentes valores ao nível milésimal foi a opção encontrada para que na mesma variável todos os valores tivessem frequência unitária, exigência da qlPCA na versão actual.

Finlândia. A proporção média de investigadores é de 8.58 por mil indivíduos na população activa, com um coeficiente de variação de cerca de 50%.

A proporção de cientistas e engenheiros na totalidade da população activa apresenta o valor mínimo de 2.5 na Eslováquia e o valor máximo de 7.7 na Bélgica. A proporção média de cientistas e engenheiros é de 4.83, com um coeficiente de variação de cerca de 36%.

A proporção de investimento em C&T por investigador em PPS apresenta o valor mínimo de 14.12 na Letónia e o valor máximo de 169.72 nos Países Baixos. A proporção média de investimento em C&T é de 75.36, com um coeficiente de variação de cerca de 68%. A mediana tem o valor de 71.04.

A diferença na taxa de acesso a fundos para investigação entre homens e mulheres apresenta o valor mínimo de -4.75 na Eslováquia e o valor máximo de 6.1 na Suécia. Valores negativos nesta variável correspondem a uma maior taxa de acesso a fundos para as mulheres. Os países que apresentaram valores negativos foram: Bélgica, Estónia, Finlândia, Lituânia, Países baixos, Eslováquia e Eslovénia. A diferença média na taxa de acesso a fundos é de 1.2, com um coeficiente de variação de cerca de 313%. A mediana desta variável tem o valor de 1.09.

A proporção de mulheres no topo da carreira, no universo das mulheres docentes no ensino superior, apresenta o valor mínimo de 1.8 na Lituânia e o valor máximo de 16.5 na Itália. A proporção média de mulheres no topo da carreira é de 5.82, com um coeficiente de variação de cerca de 67%. A mediana tem o valor de 4.2.

A proporção de homens no topo da carreira, no universo dos homens docentes no ensino superior, apresenta o valor mínimo de 9.8 na Alemanha e o valor máximo de 38.3 na Itália. A proporção média de homens no topo da carreira é de 19.4, com um coeficiente de variação de cerca de 38%. A mediana tem o valor de 17.8.

O *Glass Ceiling Index* apresenta o valor mínimo de 1.7 na Bélgica e o valor máximo de 3.2 na Lituânia. A média deste indicador é de 2.27, com um coeficiente de variação de cerca de 21%.

A diferença percentual entre homens doutorados e mulheres doutoradas apresenta o valor mínimo de -34.38 na Letónia e o valor máximo de 29.5

na República Checa. Valores negativos nesta variável correspondem a uma maior taxa de doutorados do sexo feminino. Os países que apresentaram valores negativos foram: Estónia, Itália, Letónia, Lituânia e Eslováquia. A diferença média é de 6.15, com um coeficiente de variação de cerca de 317%. A mediana desta variável tem o valor de 14.15.

A diferença percentual entre investigadores do sexo masculino e do sexo feminino apresenta o valor mínimo de -6.15 na Letónia e o valor máximo de 65.65 nos Países Baixos. Valores negativos nesta variável correspondem a uma maior taxa de investigadores do sexo feminino, tendo sido a Letónia o único país nesta situação. A diferença média é de 31.97, com um coeficiente de variação de cerca de 62%.

A diferença percentual entre pessoal docente do ensino superior do sexo masculino e do sexo feminino apresenta o valor mínimo de -15.46 na Letónia e o valor máximo de 41.69 na Alemanha. Valores negativos nesta variável correspondem a uma maior taxa de docentes do ensino superior do sexo feminino, tendo a Letónia sido o único país nesta situação. A diferença média é de 23.53, com um coeficiente de variação de cerca de 71%.

A diferença percentual entre homens e mulheres nas comissões científicas apresenta o valor mínimo de 4.92 na Finlândia e o valor máximo de 85.61 na Polónia. A diferença média é de 56.19, com um coeficiente de variação de cerca de 44%.

4.2 Resultados

Por motivos de visualização, usualmente pretende-se uma representação dos dados a duas dimensões. A exploração de variantes da ACP neste cenário, poderá ser justificada pela fraca percentagem de variância explicada a duas dimensões, devido à eventual existência de uma estrutura não-linear subjacente nos dados.

4.2.1 ACP

Sendo todas as variáveis deste conjunto de dados de natureza quantitativa, não há qualquer impedimento formal à sua análise através de uma ACP. Todos os resultados apresentados de seguida foram obtidos usando o MatLab (versão 2009b). A matriz de correlação entre as 11 variáveis é dada na tabela 4.2.

Tabela 4.2: Matriz de correlações.

	Investig.	Cient. e Eng.	Investimento	H-M Fundos	M top	H top	Glass Ceiling	H-M PhD	H-M investig.	H-M docentes	H-M CC
Investig.	1.00										
C. e E.	**0.67**	1.00									
Investo	**0.29**	**0.69**	1.00								
H-M F.	0.14	-0.04	0.08	1.00							
M top	-0.15	-0.23	0.12	0.08	1.00						
H top	-0.21	**-0.28**	0.05	-0.10	**0.94**	1.00					
Glass	**-0.31**	**-0.53**	**-0.55**	**-0.27**	**-0.42**	**-0.27**	1.00				
H-M PhD	0.24	**0.47**	**0.59**	0.08	-0.04	-0.10	**-0.34**	1.00			
H-M investig.	0.23	**0.55**	**0.85**	-0.04	0.03	-0.07	**-0.43**	**0.82**	1.00		
H-M doc.	0.00	0.29	**0.64**	-0.01	**0.26**	0.17	**-0.38**	**0.87**	**0.86**	1.00	
H-M CC	**-0.85**	**-0.60**	**-0.27**	-0.21	0.08	0.10	0.23	0.00	-0.06	0.21	1.00

Como se pode observar na tabela 4.2 existem correlações lineares fortes[6] entre os seguintes pares de variáveis:

- *Investigadores* e *Cientistas e Engenheiros*;
- *Investigadores* e *H-M CC*;
- *Cientistas e Engenheiros* e *Investimento*;
- *Investimento* e *H-M investigadores*;
- *M top* e *H top*;

[6]Superiores em valor absoluto a 0.66.

- *H-M PhD* e *H-M investigadores*;
- *H-M PhD* e *H-M docentes*;
- *H-M investigadores* e *H-M docentes*.

Existem ainda relações moderadas[7] entre os seguintes pares de variáveis:

- *Cientistas e Engenheiros* e *Glass Index*;
- *Cientistas e Engenheiros* e *H-M PhD*;
- *Cientistas e Engenheiros* e *H-M investigadores*;
- *Cientistas e Engenheiros* e *H-M CC*;
- *Investimento* e *Glass Index*;
- *Investimento* e *H-M PhD*;
- *Investimento* e *H-M docentes*;
- *M top* e *Glass Index*;
- *Glass Index* e *H-M investigadores*.

Na tabela 4.3 encontra-se informação sobre os valores próprios e importância relativa de cada uma das componentes principais. Tal como aí se constata, as duas primeiras componentes principais - que irão definir a representação bi-dimensional pretendida - explicam, no seu conjunto, 61.89% da variabilidade contida na matriz de dados. Ou seja, 61.89% da variabilidade total dos dados é preservada pela projecção da nuvem de pontos sobre o subespaço bidimensional de IR^{11} gerado pelas duas primeiras componentes principais.

O comportamento da sequência dos valores próprios sugere a retenção de 4 componentes principais, tanto segundo o critério de reter todas as componentes com valores próprios superiores à unidade, como também segundo

[7] valor absoluto entre 0.4 e 0.65

o critério do *scree plot*. Como se pode observar na figura 4.1, é sugerida a retenção de 4 componentes principais pois esta apresenta um ponto de "quebra" na quinta componente, evidenciando uma estabilização dos valores próprios a partir desse ponto.

Tabela 4.3: Valores próprios e percentagem de variância explicada.

Componente	Val. Próprio	% de variância	Acumulada %
1	4.276	38.87	38.87
2	2.531	23.01	61.89
3	1.878	17.07	78.96
4	1.037	9.43	88.39
5	.574	5.22	93.60
6	.387	3.52	97.12
7	.148	1.35	98.47
8	.071	.64	99.11
9	.055	.50	99.61
10	.027	.24	99.85
11	.017	.15	100

Na tabela 4.4 apresentam-se os *loadings*[8] das 4 primeiras componentes principais para as 11 variáveis em estudo. A primeira componente principal, que explica cerca de 40% da variância total, apresenta correlações lineares negativas fortes com as variáveis *Cientistas e Engenheiros*, *Investimento*, *H-M PhD*, *H-M investigadores* e *H-M docentes*, correlação linear negativa moderada com a variável *Investigadores* e correlação linear positiva forte com a variável *Glass Index*. A segunda componente principal, que explica cerca de 20% da variância total, apresenta correlações lineares negativas moderadas com as variáveis *Investigadores* e *Cientistas e Engenheiros* e correlações lineares positivas fortes com as variáveis *M top* e *H top* e moderadas com as variáveis *H-M docentes* e *H-M CC*. A terceira componente principal, que explica cerca de 20% da variância total, apresenta correlações lineares posi-

[8]De sublinhar que estes resultados, obtidos no MatLab, são os simétricos dos obtidos no SPSS para a primeira e terceira colunas, sendo iguais nas duas restantes. Os *Loadings* superiores em valor absoluto a 0.4 estão destacados.

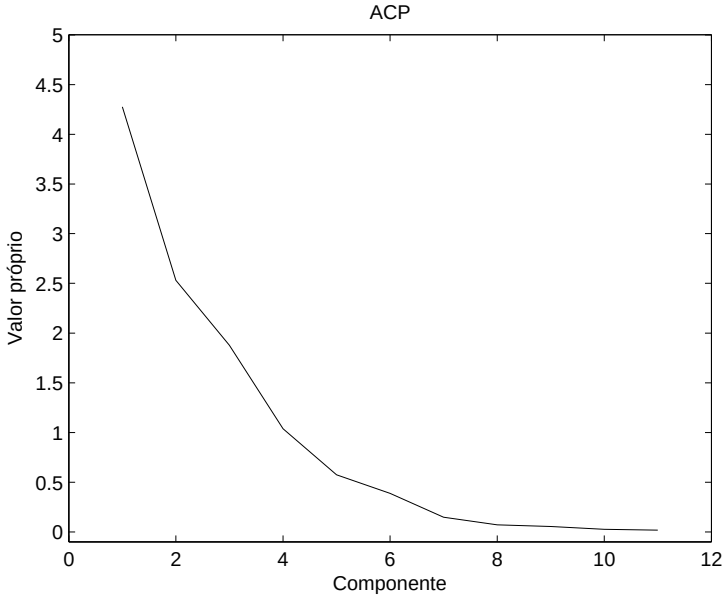

Figura 4.1: *Scree plot* para a ACP.

tivas moderadas com as variáveis *Investigadores*, *M top*, *H top* e negativas moderadas com as variáveis *Glass Index* e *H-M CC*. A quarta componente principal, que explica cerca de 10% da variância total, apresenta correlação linear positiva forte com a variável *H-M Fundos*.

4.2.2 qlPCA

Este conjunto de dados apresenta todas as variáveis na mesma escala de medida e sendo estas contínuas, justifica-se uma abordagem através da qlPCA. Todas as variáveis irão ser submetidas a transformações *spline* de grau 1. A escolha do número de nós interiores está sujeita a existência de um mínimo de $5w$ observações, sendo w a soma do grau do *spline* com o número de nós interiores (ver secção 3.1.4). Assim, para um conjunto de dados com 15 observações pode optar-se por *splines* de grau 1 com no máximo dois nós interiores.

Na tabela 4.5 encontra-se informação sobre os valores próprios e importância relativa das duas primeiras componentes principais, nas soluções

Tabela 4.4: *Loadings* das componentes principais dos dados referentes aos 15 países.

	Componente Principal			
Variável	1	2	3	4
Investig.	**-0.53**	**-0.62**	**0.41**	-0.12
Cient. e Eng.	**-0.80**	**-0.44**	0.09	-0.22
Investimento	**-0.88**	0.10	-0.02	-0.05
H-M Fundos	-0.12	-0.10	0.32	**0.92**
M top	-0.08	**0.77**	**0.61**	-0.05
H top	0.03	**0.76**	**0.57**	-0.24
Glass Ceiling	**0.66**	-0.20	**-0.45**	-0.15
H-M PhD	**-0.80**	0.16	-0.38	0.13
H-M investig.	**-0.89**	0.19	-0.33	-0.03
H-M docentes	**-0.75**	**0.51**	-0.34	0.06
H-M CC	0.38	**0.64**	**-0.58**	0.12
% de var. explicada	38.87	23.01	17.07	9.43

obtidas com a qlPCA de grau 1 com 1 e 2 nós interiores. Verifica-se, neste caso, que as duas primeiras componentes principais, que irão definir a representação bidimensional pretendida, explicam 68.9% e 76.33% da variabilidade contida na matriz de dados transformada, consoante se opta por *splines* de grau 1 com 1 ou com 2 nós interiores, respectivamente.

Tabela 4.5: Valores próprios e percentagem de variância explicada (qlPCA).

	Grau 1, 1 nó		Grau 1, 2 nós	
Componente	Val. Próprio	% de variância	Val. Próprio	% de variância
1	4.582	41.65	4.760	43.27
2	2.998	27.25	3.637	33.06
Total	7.580	68.90	8.396	76.33

Para a construção do *scree plot* associado à qlPCA, são usados os valores próprios da matriz de correlações das variáveis transformadas (conforme parágrafo A da secção 3.1.4). As figuras 4.2 e 4.3 mostram os *scree plots*

para a qlPCA associada a *splines* de grau 1 com 1 nó interior com duas e três componentes retidas, respectivamente.

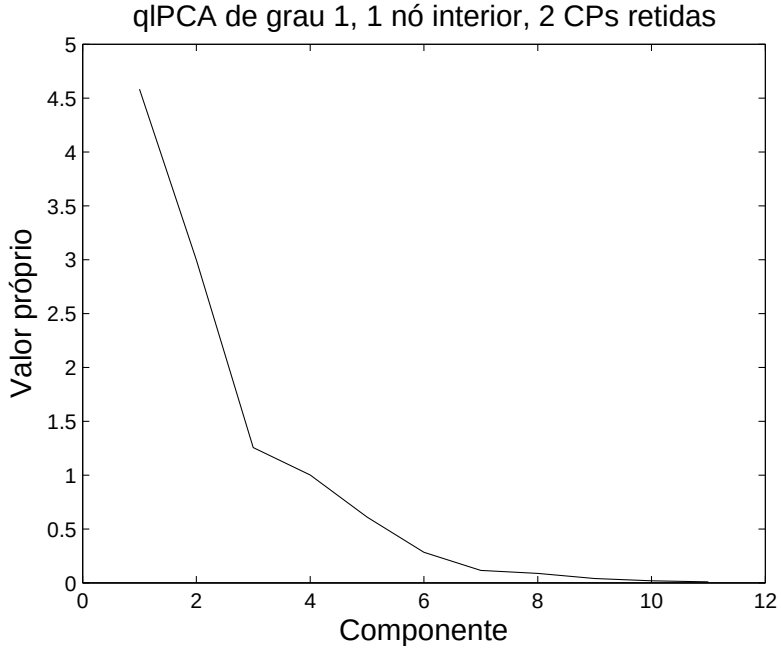

Figura 4.2: *Scree plot* para a qlPCA com um nó interior com duas componentes retidas.

Na solução bi-dimensional (figura 4.2), observa-se que a "quebra" está localizada na terceira componente. Como as soluções da qlPCA não são "encaixadas" (conforme parágrafo B da secção 3.1.4), um *scree plot* para a solução com três componentes principais retidas, pode ser diferente do *scree plot* com duas, como aliás se verifica neste caso.

Na solução tri-dimensional (figura 4.3), observa-se que a "quebra" está localizada na quarta componente. Não observada uma consistência na localização da "quebra" não é possível por esta via decidir o número apropriado de componentes principais a reter, tanto é defensável para a qlPCA de grau 1 com 1 nó interior uma retenção de duas como de três componentes principais na solução.

Nas figuras 4.6 e 4.7 podem observar-se os *scree plots* para a qlPCA asso-

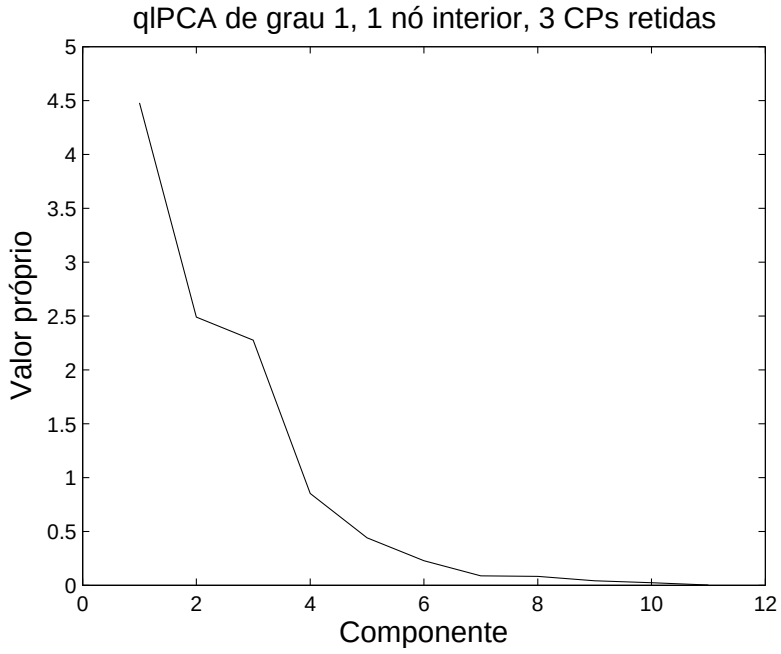

Figura 4.3: *Scree plot* para a qlPCA com um nó interior com três componentes retidas.

ciada a *splines* de grau 1 com 2 nós interiores com duas e três componentes retidas. A não consistência da localização do ponto de "quebra" também não permite decidir o número apropriado de componentes a reter.

Na tabela 4.6 apresentam-se os *loadings*[9] das 11 variáveis transformadas associados à solução com duas componentes principais obtidas através da qlPCA de grau 1 com 1 ou 2 nós interiores.

A primeira componente principal, que explica cerca de 40% da variância total, apresenta correlações semelhantes independentemente de se considerar 1 ou 2 nós interiores. Esta apresenta correlações lineares negativas fortes com as variáveis transformadas *Cientistas e Engenheiros*, *Investimento*, *H-M PhD*, *H-M investigadores* e *H-M docentes*. Esta componente apresenta ainda correlações lineares positivas, no limiar do moderado/forte, com a variável transformada *Glass Ceiling* em ambas as soluções e com a variável transfor-

[9] *Loadings* superiores em valor absoluto a 0.4 estão destacados.

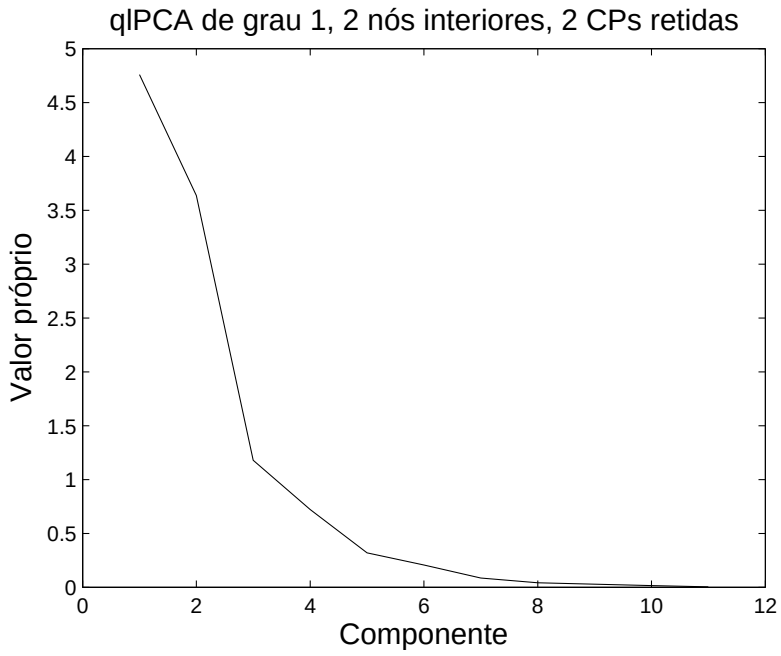

Figura 4.4: *Scree plot* para a qlPCA com 2 nós interiores e duas componentes retidas.

mada *H-M CC*, mas neste caso apenas na solução com 2 nós interiores.

A segunda componente principal, que explica cerca de 30% da variância total, apresenta uma estrutura de correlações com algumas diferenças entre a solução com um ou dois nós interiores. A diferença mais notória reside na variável transformada *H-M Fundos*, na qual se verifica um sentido oposto de associação, apresentando uma correlação linear positiva moderada na solução com um nó interior e uma correlação linear negativa forte na solução com dois nós interiores. As restantes diferenças não se devem ao sentido da correlação mas à sua intensidade. Assim, esta componente apresenta correlações lineares negativas fortes em ambas as soluções com as variáveis transformadas *M top* e *H top*, mas intensidades diferentes no que diz respeito às variáveis *H-M docentes* e *H-M CC*. Apresenta ainda correlações lineares positivas entre o moderado e o forte em ambas as soluções com as variáveis transformadas *Investigadores* e *Cientistas e Engenheiros*. Sublinhe-se, uma vez mais, que

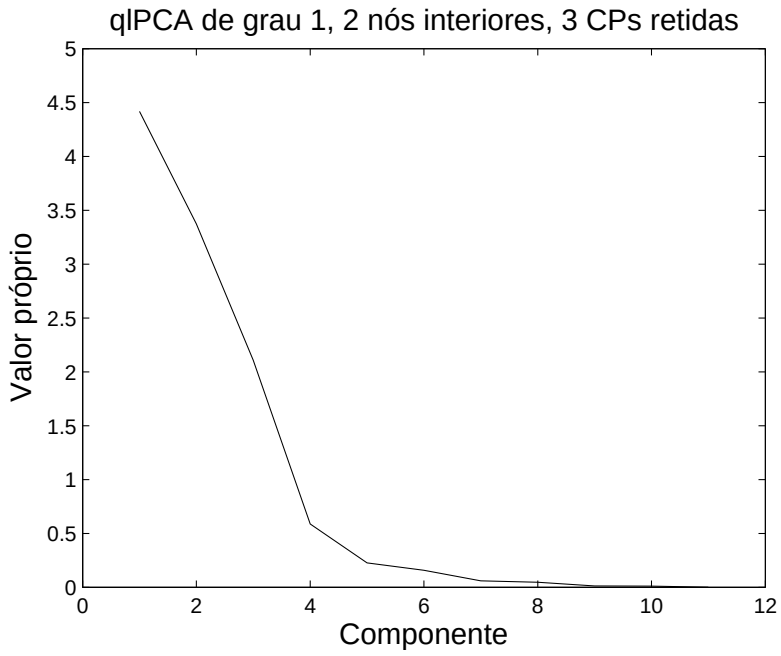

Figura 4.5: *Scree plot* para a qlPCA com 2 nós interiores e três componentes retidas.

todos estes resultados se referem às correlações entre as variáveis transformadas e as componentes retidas.

Na figura 4.6 apresentam-se as 11 transformações óptimas obtidas através da qlPCA linear com 1 nó interior e com duas componentes principais retidas. Os gráficos apresentam no eixo das abcissas a variável original standardizada e no eixo das ordenadas a variável transformada. O único nó interior corresponde ao ponto cuja abcissa é a mediana da variável original standardizada. Os valores das variáveis transformadas são as quantificações óptimas relativas a este conjunto de dados no contexto de *splines* lineares com um nó interior.

Como se pode observar na figura 4.6, com excepção das variáveis *Investigadores* e *Glass Ceiling*, todas as variáveis estão associadas a transformações óptimas claramente distintas da respectiva transformação linear[10]. A tabela

[10]Embora ainda não esteja disponível uma medida do afastamento entre a transformação

Tabela 4.6: *Loadings* das componentes principais dos dados referentes aos 15 países para a qlPCA de grau 1 com 1 (qlPCA1) ou 2 (qlPCA2) nós interiores.

	Componente Principal			
	qlPCA1		qlPCA2	
Variável Transf.	1	2	1	2
Investig.	**-0.53**	**0.70**	**-0.56**	**0.69**
Cient. e Eng.	**-0.72**	**0.55**	**-0.67**	**0.66**
Investimento	**-0.94**	-0.08	**-0.95**	0.06
H-M Fundos	-0.09	**0.49**	-0.01	**-0.89**
M top	-0.23	**-0.75**	-0.29	**-0.83**
H top	-0.31	**-0.75**	-0.16	**-0.80**
Glass Ceiling	**0.67**	-0.07	**0.56**	0.12
H-M PhD	**-0.83**	-0.11	**-0.90**	-0.30
H-M investig.	**-0.92**	-0.19	**-0.94**	-0.23
H-M docentes	**-0.79**	**-0.49**	**-0.84**	-0.31
H-M CC	0.39	**-0.72**	**0.52**	**-0.59**
% de var. explicada	41.65	27.25	43.27	33.06

4.7 mostra que a *comunalidade* é superior ou igual a 0.25 em todas as variáveis transformadas, ou seja, pelo menos 25% da variância de cada variável transformada é explicada pelas duas componentes principais não-lineares retidas. Conjugando o facto dos gráficos da figura 4.6 apresentarem transformações não-lineares com os valores elevados das *comunalidades*, justifica-se a relevância desta variante não-linear da ACP para a análise deste conjunto de dados.

Nas transformações das variáveis *Cientistas e Engenheiros*, *H-M PhD* e *H-M CC*, observa-se na figura 4.6 que um dos segmentos é aproximadamente horizontal, sugerindo que a gama de valores nos referidos segmentos é irrelevante e indistinguível no seio da estrutura multivariada transformada. Assim, por exemplo na variável *Cientistas e Engenheiros*, o facto da sua

óptima e a transformação linear, a mudança de declive à esquerda e à direita do nó interior são um bom indicador desse afastamento. Neste caso, está-se a proceder a uma análise visual dessa mudança.

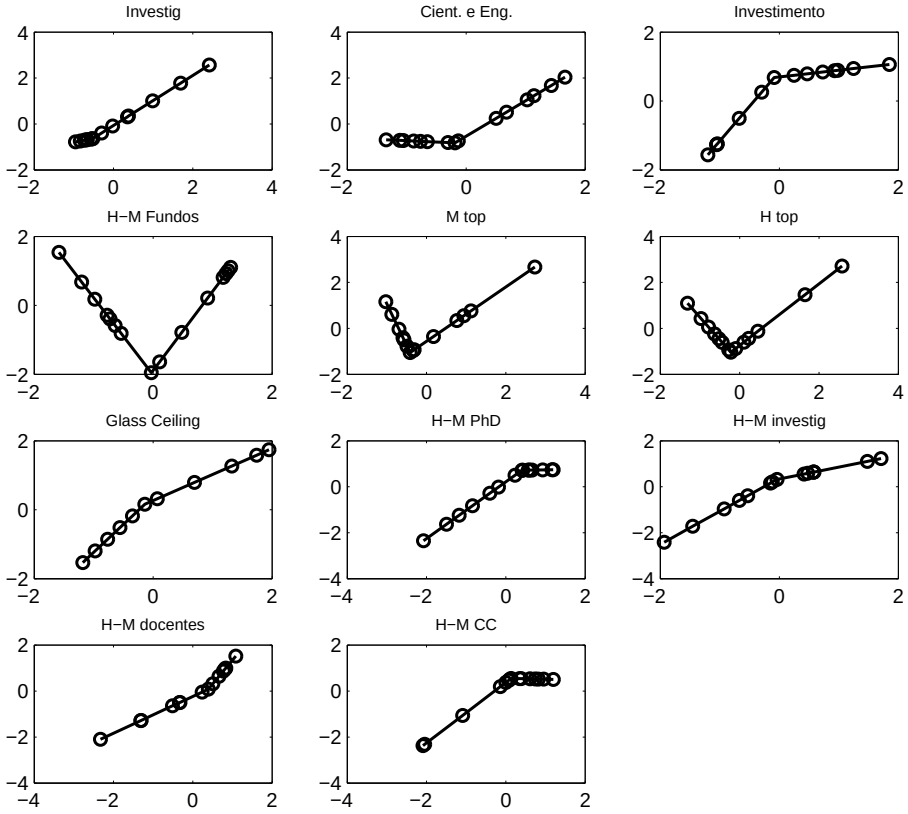

Figura 4.6: Transformações *spline* de grau 1 com 1 nó interior, obtidas com a qlPCA sobre 15 países observados em 11 dimensões.

transformação óptima apresentar um segmento aproximadamente horizontal para valores inferiores à mediana, permite afirmar que dados dois países com os mesmo valores nas restantes variáveis, estes serão indistinguíveis na estrutura transformada se ambos tiverem valores inferiores à mediana da variável *Cientistas e Engenheiros*.

Os gráficos das variáveis *H-M Fundos*, *M top* e *H top* apresentam aproximadamente a forma de "V", indicando que os países com valores mais afastados da mediana nestas variáveis são semelhantes na estrutura multivariada transformada e diferem de alguma forma dos países com valores

Tabela 4.7: *Comunalidade* para cada uma das variáveis transformadas via *splines* lineares com um nó interior.

Variável Transf.	*Comunalidade*
Investig.	0.77
Cient. e Eng.	0.82
Investimento	0.89
H-M Fundos	0.25
M top	0.62
H top	0.66
Glass Ceiling	0.45
H-M PhD	0.70
H-M investig.	0.88
H-M docentes	0.87
H-M CC	0.67

próximos da mediana nestas variáveis. Assim, por exemplo na variável *H-M Fundos*, o facto da sua transformação óptima apresentar a forma de "V", permite afirmar que dados dois países com os mesmo valores nas restantes variáveis, estes serão indistinguíveis na estrutura transformada se ambos tiverem valores sensivelmente à mesma distância da mediana da variável *H-M Fundos*.

Os gráficos das variáveis *Investimento*, *H-M investigadores* e *H-M docentes* apresentam uma notória mudança de declive na mediana, estando, no entanto, associadas a transformações monótonas crescentes.

Conforme referido na secção 3.1.4, parágrafo E, a estrutura em "V" do gráfico da transformação óptima indica que a correlação entre o primeiro segmento da respectiva variável original e as componentes principais não-lineares retidas tem sentido oposto à correlação entre as componentes principais e a respectiva variável transformada. Note-se ainda que a manutenção ou inversão do sentido das correlações por segmento não implica qualquer padrão quanto à intensidade dessas correlações.

Na figura 4.7 apresentam-se as 11 transformações óptimas obtidas através da qlPCA linear com 2 nós interiores e com duas componentes principais re-

tidas. Os gráficos apresentam no eixo das abcissas a variável original estandardizada e no eixo das ordenadas a variável transformada. Os nós interiores correspondem aos pontos cujas abcissas são o percentil 33 e 66 da variável original estandardizada. Os valores das variáveis transformadas são as quantificações óptimas relativas a este conjunto de dados no contexto de *splines* lineares com dois nós interiores.

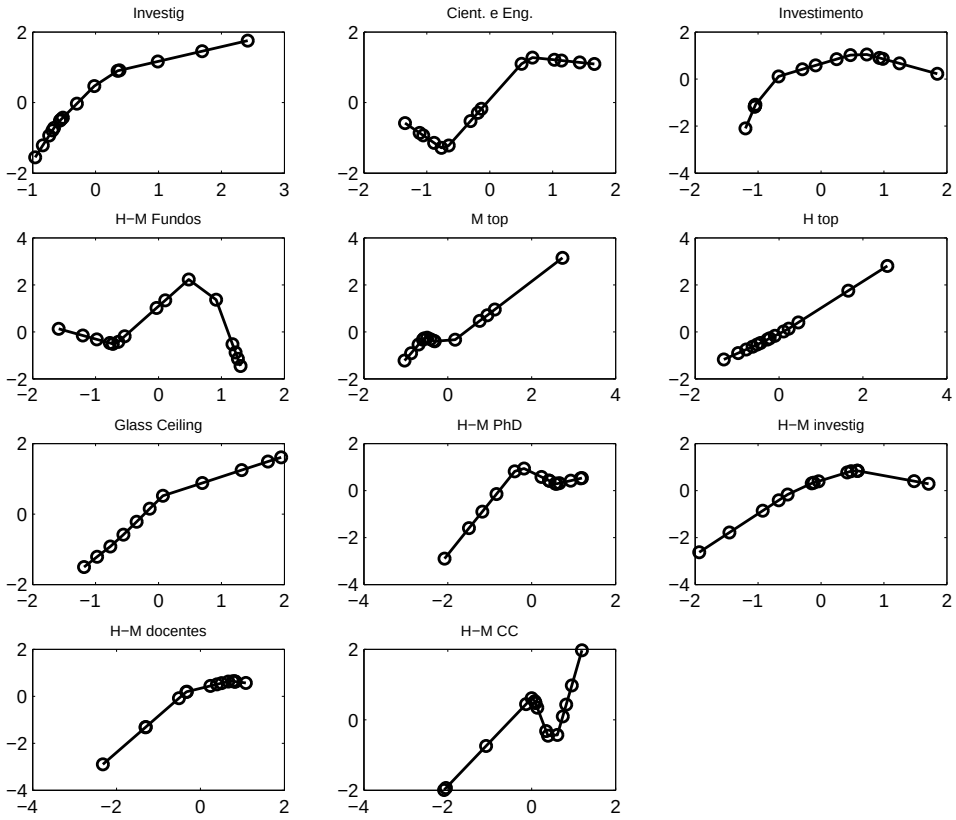

Figura 4.7: Transformações *spline* de grau 1 com 2 nós interiores, obtidas com a qlPCA sobre 15 países observados em 11 dimensões.

Como se pode observar na figura 4.7 a flexibilidade adicional associada à presença de dois nós interiores apenas foi aproveitada nas transformações

óptimas associadas às variáveis *Cientistas e Engenheiros*, *H-M Fundos* e *H--M CC*. A tabela 4.8 mostra que a *comunalidade* é superior ou igual a 0.33 em todas as variáveis transformadas, ou seja, pelo menos 33% da variância de cada variável transformada é explicada pelas duas componentes principais não-lineares retidas. De salientar que a variável transformada *H-M Fundos* sofreu um incremento da *comunalidade* de 0.25 na solução com um nó interior para 0.8 na solução com dois nós interiores. No entanto, as variáveis *Cientistas e Engenheiros* e *H-M CC* mantiveram sensivelmente a mesma *comunalidade*. Conjugando o facto dos gráficos da figura 4.7 apresentarem transformações não-lineares com os valores elevados das *comunalidades*, justifica-se a relevância desta variante não-linear da ACP para a análise deste conjunto de dados.

Tabela 4.8: *Comunalidade* para cada uma das variáveis transformadas via *splines* lineares com dois nós interiores.

Variável Transf.	*Comunalidade*
Investig.	0.79
Cient. e Eng.	0.89
Investimento	0.90
H-M Fundos	0.80
M top	0.77
H top	0.66
Glass Ceiling	0.33
H-M PhD	0.91
H-M investig.	0.93
H-M docentes	0.80
H-M CC	0.62

4.3 Discussão

Os resultados da secção anterior permitem afirmar que, para este conjunto de dados, as variantes não-lineares da ACP permitem um incremento considerável da variância explicada pelas duas primeiras componentes principais.

Assim, passou-se dos 61.89% da ACP para 68.9% na qlPCA linear com um nó interior e para 76.33% na qlPCA linear com dois nós interiores. Estes incrementos são justificados pela presença de várias transformações óptimas não-lineares, conjugadas com valores elevados das respectivas *comunalidades*.

Cada uma destas abordagens está associada a uma representação bi-dimensional distinta, que corresponde à projecção da nuvem de pontos sobre o subespaço bi-dimensional de IR^{11} gerado pelas duas primeiras componentes principais obtidas em cada solução. Para efeitos comparativos das três representações bi-dimensionais recorre-se aos respectivos *biplots*. Cada *biplot* permite uma visualização das relações entre as variáveis, bem como o posicionamento relativo dos países. A qualidade da representação bidimensional dos dados através das duas primeiras componentes principais obtidas pela ACP, pode não ser uma imagem fidedigna da nuvem de pontos original. Sendo esta apenas uma aproximação, as conclusões extraídas de leituras neste tipo de imagens devem sempre envolver a dose necessária de cautela. As variantes não-lineares devem ser exploradas tendo em vista a obtenção de uma imagem eventualmente mais fidedigna.

Nas figuras 4.8, 4.9 e 4.10 apresentam-se os *biplots*[11] das configurações bidimensionais dos dados através das duas primeiras componentes principais associadas à ACP, à qlPCA linear com um e à qlPCA linear com dois nós interiores.

No que diz respeito às variáveis, repare-se na representação gráfica dos *loadings* de cada variável nas duas componentes, indicada pelas coordenadas do vector respectivo e a consequente identificação imediata das variáveis com correlações mais elevadas em cada componente. O feixe de vectores que aponta para sentido positivo de cada eixo representa as variáveis que apresentam uma correlação positiva com a componente principal representada nesse eixo. O feixe de vectores que aponta para sentido negativo de cada

[11]Note-se o sinal de alguns *loadings* foi alterado. Extracto do Help do MatLab para a função biplot: "biplot imposes a sign convention, forcing the element with largest magnitude in each column of coefs to be positive. This flips some of the vectors in coefs to the opposite direction, but often makes the plot easier to read. Interpretation of the plot is unaffected, because changing the sign of a coefficient vector does not change its meaning".

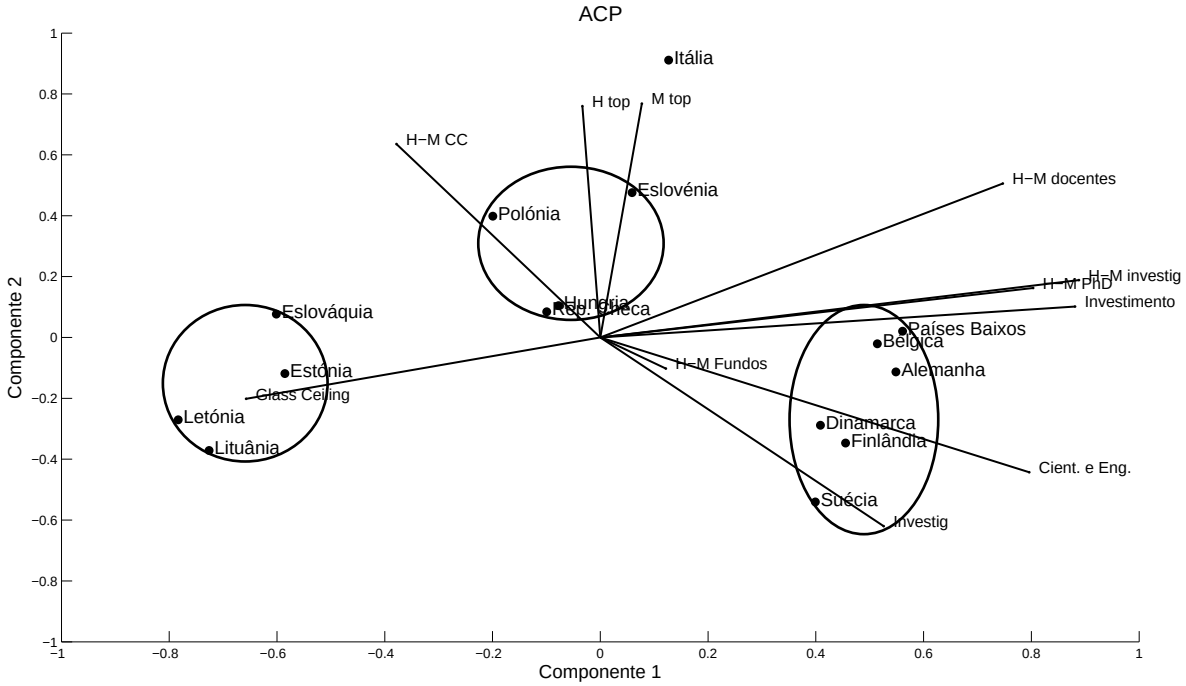

Figura 4.8: *Biplot* associado à ACP sobre 15 países observados em 11 dimensões.

eixo representa as variáveis que apresentam uma correlação negativa com a componente principal representada nesse eixo.

A alteração mais notória reside na variável *H-M fundos*. Repare-se na representação gráfica da *comunalidade* dessa variável nas duas componentes retidas pela ACP, traduzida através de um comprimento diminuto do vector que lhe está associado, comprimento esse que aumenta significativamente na qlPCA com um nó interior e na qlPCA com dois nós interiores. De facto, a *comunalidade* desta variável na ACP tem o valor de 0.0244, sendo que apenas 2.44% da variância da variável *H-M fundos* é explicada pelas duas componentes principais retidas. Já na qlPCA com um nó interior a *comunalidade* da variável transformada tem o valor de 0.25, sendo agora 25% da variância dessa variável explicada pelas duas componentes principais não-lineares retidas. Na qlPCA com dois nós interiores a *comunalidade* da

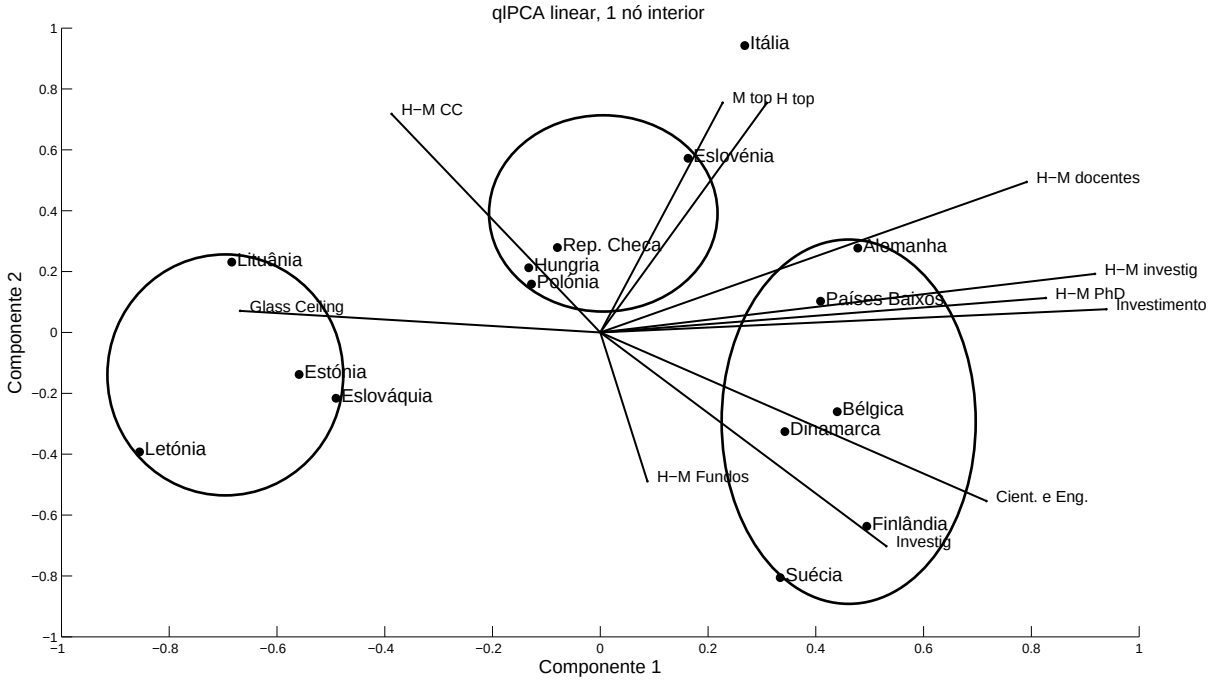

Figura 4.9: *Biplot* associado à qlPCA com 1 nó interior sobre 15 países observados em 11 dimensões.

variável transformada tem o valor de 0.80, sendo que agora 80% da variância dessa variável é explicada pelas duas componentes principais não-lineares retidas. A variável *H-M fundos* e as respectivas variáveis transformadas que lhe estão associadas são também as que sofrem alterações mais substanciais nos valores dos *loadings* como se pode observar pelo posicionamento relativo do vector que a representa nos três *biplots*. As alterações, no posicionamento relativo nos *biplots*, que a variável *H-M fundos* protagoniza, fazem sentido tendo em conta que esta variável foi a que sofreu a transformação não-linear mais intensa.

O feixe de vectores associados às variáveis *H-M docentes*, *H-M Investigadores*, *H-M PhD* e *Investimento* e o feixe de vectores associados às variáveis *Cient. e Eng.* e *Investigadores* mantêm aproximadamente o seu posicionamento nos três *biplots*, traduzindo desta forma a aproximação das trans-

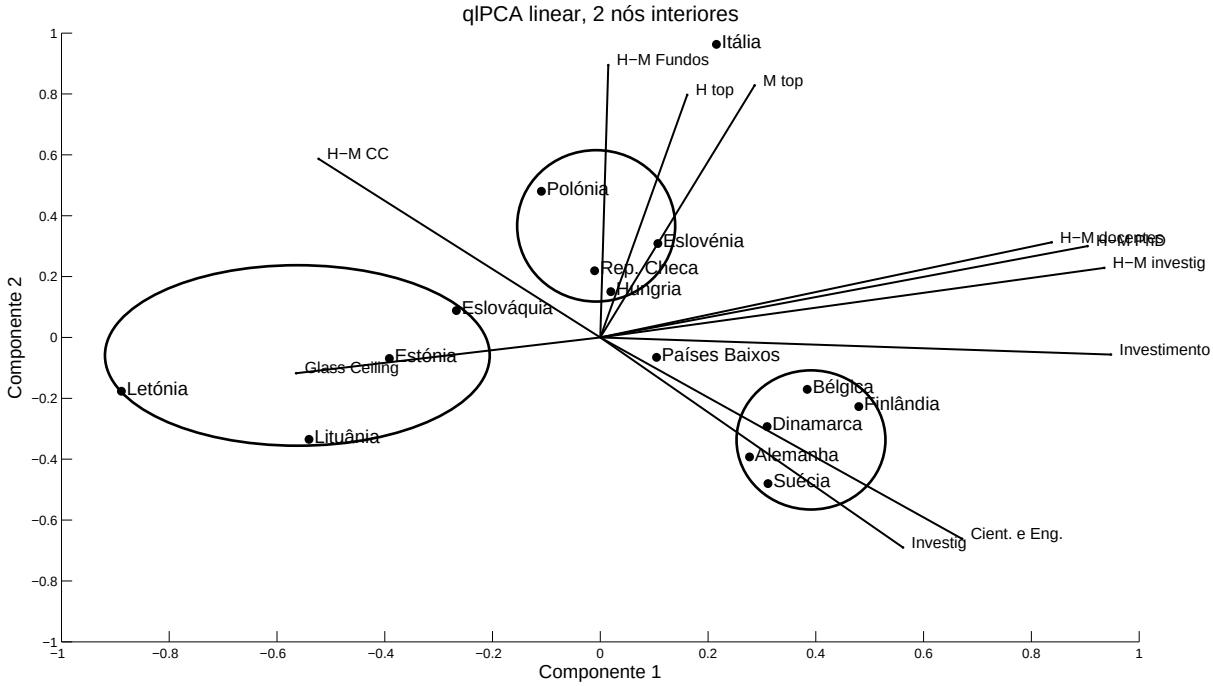

Figura 4.10: *Biplot* associado à qlPCA com 2 nós interiores sobre 15 países observados em 11 dimensões.

formações óptimas a transformações lineares e o fraco incremento das respectivas *comunalidades*.

Os vectores que representam as variáveis *H-M CC* e *Glass Ceiling* e as suas transformadas, mantêm aproximadamente a mesma direcção e o mesmo sentido nos três *biplots* com ligeiras alterações de comprimento. Já o feixe de vectores que representa as variáveis *H top* e *M top* apresenta uma rotação no sentido positivo da primeira componente principal não-linear tanto no caso com 1 nó interior como no caso com 2 nós interiores, bem como um incremento do comprimento dos vectores.

Em qualquer um dos *biplots*, é possível identificar três grupos de países relativamente bem separados na projecção dos dados sobre o plano formado pelas duas primeiras componentes principais. O grupo mais à esquerda, constituído pela Letónia, Lituânia, Estónia e Eslováquia, distingue-se dos

restantes por apresentar valores negativos na primeira componente principal e valores negativos para a segunda componente. O seu posicionamento sugere que são países que apresentam um sistema de Ciência e Tecnologia menos desenvolvido, traduzido por valores reduzidos nas variáveis *Investimento*, *Cientistas e Engenheiros* e *Investigadores*, sendo também países com menores níveis de discriminação em todas essas variáveis com excepção da discrepância entre homens e mulheres nas Comissões Científicas. Note-se que todos estes países apresentam valores negativos na variável *H-M PhD*, que se traduz num número superior de mulheres doutoradas, bem como diferenças de género reduzidas nas variáveis relativas aos investigadores e docentes. Os países deste grupo apresentam ainda valores elevados na variável *Glass Ceiling Index* e uma proporção reduzida tanto de homens como mulheres no topo da carreira.

O grupo central, constituído pela Eslovénia, Polónia, Hungria e República Checa, distingue-se dos restantes por apresentar valores mais próximos da origem para a primeira componente principal e valores positivos para a segunda componente. Este posicionamento sugere uma proporção relativamente elevada tanto de homens como mulheres no topo da carreira e valores médios em termos de desenvolvimento do sistema de Ciência e Tecnologia e dos valores de discriminação. Os três *biplots* sugerem a não inclusão da Itália neste grupo, pois apesar de apresentar valor próximo da origem na primeira componente principal, apresenta um valor na segunda componente principal claramente superior aos países do grupo central.

O grupo mais à direita, constituído no *biplot* da ACP e no *biplot* da qlPCA com um nó interior pela Suécia, Dinamarca, Finlândia, Bélgica, Alemanha e Países Baixos, distingue-se dos restantes por apresentar valores positivos na primeira componente principal e valores negativos para a segunda componente. O seu posicionamento sugere que são países que apresentam os sistema de Ciência e Tecnologia mais desenvolvido, traduzido por valores elevados nas variáveis *Investimento*, *Cientistas e Engenheiros* e *Investigadores*, sendo também países com maiores níveis de discriminação em todas essas variáveis com excepção da discrepância entre homens e mulheres nas Comissões Científicas. O *biplot* da qlPCA com dois nós interiores sugere a

saída dos Países Baixos deste grupo.

Com a excepção dos Países Baixos a constituição dos grupos é mantida independentemente da solução que se está a considerar. No entanto, o posicionamento relativo dos países intra-grupos sofre alterações. No grupo central e no grupo mais à direita essas alterações são devidas principalmente às alterações das projecções na segunda componente principal. No grupo mais à esquerda as alterações de posicionamento tanto se devem às diferentes projecções na primeira como na segunda componente principal.

Conclusão

Nesta tese foram apresentados desenvolvimentos originais no domínio das variantes não-lineares da Análise em Componentes Principais (ACP). Um novo algoritmo, designado *quasi-linear PCA* (qlPCA), foi introduzido e implementado em MatLab. Este teve por base o *sistema Gifi* e transformações *spline*, tendo sido especialmente concebido para estruturas não-lineares de variáveis contínuas.

A ideia fundamental da ACP tradicional é projectar os dados originais, que incluem ruído e variáveis redundantes, num espaço latente de dimensão inferior, com o objectivo de revelar a verdadeira dimensionalidade dos dados. Este objectivo foi mantido na qlPCA, sendo a interpretação geométrica a mesma da ACP tradicional, mas traduzida em termos das variáveis transformadas associadas à designada *quantificação óptima*. Um processo de *Mínimos Quadrados Alternados* (MQL) foi usado para minimizar uma *função perda*, tendo como resultado as *quantificações óptimas* e as componentes principais não-lineares. A perda a ser minimizada é a perda de informação inerente à representação das variáveis por um número reduzido de componentes principais, tendo o MQL sido usado até à convergência, alternando entre a actualização das quantificações tendo os *object scores* fixos e a actualização dos *object scores* tendo as quantificações fixas.

Usualmente a ACP tradicional é obtida através de uma decomposição em valores singulares da matriz de dados estandardizada, ou através de uma decomposição em valores e vectores próprios da matriz de correlação. No entanto, os mesmos resultados podem ser obtidos através de um processo iterativo no qual uma *função perda* é minimizada. Nas variantes da ACP as-

sociadas ao *sistema Gifi*, nas quais se inclui a qlPCA, o processo iterativo para minimização da perda permite adicionalmente transformações não-lineares das variáveis originais. Estas variantes são por isso uma generalização da ACP tradicional, ao permitirem, no mesmo modelo, a substituição da matriz original dos dados observados pela matriz das variáveis transformadas.

No capítulo 1 foram introduzidas detalhadamente várias formulações para a *função perda*, partindo da associada à ACP tradicional até uma versão suficientemente abrangente para incluir como casos particulares todas as vertentes abordadas nesta tese. Introduziram-se também as transformações *spline*, que não são mais do que funções seccionalmente polinomiais, com ligações suaves nos pontos de junção. Foi usado um resultado fundamental que refere que o conjunto dos *splines* de determinado grau, com pontos de junção fixos constitui um espaço linear de funções, sendo conhecidos os elementos da base. No capítulo 3, mostrou-se que na qlPCA, determinar a *quantificação óptima* de uma variável é equivalente a determinar a combinação linear óptima dos elementos da base e assim obter o *spline* óptimo a aplicar à variável.

Dois algoritmos que emergiram do *sistema Gifi*, e que por isso também recorrem a um processo de *Mínimos Quadrados Alternados* para minimizar determinadas funções perda, foram essenciais para este trabalho: a HOMALS (HOMogeneity analysis by Alternating Least Squares) e a CATPCA (CATegorical Principal Components Analysis). No capítulo 2 foi apresentada um breve revisão da sua fundamentação teórica e respectiva implementação. Há em ambos um nítido investimento na quantificação de variáveis categoriais como forma de estender a ACP tradicional para estruturas com este tipo de dados e, na CATPCA, também para ambientes com variáveis das várias escalas de medida.

A definição da classe de transformações admissíveis para cada variável a incluir na função perda assume um papel preponderante neste tipo de análises. Esta definição deve resultar da dialéctica entre a qualidade do ajustamento e o respeito pela natureza das variáveis originais, traduzido pelas restrições que se impõem nas transformações permitidas. A própria ACP tradicional pode ser considerada um caso particular destas abordagens, em que determinada função perda é minimizada sendo apenas permitidas trans-

formações lineares para todas as variáveis. A transformação linear é a mais rígida pois apenas permite que as distâncias relativas entre os valores originais sejam alteradas de forma proporcional, sendo a constante de proporcionalidade a mesma em todo o domínio da variável. As generalizações da ACP referidas nesta tese (HOMALS, CATPCA e qlPCA) abordam o problema da não-linearidade e do tratamento de variáveis categoriais relaxando as restrições lineares das transformações.

De todas as técnicas provenientes do *sistema Gifi*, a HOMALS é a mais potente devido à flexibilidade permitida para as quantificações. A minimização da função perda da HOMALS, estando associada à classe das transformações seccionalmente constantes, trabalha todas as variáveis tendo apenas em consideração quais os objectos que estão em cada categoria. Como se mostrou no capítulo 2, até mesmo a ordem das categorias de uma variável pode ser alterada nas quantificações óptimas. Assim, a HOMALS está especialmente indicada para a análise de dados provenientes de variáveis de natureza nominal. Como foi referido, esta pode também ser usada para casos em que, existindo também variáveis de natureza ordinal, o investigador decida pela eventual flexibilização da ordem inicial tendo em vista a maior qualidade do ajustamento do modelo.

A CATPCA surge, histórica e tecnicamente, como uma generalização da HOMALS por forma a permitir alargar a classe de transformações admissíveis e restringir a flexibilidade permitida para as quantificações na HOMALS. Como foi mostrado no capítulo 2, a definição das classes de transformações admissíveis é, na CATPCA, definida variável a variável. Estão disponíveis várias opções, definidas entre dois extremos, o mais rígido - transformações lineares, como na ACP tradicional - e o mais flexível - transformações seccionalmente constantes, como na HOMALS. Entre as opções disponíveis estão as transformações *spline*, sendo a CATPCA o único algoritmo associado ao *sistema Gifi* que as disponibiliza. No capítulo 2 foi introduzida a função perda da CATPCA, como uma generalização da HOMALS, tendo sido analisada a implementação das várias classes de transformações disponíveis.

No entanto, mostrou-se que a CATPCA, algoritmo originalmente concebido para variáveis categoriais (ordinais e nominais), necessita, *a priori*, de

um processo de discretização quando aplicada a variáveis contínuas, sendo os *splines* aplicados *a posteriori*. Assim, esta tese apresenta-se como uma resposta para o seguinte problema:

> Tendo estas variantes da ACP sido desenvolvidas para variáveis categoriais, como adaptá-las de modo a serem consideradas como uma abordagem não-linear em contexto de variáveis contínuas?

O capítulo 3 introduziu a fundamentação teórica da qlPCA e a sua implementação em MatLab, sendo a resposta para o problema enunciado. A qlPCA opera exclusivamente num contexto de variáveis contínuas, implementando a base dos *splines* directamente através duma matriz pseudo-indicatriz, evitando assim o processo de discretização e preservando todas as características das variáveis contínuas (ordem e distâncias relativas).

O adjectivo *quasi* na qlPCA pretende sublinhar as vantagens que advêm do uso de *splines* lineares (transformações seccionalmente lineares). Uma transformação linear é um caso particular de um *spline* linear sem partições, ou seja, sem nós interiores. Mantendo o *spline* linear e incrementando o número de nós interiores é possível aumentar a flexibilidade das *quantificações*. As distâncias relativas entre os valores originais são alteradas de forma proporcional, mas a constante de proporcionalidade é variável por segmentos. Esta característica permitiu uma extensão do conceito de *loading*, que se propôs designar *piecewise loadings*, sendo definidos através das correlações "segmentadas" entre as componentes principais não-lineares e as variáveis originais. Foram ainda apresentadas as principais propriedades da qlPCA, nomeadamente em termos do sumário do modelo, escolha do número de componentes a reter, projecção de novas observações no espaço das componentes principais não-lineares e reconstrução dos valores originais através das projecções.

No final do capítulo 3 foi apresentado um exemplo para ilustrar o potencial do algoritmo proposto. A qlPCA em comparação com a ACP, revelou um desempenho notável na captação de uma estrutura não-linear conhecida. Por outro lado, verificou-se que, nesse exemplo, a qlPCA apresentou

transformações *spline* óptimas que se aproximavam das transformações linearizantes conhecidas. No capítulo 4 foi apresentado um exemplo com base em dados reais onde também foi possível comparar resultados entre a qlPCA e a ACP.

Esta tese para além de responder a algumas questões, proporcionou o levantamento de outras que se perspectivam como trabalho futuro.

Uma delas está relacionada com a comparação entre a ACP tradicional e a qlPCA no que diz respeito à qualidade do ajustamento e à necessidade de uma abordagem não-linear. Como foi referido uma medida baseada na percentagem de variância explicada pelas componentes retidas pode ser enganadora, pois na ACP tradicional esta refere-se à variância das variáveis originais e na qlPCA à variância das variáveis transformadas. Seria importante encontrar outras medidas para este propósito. Relativamente à necessidade de uma abordagem não-linear, foram apresentados resultados para cada variável, no entanto, seria relevante dispor de um indicador global.

Outra questão está relacionada com a relação entre as componentes principais não-lineares e as variáveis originais. Apresentaram-se alguns resultados, nomeadamente os designados *piecewise loadings*, mas estes apresentam informação por segmentos e ainda não é claro como adaptá-los por forma a que "o todo seja a soma das partes". Por este motivo, optou-se por não referir os *piecewise loadings* nas aplicações apresentadas.

Uma terceira questão está relacionada com a implementação da qlPCA. Por motivos pessoais, foi usado o MatLab na implementação. No entanto, dado o crescimento notável de utilizadores do *software*/linguagem R na comunidade científica associada à Estatística e Análise de Dados, pretende-se traduzir o código para R. Associado ao código em R, pretende-se desenvolver a respectiva documentação de apoio e submeter este algoritmo para uma das comunidades R, partilhando-o desta forma a nível mundial.

Dum ponto de vista mais teórico, pretende-se estudar os designados *P--splines* (*Penalized splines*, que usam um elevado número de nós interiores, penalizando posteriormente os nós associados a alterações bruscas) e analisar a sua incorporação na qlPCA. Finalmente, ainda há a questão da localização dos nós interiores, sob a qual não é realizada qualquer optimização. Seria

importante realizar um estudo para analisar o efeito de diferentes localizações na solução da qlPCA.

Apêndice A

Apêndices

A.1 Dados dos países da União Europeia

País	Investig.	Cient. e Eng.	Investimento	H-M Fundos	M top	H top	Glass Ceiling	H-M PhD	H-M investig.	H-M docentes	H-M CC
Bélgica	10.20	7.70	112.56	-2.48	3.60	17.80	1.70	28.91	43.47	34.62	71.19
Rep. Checa	6.25	3.30	60.29	1.61	3.50	15.60	3.10	29.50	43.32	31.99	76.36
Dinamarca	12.79	6.00	99.27	5.62	3.90	14.80	2.301	19.03	43.27	36.44	29.50
Estónia	8.50	2.90	21.54	-.82	4.50	21.101	2.60	-16.81	13.72	1.58	56.20
Finlândia	18.87	7.30	88.04	-1.70	8.80	22.80	1.801	2.56	40.16	18.10	4.92
Alemanha	10.08	5.70	125.46	5.94	2.40	9.80	1.901	24.28	61.12	41.69	65.72
Hungria	7.32	4.30	40.93	1.09	6.50	20.20	2.30	14.15	29.70	27.49	59.46
Itália	4.50	3.50	139.00	3.01	16.50	38.30	1.90	-1.75	41.40	37.50	74.34
Letónia	5.02	3.70	14.12	5.80	4.60	17.40	2.20	-34.38	-6.15	-15.46	52.86
Lituânia	6.37	4.50	21.76	-1.52	1.80	12.30	3.20	-23.02	3.32	1.70	64.66
Países Baixos	5.45	6.80	169.72	-3.31	3.10	13.70	2.00	17.72	65.65	37.29	58.45
Polónia	5.67	3.00	22.23	4.64	9.50	21.10	1.80	10.68	21.50	30.11	85.61
Eslováquia	6.18	2.50	21.81	-4.75	4.20	18.70	2.90	-10.25	18.76	17.88	79.62
Eslovénia	5.78	4.60	71.04	-1.20	10.20	31.50	2.201	17.17	31.22	37.10	58.02
Suécia	15.79	6.60	122.68	6.10	4.201	16.20	2.10	14.45	29.07	14.97	6.02

A.2 Dados discretizados dos países da União Europeia

País	Investig.	Cient. e Eng.	Investimento	H-M Fundos	M top	H top	Glass Ceiling	H-M PhD	H-M investig.	H-M docentes	H-M CC
Bélgica	12	15	11	3	5	8	1	14	13	10	11
Rep. Checa	7	4	7	9	4	5	14	15	12	9	13
Dinamarca	13	11	10	12	6	4	11	12	11	11	3
Estónia	10	2	2	7	9	11	12	3	3	2	5
Finlândia	15	14	9	4	12	13	3	6	9	6	1
Alemanha	11	10	13	14	2	1	5	13	14	15	10
Hungria	9	7	6	8	11	10	10	8	7	7	8
Itália	1	5	14	10	15	15	4	5	10	14	12
Letónia	2	6	1	13	10	7	9	1	1	1	4
Lituânia	8	8	3	5	1	2	15	2	2	3	9
Países Baixos	3	13	15	2	3	3	6	11	15	13	7
Polónia	4	3	5	11	13	12	2	7	5	8	15
Eslováquia	6	1	4	1	8	9	13	4	4	5	14
Eslovénia	5	9	8	6	14	14	8	10	8	12	6
Suécia	14	12	12	15	7	6	7	9	6	4	2

Referências bibliográficas

[1] Z. Asan and M. Greenacre. Biplots of fuzzy coded data. Economics Working Papers 1077, Department of Economics and Business, Universitat Pompeu Fabra, revised Jan 2011. http://www.econ.upf.edu/en/research/onepaper.php?id=1077. {citado na p. 61}

[2] P. Bekker and J. De Leeuw. Relations between variants of non-linear principal components analysis. In J. Van Rijckevorsel and J. De Leeuw, editors, *Component and correspondence analysis*, pages 1–31. Wiley, 1988. {citado nas pp. 2, 17, 18 e 25}

[3] H. Carvalho. *Variáveis qualitativas na análise sociológica: exploração de métodos multidimensionais*. PhD thesis, ISCTE, 1998. {citado na p. 18}

[4] European Commission. She figures 2006. Acedido em Abril, 2012, de `http://www.kif.nbi.dk/She_Figures_2006.pdf`, 2006. {citado na p. 81}

[5] C. De Boor. *A Practical Guide to Splines*. Springer, 1978. {citado nas pp. 8 e 9}

[6] J. De Leeuw. Here's looking at multivariables. In *Keynote address, presented at the Conference on Visualization of Categorical Data, Koln, May 17-19*, 1995. {citado na p. 3}

[7] J. De Leeuw. Nonlinear principal component analysis and related techniques. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*, pages 107–134. Chapman & Hall/CRC, Boca Raton, FL, 2006. {citado na p. 2}

[8] J. De Leeuw and P. Mair. Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, 31(4):1–21, 2009. {citado na p. 3}

[9] J. De Leeuw and J. Van Rijckevorsel. Beyond homogeneity analysis. In J. Van Rijckevorsel and J. De Leeuw, editors, *Component and correspondence analysis*, pages 55–80. Wiley, 1988. {citado nas pp. 4, 15 e 16}

[10] B. Escofier and J. Pagès. *Analyse factorielles simples et multiples: objectifs, méthodes et interprétation*. Dunod, 1998. {citado nas pp. 17 e 18}

[11] A. Gifi. PRINCALS. Report UG-85-03, Department of Data Theory, University of Leiden, 1985. {citado na p. 20}

[12] A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, 1991. {citado nas pp. 2, 15, 16, 25, 48 e 74}

[13] T. Hastie. Principal curves and surfaces. Technical report 11, Department of Statistics, Stanford University, 1984. {citado nas pp. 5 e 50}

[14] Research Group ISSP. *International Social Survey Programme 1995: National Identity I (ISSP 1995)*. GESIS Data Archive, Cologne. ZA2880 Data file Version 1.0.0, doi:10.4232/1.2880, 1998. {citado na p. 28}

[15] J. Jackson. *A User's Guide to Principal Components*. Wiley, 1991. {citado na p. 63}

[16] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. {citado nas pp. 5 e 50}

[17] U. Kruger, J. Zhang, and L. Xie. Developments and applications of nonlinear principal component analysis - a review. In A. Gorban, B. Kégl, D. Wunsch, and A. Zinovyev, editors, *Principal Manifolds for Data Visualization and Dimension Reduction*, pages 1–43. Springer, 2008. {citado na p. 50}

[18] W. Krzanowski and F. Marriott. *Multivariate Analysis Part I - Distributions, Ordination and Inference*. Edward Arnold, 1994. {citado na p. 49}

[19] N. Lavado. Análise em Componentes Principais Não-Linear. Master's thesis, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, 2004. {citado nas pp. 6, 17, 18 e 25}

[20] N. Lavado and T. Calapez. Um enquadramento das variantes não-lineares da ACP via transformações spline. In C. Braumann, P. Infante, M. Oliveira, R. Alpízar-Jar, and F. Rosado, editors, *Estatística Jubilar. Actas do XII Congresso da Sociedade Portuguesa de Estatística*, pages 391–402, 2005. {citado nas pp. 6 e 7}

[21] N. Lavado and T. Calapez. Matrizes de codificação difusa na ACP não-linear. In I. Oliveira, E. Correia, F. Ferreira, S. Dias, and C. Braumann, editors, *Estatística Arte de Explicar o Acaso. Actas do XVI Congresso Anual da Sociedade Portuguesa de Estatística*, pages 353–364, 2009. {citado na p. 6}

[22] N. Lavado and T. Calapez. *Quasi − Linear* PCA: Low order spline's approach to non-linear principal components. In *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2011, WCE 2011, 6-8 July, 2011, London, U.K.*, pages 360–364, 2011a. {citado nas pp. 6 e 60}

[23] N. Lavado and T. Calapez. Principal components analysis with spline optimal transformations for continuous data. *IAENG Int. J. App. Math.*, 41(4):367–375, 2011b. {citado nas pp. 6, 19, 48 e 60}

[24] L. Lebart. Validation techniques in multiple correspondence analysis,. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*, pages 179–195. Chapman & Hall/CRC, Boca Raton, FL, 2006. {citado na p. 60}

[25] M. Linting and A. Kooij. Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment*, 94:1:12–25, 2012. {citado nas pp. 3, 4, 5, 34, 60, 61 e 62}

[26] M. Linting, J. Meulman, P. Groenen, and A. Kooij. Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12(3):336–358, 2007. {citado nas pp. 3, 4, 63 e 64}

[27] M. Linting, J. Meulman, P. Groenen, and A. Kooij. Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, 12(3):359–379, 2007. {citado nas pp. 3 e 60}

[28] P. Mair and J. De Leeuw. A general framework for multivariate analysis with optimal scaling: The r package aspect. *Journal of Statistical Software*, 32(9):1–23, 2010. {citado nas pp. 3 e 4}

[29] J. Meulman and W. Heiser. *IBM SPSS Categories 20*. IBM Corp., 2011. {citado na p. 2}

[30] J. Meulman, A. Kooij, and W. Heiser. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In Kaplan D., editor, *The Sage Handbook of Quantitative Methodology for the Social Sciences*, pages 49–70. Sage Publications, 2004. {citado nas pp. 2, 3, 15, 16, 17, 50 e 60}

[31] L. Oliveira and H. Carvalho. The segmentation of the S&T space and gender discrimination in Europe. In P. Katarina, L. Oliveira, and S. Hemlin, editors, *Women in Science and Technology*, pages 27–51. Institute for Social Research, Zagreb, Sociology of Science and Technology Network of the European Sociological Association, 2009. {citado na p. 80}

[32] B. Schoolkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. {citado nas pp. 5, 49 e 50}

[33] L. Schumaker. *Spline Functions: Basic Theory*. Wiley, 1981. {citado nas pp. 7, 8, 9 e 10}

[34] M. Tenenhaus and F. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119, 1985. {citado na p. 17}

[35] J. Van Rijckevorsel. Fuzzy coding and b-splines. In J. Van Rijckevorsel and J. De Leeuw, editors, *Component and correspondence analysis*, pages 33–54. Wiley, 1988. {citado na p. 4}

[36] S. Winsberg and J. Ramsay. Monotone spline transformations for dimension reduction. *Psychometrika*, 48:575–595, 1983. {citado nas pp. 11, 54, 55, 59 e 74}

Índice remissivo