CARLOS JOSÉ DOS REIS

PROPOSIÇÃO DE DOIS NOVOS MÉTODOS PARA ANÁLISE DE COMPONENTES PRINCIPAIS

LAVRAS – MG

2020

CARLOS JOSÉ DOS REIS

PROPOSIÇÃO DE DOIS NOVOS MÉTODOS PARA ANÁLISE DE COMPONENTES PRINCIPAIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Lucas Monteiro Chaves
Orientador

Prof. Dr. Devanil Jaques de Souza
Coorientador

LAVRAS – MG

2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Reis, Carlos José dos
Proposição de dois novos métodos para análise de componentes principais / Carlos José dos Reis.
2020.
187 p. : il.

Orientador: Lucas Monteiro Chaves.
Coorientador: Devanil Jaques de Souza.
Tese (Doutorado)–Universidade Federal de Lavras, 2020.

Bibliografia.
1. PCA. 2. SPCA. 3. Agrupamento. 4. Esparsidade. I. Universidade Federal de Lavras. II. Título.

CARLOS JOSÉ DOS REIS

PROPOSIÇÃO DE DOIS NOVOS MÉTODOS PARA ANÁLISE DE COMPONENTES PRINCIPAIS
PROPOSITION OF TWO NEW METHODS FOR PRINCIPAL COMPONENT ANALYSIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 04 de Agosto de 2020.

Dr. Denismar Alves Nogueira	UNIFAL-MG
Dr. Daniel Furtado Ferreira	UFLA
Dr. Paulo Henrique Sales Guimarães	UFLA

Prof. Dr. Lucas Monteiro Chaves
Orientador

Prof. Dr. Devanil Jaques de Souza
Coorientador

LAVRAS – MG
2020

À minha prima Daniela Cândida Maranesi (in memorian).

A todos que, por necessidade ou inclinação, desejam conhecer mais sobre esse fascinante método da Estatística Multivarida, a análise de componentes principais.

Dedico

AGRADECIMENTOS

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística, pela oportunidade concedida para realização do doutorado.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão das bolsas de estudo, que me forneceram as condições necessárias para que eu pudesse dedicar-me ao curso.

Ao povo brasileiro. Quero expressar nessas linhas a minha profunda gratidão aos trabalhadores de nosso país. Foi graças a esses meus compatriotas que obtive condições financeiras para dedicar-me integralmente aos meus estudos nas Universidades Federais de Alfenas, de Recife e de Lavras, ao longo de toda a minha formação acadêmica.

Aos professores do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária pelos ensinamentos transmitidos para minha formação acadêmica e pessoal. Agradeço o respeito e a harmoniosa convivência nesses quatro anos de doutorado.

A todos os funcionários dos Departamentos de Estatística e de Ciências Exatas. Sou muito grato pelo carinho, pela amizade e pela presteza com que sempre me atenderam. Muito obrigado!

Aos professores Daniel Furtado Ferreira, Denismar Alves Nogueira, Maria do Carmo Pacheco de Toledo Costa e Paulo Henrique Sales Guimarães, que contribuíram de forma valiosa com suas sugestões e comentários para que se chegasse ao documento final deste trabalho.

Aos meus orientadores, Lucas Monteiro Chaves e Devanil Jaques de Souza, deixo aqui registrado o meu eterno carinho. Sinto-me ao final desse trabalho muito feliz e honrado em ter sido orientado por vocês. A disposição de vocês em ensinar e o respeito que dedicam aos seus alunos e orientados são exemplos que desejo praticar. Após seis anos, posso afirmar que saio mais edificado por tudo aquilo que recebi de vocês. Reservo ao final meus sinceros agradecimentos pelos ensinamentos e pelo respeito com que sempre me trataram.

Aos meus colegas pelos bons momentos que passamos juntos. Em especial, agradecimentos são devidos aos meus colegas de doutorado, nas Universidades Federais de Lavras e de Recife. Alguns colegas tornaram-se amigos chegados.

Gostaria de expressar meus agradecimentos também a Jéssica Gracielle Silva e a Laerte dias de Carvalho pela boa convivência e pelo apoio em diferentes momentos ao longo do doutorado.

A Kathylce Jaqueline Vital Vieira, um presente que me foi dado por Deus. O seu apoio e a sua presença foram reconfortantes. Muito obrigado por se fazer presente, por sua paciência e por sua compreensão nos momentos em que precisei me ausentar para dedicar-me ao doutorado.

Aos meus pais, José de Souza Reis e Selma Batista dos Reis que, não somente formam a base desses agradecimentos, mas também formam a base de minha vida. Amo muito vocês!

A Deus, pela oportunidade que me foi concedida. O amor, a graça e a misericórdia do Senhor seguiram-me em todos esses anos. Ao ouvir em meu coração a Sua palavra, senti minhas forças recobrarem, me senti disposto para continuar. “Sê forte e corajoso. Não temas

e nem te espantes, porque o Senhor, teu Deus, é contigo por onde quer que andares" (JOSUÉ 1.9).

De antemão, quero também expressar os meus agradecimentos a todos aqueles que venham a dedicar o seu tempo a leitura desse trabalho. Que o presente trabalho possa produzir bons frutos, assim como o fez comigo ao longo desses anos de dedicação. Que o presente trabalho seja uma nascente a todos que venham a se dedicar a ele.

Encerro, com um trecho do livro *A Vida Intelectual*, de Antonin-Gilbert Sertillanges:

> "Escrevendo, é preciso publicar. Desde que bons juízes achem está capacitado para isso e que você mesmo sinta a aptidão para o voo. O pássaro sabe bem quando pode afrontar o espaço. Sua mãe o sabe com maior segurança. Apoiado em si mesmo e em uma sábia maternidade espiritual, voe logo que puder. O contato com o público vai obrigá-lo a produzir melhor. Os elogios merecidos o encorajarão. As críticas exercerão seu controle. O progresso lhe será por assim dizer imposto, em lugar da estagnação que poderia resultar de um perpétuo silêncio. A paternidade intelectual é uma semeadura de bens. Toda obra é uma nascente."

"Quando todo fato, todo fenômeno presente ou passado desse universo, todas as fases da vida presente ou passada, tiverem sido examinadas, classificadas e coordenadas com o resto, a missão da ciência será concluída. O que é isso senão dizer que a tarefa da ciência nunca pode terminar até que o homem deixe de existir, até que a história não seja mais feita e o próprio desenvolvimento cesse?"

"As medalhas são um grande incentivo para os rapazes e os levam a sentir que seu trabalho é valioso. Lembro-me de como senti isso quando nos anos de 1890 recebi as medalhas Darwin e a Huxley. Quando alguém é velho, não quer encorajamento e continua seu trabalho na extensão de seu poder, porque se tornou habitual."

(Karl Pearson)

RESUMO

A análise de componentes principais (PCA, do inglês "*Principal Component Analysis*") é um método multivariado amplamente utilizado, principalmente por sua capacidade de conter em poucas variáveis latentes, conhecidas como componentes principais, uma grande proporção da variância total de todas as variáveis originais. Entretanto, a PCA sofre pelo fato de cada componente principal ser a combinação linear de todas as variáveis originais, o que frequentemente ocasiona dificuldades na interpretação dos resultados. Uma das formas adotadas para contornar essa dificuldade é observar os *loadings* que acompanham cada variável e ignorar aqueles cujos valores sejam pequenos. O componente assim obtido passa a ser a combinação linear envolvendo as variáveis remanescentes. Embora essa prática seja muito utilizada, este procedimento é potencialmente enganoso por se basear na subjetividade. A análise de componentes principais esparsos (SPCA, do inglês "*Sparse Principal Component Analysis*") surgiu como um método que pode ser aplicado para melhorar essa desvantagem da PCA. Sendo um tema de intensa pesquisa por mais de uma década, o método SPCA proposto por Zou, Hastie e Tibshirani em 2006 modifica a formulação original da PCA por tratá-la como um problema de regressão pela introdução da penalidade LASSO, acrônimo de *Least Absolute Shrinkage and Selection Operator*, que é útil por induzir a esparsidade (*loadings* nulos) nos componentes principais. Diante do que foi exposto, são propostos dois novos métodos com o objetivo de facilitar a interpretação dos resultados na PCA, principalmente para cenários em que o problema sob investigação possua um número muito elevado de variáveis. Os métodos propostos foram denominados *Sparse Group for Principal Component Analysis* (SGPCA) e *Pairwise Absolute Clustering and Sparsity for Principal Component Analysis* (PACSPCA). Os métodos SGPCA e PACSPCA se baseiam nos métodos de regressão *Octogonal Shrinkage and Clustering Algorithm for Regression* (OSCAR) e *Pairwise Absolute Clustering and Sparsity* (PACS), respectivamente. Os dois novos métodos propostos, além de também induzirem a esparsidade nos componentes como o método SPCA, também possuem a capacidade de agrupar variáveis utilizando-se da correlação entre as mesmas pela igualdade dos seus *loadings*. Como ilustração, os métodos propostos SGPCA e PACSPCA foram aplicados a dados reais e simulados, visando elucidar algumas de suas características.

Palavras-chave: PCA. SPCA. Esparsidade. OSCAR. PACS. Agrupamento.

ABSTRACT

Principal component analysis (PCA) is a multivariate method widely used, mainly because of its ability to synthesize in a few latent variables, known as principal components, a large proportion of the total variance of all original variables. However, PCA suffers from the fact that each principal component is the linear combination of a very large number of original variables, which often causes difficulties in interpreting the results. One of the ways adopted to overcome this difficulty is to observe the loadings that accompany each variable and ignore those whose values are small. The component thus obtained becomes the linear combination involving the remaining variables. Although this practice is widely used, this procedure is potentially misleading as it is based on subjectivity. Sparse principal component analysis (SPCA) has emerged as a method that can be applied to improve this disadvantage of PCA. Being a subject of intense research for over a decade, the SPCA method proposed by Zou, Hastie and Tibshirani in 2006 modifies the original formulation of the PCA by treating it as a regression problem by introducing the LASSO penalty, acronym for Least Absolute Shrinkage and Selection Operator, which is useful for inducing sparse (null loadings) in the principal components. Because of the above, two new methods are proposed in order to facilitate the interpretation of results in the PCA, mainly for scenarios in which the problem under investigation has a very large number of variables. The proposed methods were called Sparse Group for Principal Component Analysis (SGPCA) and Pairwise Absolute Clustering and Sparsity for Principal Component Analysis (PACSPCA). The SGPCA and PACSPCA methods are based on the Octogonal Shrinkage and Clustering Algorithm for Regression (OSCAR) and Pairwise Absolute Clustering and Sparsity (PACS) regression methods, respectively. The two new methods proposed, in addition to also inducing the sparsity in the components such as the SPCA method, also can group variables using the correlation between them by the equality of their loadings. As an illustration, the proposed SGPCA and PACSPCA methods were applied to real and simulated data, aiming to elucidate some of their characteristics.

Keywords: PCA. SPCA. Esparsity. OSCAR. PACS. Grouping.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

arg	Argumento
cu. in.	*Cubic inch*
lbs	Libras
max	Máximo
mpg	*Miles per gallon*
min	Mínimo
US	*United States*

LISTA DE SIGLAS

BLUE	*Best Linear Unbiased Estimator*
LASSO	*Least Absolute Shrinkage and Selection Operator*
en	*Elastic Net*
nen	*Naïve Elastic Net*
ols	*Ordinary Least Squares*
OSCAR	*Octogonal Shrinkage and Clustering Algorithm for Regression*
PACS	*Pairwise Absolute Clustering and Sparsity*
PACSPCA	*Pairwise Absolute Clustering and Sparsity for Principal Component Analysis*
PCA	*Principal Component Analysis*
PCR	*Principal Component Regression*
SGPCA	*Sparse Group for Principal Component Analysis*
SPCA	*Sparse Principal Component Analysis*

LISTA DE SÍMBOLOS

$\boldsymbol{x}, \boldsymbol{X}, \ldots$	vetores e matrizes
$\boldsymbol{X}$	Matriz de observações ou matriz de delineamento
$'$	Transposto de um vetor ou uma matriz
$\sim$	"Segue a distribuição"
$N\left(0, \sigma^2\right)$	Distribuição Normal com média 0 e variância σ^2
$\boldsymbol{y}$	Vetor resposta
$\boldsymbol{\beta}$	vetor de parâmetros desconhecidos de regressão
$\boldsymbol{\varepsilon}$	Vetor de erros aleatórios
$\mathbb{R}$	Conjunto dos números reais
$\|\cdot\|$	Função norma
$\mathrm{Im}\left(\boldsymbol{X}\right)$	Imagem de $\boldsymbol{X}$
$P_{\mathrm{Im}(\boldsymbol{X})}\left(\boldsymbol{y}\right)$	Projeção de $\boldsymbol{y}$ na imagem de $\boldsymbol{X}$
$\mathrm{rank}\left[\cdot\right]$	Posto de uma matriz
$\dim\left[\cdot\right]$	Dimensão de uma matriz
$E\left[\cdot\right]$	Esperança de uma variável aleatória
$\mathrm{var}\left[\cdot\right]$	Variância de uma variável aleatória
$\mathrm{cov}\left[\cdot\right]$	Covariância de um vetor ou uma matriz
$\boldsymbol{I}$	Matriz identidade
$\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$	Estimador de mínimos quadrados ordinários
$\in$	"Pertence a"
$\subset$	"Está contido em"
$\ker\left(\cdot\right)$	Núcleo de uma matriz
$\left\langle \cdot, \cdot \right\rangle$	Produto interno
$\mathrm{var}_{\mathrm{total}}\left[\cdot\right]$	Variância total de uma matriz de covariâncias
$\mathrm{tr}\left[\cdot\right]$	Traço de uma matriz
$\boldsymbol{X}_{\mathrm{c}}$	Matriz de observações em sua forma centrada
$\boldsymbol{J}$	Matriz quadrada de 1's
$\boldsymbol{X}_{\mathrm{p}}$	Matriz de observações em sua forma padronizada
$\boldsymbol{j}$	Vetor coluna de 1's
$\boldsymbol{X}_1$	Matriz de observações sem o vetor coluna de 1's
$\boldsymbol{0}$	Vetor ou matriz de 0's

$\perp$	Ortogonal
$\oplus$	Soma direta
$\lvert\cdot\rvert$	Valor absoluto de um escalar ou determinante de uma matriz
$\lVert\cdot\rVert_1$	Norma L_1 ou de Manhattan
$\lVert\cdot\rVert_2$	Norma L_2 ou Euclidiana
$\lVert\cdot\rVert_3$	Norma L_3 ou de Chebyshev
$\arg\min_{\boldsymbol{\beta}}$	Argumento mínimo em $\boldsymbol{\beta}$
$\hat{\boldsymbol{\beta}}_{\text{Ridge}}$	Estimador *Ridge*
$\hat{\boldsymbol{\beta}}_{\text{LASSO}}$	Estimador LASSO
$\gg$	Muito maior que
$\hat{\boldsymbol{\beta}}_{\text{nen}}$	Estimador *Naïve Elastic Net*
$\hat{\boldsymbol{\beta}}_{\text{en}}$	Estimador *Elastic Net*
$\cos(\theta)$	Cosseno de θ
∂	Derivada parcial
$\boldsymbol{\Sigma}$	Matriz de covariâncias populacionais
$\text{diag}(\cdot)$	Matriz diagonal
$\boldsymbol{e}_i$	Vetor canônico com argumento unitário na posição i
$\hat{\boldsymbol{\beta}}_{\text{OSCAR}}$	Estimador OSCAR
$\hat{\boldsymbol{\beta}}_{\text{PACS}}$	Estimador PACS

SUMÁRIO

1 INTRODUÇÃO

Com o avanço e desenvolvimento de novas tecnologias na área de computação tem-se tornado cada vez mais comum a exigência da análise da ocorrência de determinados fenômenos, que estão relacionados com muitas variáveis de interesse. Dessa maneira, para uma melhor compreensão dos fenômenos que estejam em investigação, não é de interesse que as variáveis sejam tratadas separadamente ou que muitas não sejam levadas em consideração. Na estatística, isso exigiu nas décadas recentes o desenvolvimento de teorias e métodos que possibilitem a análise nesse cenário envolvendo dados superdimensionados.

A característica de alta dimensionalidade dos dados tem sido frequentemente verificada nas modelagens estatísticas. Por sua vez, em decorrência desse fato, alguns problemas como o alto custo computacional e dificuldades no processo de inferência estatística em altas dimensões acabam sendo observados. Diante disso, uma propriedade desejada nessa situação é a esparsidade das soluções, que consiste basicamente em descartar as variáveis que sejam pouco relevantes ao modelo. Em termos numéricos, a esparsidade em um modelo está relacionada a identificação, de forma eficiente, de estimativas para os parâmetros populacionais com valores aproximadamente nulos ou que ainda possam ser consideradas estatisticamente nulas pela utilização de um teste apropriado.

A propriedade de esparsidade é muito desejada em modelos para dados superdimensionados, pois ela permite após a identificação e remoção de variáveis pouco representativas para o modelo, que ocorra uma consequente diminuição da dimensão do problema. Isso permite uma maior facilidade na interpretação do modelo ajustado e também que boas predições possam ser obtidas. A busca de soluções esparsas associada ao ajuste e a seleção de variáveis em um modelo têm sido objeto de investigação em muitos trabalhos. Em algumas situações é desejável também que algumas variáveis sejam agrupadas. Geralmente, para o agrupamento supervisionado leva-se em consideração a correlação entre as variáveis e o fato dessas variáveis apresentarem efeitos semelhantes em relação a uma determinada variável resposta. Nesse sentido, os agrupamentos obtidos podem ser avaliados para se determinar o que contribui para que as variáveis, presentes em cada grupo, tenham um comportamento semelhante.

Em particular, para a teoria da regressão linear o número elevado de covariáveis demanda, além de um robusto método de estimação dos parâmetros, também um método de seleção de covariáveis. O método de estimação dos mínimos quadrados apresenta problemas no tratamento de covariáveis que sejam altamente correlacionadas. A regressão *Ridge* foi original-

mente introduzida para ser utilizada em situações de multicolinearidade e, consequentemente, esse método de regressão pode ser pensado como uma versão estabilizada para o estimador de mínimos quadrados (HOERL; KENNARD, 1970). A regressão *Ridge* é mais estável, mas raramente gera estimativas nulas para os coeficientes dos parâmetros, o que é uma limitação ao seu uso num cenário de investigação contendo muitas covariáveis. Em 1996, Tibshirani introduziu o estimador "operador de seleção e encolhimento absoluto mínimo" (LASSO, do inglês "*Least Absolute Shrinkage and Selection Operator*") e a principal diferença entre essa nova proposta e a regressão *Ridge* é que o método LASSO tem a capacidade de fornecer estimativas esparsas (nulas). Esse método também encolhe estimativas de forma similar à regressão *Ridge*. Porém, a principal motivação para o seu uso é a seleção de covariáveis.

No sentido de se obter um método que possua as propriedades do LASSO, mas que consiga minimizar suas deficiências, Zou e Hastie (2005) propuseram o método *Elastic Net*. Esse método pode ser pensado como um método "híbrido" entre as regressões *Ridge* e LASSO, por combinar as penalizações utilizadas nesses dois métodos. Além das propriedades de encolhimento e de esparsidade, o *Elastic Net* possui também a vantagem de fazer seleção por grupo, isto é, possui a tendência de selecionar simultaneamente covariáveis altamente correlacionadas. Além do interesse em técnicas que forneçam modelos com menor erro quadrático médio (encolhimento dos coeficientes) e modelos mais parcimoniosos devido a esparsidade, têm-se tornado cada vez mais populares as técnicas que sejam capazes de fornecer agrupamentos supervisionados de covariáveis. Nesse cenário, Bondell e Reich (2008) e Sharma, Bondell e Zhang (2013) propuseram, respectivamente, os métodos "algoritmo de encolhimento e de agrupamento octogonal para regressão" (OSCAR, do inglês "*Octogonal Shrinkage and Clustering Algorithm for Regression*") e "agrupamento e esparsidade absoluta de pares" (PACS, do inglês "*Pairwise Absolute Clustering and Sparsity*"), que executam a seleção de covariáveis, enquanto agrupam preditores automaticamente.

Uma ferramenta de análise exploratória de dados e para a construção de modelos preditivos é a técnica denominada de análise de componentes principais (PCA, do inglês "*Principal Component Analysis*"), que foi introduzida por Karl Pearson em 1901 e está fundamentada no artigo de Hotelling de 1933. Essa é uma técnica de redução de dimensão muito popular, que tem sido aplicada e utilizada com muito sucesso em todas as áreas em que dados multivariados são encontrados. Entre as áreas da ciência em que a técnica PCA tem sido utilizada, podem-se citar a estatística, aprendizado de máquinas, genética, entre outras. O principal objetivo ao

se aplicar esse método é encontrar combinações lineares de variáveis originais, chamadas de componentes principais, que sejam não correlacionadas e que expliquem o máximo possível a estrutura de variância e covariância das variáveis originais. Entretanto, a interpretabilidade dos componentes principais torna-se problemática, mesmo para um número moderado de variáveis.

Diante desse problema de interpretabilidade dos componentes principais, Zou, Hastie e Tibshirani (2006) introduziram um novo método chamado análise de componentes principais esparsos (SPCA, do inglês "*Sparse Principal Component Analysis*"), utilizando o LASSO (*Elastic Net*) para produzir componentes principais com *loadings* nulos. Os autores mostraram primeiramente que a PCA pode ser formulada como um problema de regressão. Os *loadings* esparsos são então obtidos pela imposição da restrição LASSO (*Elastic Net*) nos componentes principais. A ideia utilizada na concepção do método SPCA forneceu os fundamentos teóricos para a formulação de dois métodos propostos neste trabalho.

Diante do que foi exposto, o presente trabalho tem por objetivo geral propor dois novos métodos a PCA, que fornecem *loadings* esparsos e/ou iguais (agrupamento), utilizando para isso os métodos de regressão OSCAR e PACS. Por sua vez, os objetivos específicos são:

1. Apresentar algumas definições e alguns resultados sobre a PCA e os métodos de regressão *Ridge*, LASSO, *Elastic Net*, OSCAR e PACS.

2. Apresentar os principais resultados do método SPCA, com ênfase em aspectos geométricos.

3. Exemplificar a utilização dos métodos propostos com aplicações em exemplos contendo dados simulados e reais.

Além dessa breve introdução, o presente trabalho está organizado como descrito a seguir. Entre as seções 2.1 e 2.10 apresenta-se a fundamentação teórica do trabalho. Na seção 2.1 discuti-se a organização dos dados no contexto da estatística multivariada e da teoria de regressão. A seção 2.2 é reservada à regressão linear múltipla, que recebe uma abordagem geométrica. Nessa seção, o estimador de mínimos quadrados é definido utilizando-se conceitos geométricos como matriz de projeção e projetor ortogonal. A seção 2.3 versa sobre a invariância do estimador de mínimos quadrados quando a matriz de observações está em sua forma centralizada, enquanto na seção 2.4 são destacados os problemas que decorrem desse estimador em cenários envolvendo multicolinearidade dos dados. Na seção 2.5 são apresentadas algumas normas de vetores que são utilizadas no contexto de regressão penalizada. A seção 2.6 é dedicada a três importantes métodos da regressão penalizada, a saber, os métodos *Ridge*, LASSO

e *Elastic Net*. Os métodos são abordados com ênfase geométrica, destacando-se como atuam as suas respectivas penalidades em relação às estimativas de mínimos quadrados. Na seção 2.7 aborda-se a PCA, apresentando-se a definição e algumas propriedades dos componentes principais. Na seção 2.8 discuti-se a relação entre os componentes principais e o estimador *Ridge*. Nessa seção, apresenta-se o desenvolvimento do método SPCA, que forneceu a base teórica para a formulação dos métodos propostos no presente trabalho. Nas seções 2.9 e 2.10 são apresentados os métodos de regressão OSCAR e PACS, que são as técnicas de regressão subjacentes aos novos métodos formulados. Na seção 3 apresenta-se a proposta deste trabalho, que consiste em reformular a PCA como um problema de regressão, utilizando os métodos OSCAR e PACS para induzir a esparsidade e a igualdade nos coeficientes (*loadings*) dos componentes principais. Ainda nessa seção, são apresentados os exemplos simulados e envolvendo dados reais para avaliar o desempenho dos dois métodos propostos, bem como a metodologia utilizada para atingir essa finalidade. Por último, as seções de 4 a 4.6 foram reservadas a análise dos resultados, à discussão e à conclusão do trabalho.

Cabe ressaltar que as seções 2.2, 2.6.1, 2.7.5 e 2.7.7 são fortemente fundamentadas nos trabalho de Silveira (2014) e Pereira (2017), sendo dado os devidos créditos aos autores. Nessas seções, as citações dessas referências serão usualmente omitidas.

2 REFERENCIAL TEÓRICO

2.1 A organização dos dados

Uma observação multivariada pode ser entendida como a coleção de medidas de p diferentes variáveis em relação a um mesmo item, indivíduo, unidade amostral ou unidade experimental. Essas medidas podem ser dispostas em uma matriz $\boldsymbol{X}_{n\times p}$, em que n representa o número de observações (n itens, indivíduos, unidades amostrais ou unidades experimentais) e p é o número de variáveis (JOHNSON; WICHERN, 2007, p. 5). Dessa maneira, no presente trabalho adotou-se como disposição para as observações multivariadas e para cada variável, as linhas e as colunas, respectivamente.

Essa disposição na matriz de dados pode ser avaliada de dois pontos de vista alternativos. O primeiro é concernente a disposição das variáveis nas colunas, sendo que em cada coluna tem-se n repetições para cada uma das p variáveis. Logo, a comparação de duas colunas de $\boldsymbol{X}$ consiste basicamente na análise da relação entre duas variáveis. Por sua vez, cada uma das n linhas representa uma observação multivariada, contendo p valores observados. Neste sentido, uma comparação de duas linhas de $\boldsymbol{X}$ envolve a análise da relação entre diferentes itens, indivíduos, unidades amostrais ou unidades experimentais.

Diante do que foi exposto, a matriz geral de dados $\boldsymbol{X}$, de dimensões $(n \times p)$, é dada por:

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}. \tag{2.1}$$

Consequentemente, a notação x_{ij} é utilizada para indicar o valor particular da j-ésima variável que é observado no i-ésimo item, sendo $i = 1,2,\dots,n$ e $j = 1,2,\dots,p$. Nesse sentido, cada linha de $\boldsymbol{X}$ representa uma observação multivariada e a amostra multivariada consiste de n repetições, cada uma contendo p medidas. Diante disso, tem-se:

$$\boldsymbol{X}_{n\times p} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \begin{matrix} X_1 & X_2 & \dots & X_p \end{matrix} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \end{matrix} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \begin{matrix} \\ \rightarrow \text{ Primeira observação multivariada} \\ \rightarrow \text{ Segunda observação multivariada} \\ \vdots \\ \rightarrow n\text{-ésima observação multivariada} \end{matrix} \tag{2.2}$$

Conforme Mardia, Kent e Bibby (1979, p. 8-9), a matriz $\boldsymbol{X}$ também pode ser escrita como $\boldsymbol{X} = \left(x_{ij}\right)$, em que x_{ij} é o elemento que se encontra na linha i e na coluna j. No presente trabalho, as linhas de $\boldsymbol{X}$ serão denotadas por $(\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_n)$. Observe que $\boldsymbol{x}_i$ denota a i-ésima linha de $\boldsymbol{X}$ escrita como uma coluna. Por sua vez, as colunas de $\boldsymbol{X}$ serão denotadas com subscritos entre parênteses, como $\left(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \ldots, \boldsymbol{x}_{(p)}\right)$. Dessa forma, a matriz $\boldsymbol{X}$ pode ser apresentada de forma equivalente como uma matriz de n vetores linha ou p vetores coluna, da seguinte maneira:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \vdots \\ \boldsymbol{x}'_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{(1)} & \boldsymbol{x}_{(2)} & \ldots & \boldsymbol{x}_{(p)} \end{bmatrix}, \tag{2.3}$$

em que

$$\boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \quad (i = 1, 2, \ldots, n) \ , \quad \boldsymbol{x}_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad (j = 1, 2, \ldots, p) \, . \tag{2.4}$$

Pode ser observado que na matriz de dados $\boldsymbol{X}_{n \times p}$, a massa de informações é organizada no formato retangular. Conforme Lebart, Morineau e Piron (1995, p. 8), para entender o princípio dos métodos estatísticos exploratórios multidimensionais, é útil representar geometricamente as n linhas e p colunas da matriz $\boldsymbol{X}$ por pontos cujas coordenadas são precisamente os elementos dessa matriz (FIGURA 2.1).

Dessa maneira, duas nuvens de pontos podem ser consideradas:

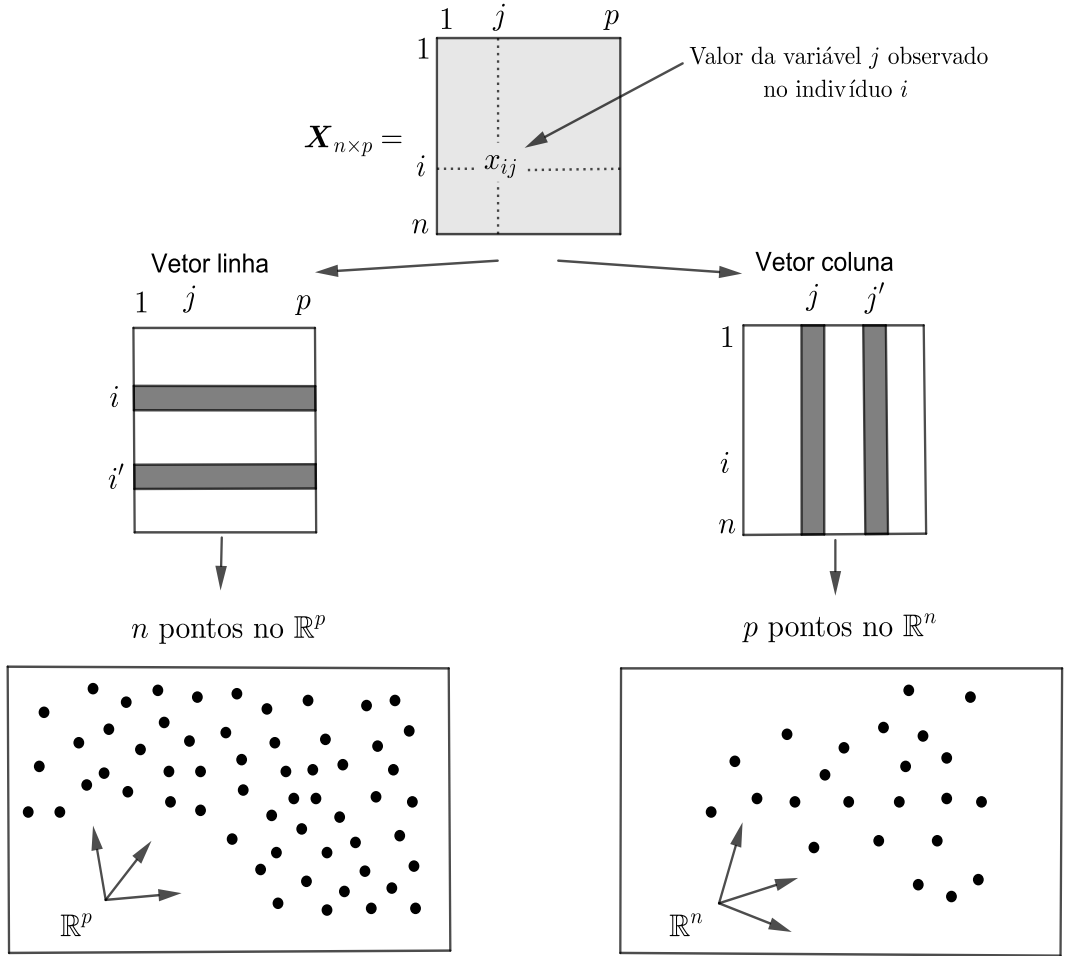
1. Uma nuvem de n indivíduos (n linhas de $\boldsymbol{X}$) localizada no espaço p-dimensional $\mathbb{R}^p$ das variáveis (colunas de $\boldsymbol{X}$). Cada uma das n linhas é representada por um ponto com p coordenadas.

2. Uma nuvem de p variáveis (p colunas de $\boldsymbol{X}$) localizada no espaço n-dimensional $\mathbb{R}^n$ dos indivíduos (linhas de $\boldsymbol{X}$). Cada uma das p colunas é representada por um ponto com n coordenadas.

Ainda segundo Lebart, Morineau e Piron (1995, p. 9), cada uma das duas dimensões da matriz de observações permite definir distâncias (ou proximidades) entre os elementos que

definem a outra dimensão. As proximidades geométricas usuais nos pontos referentes a linhas ou entre os pontos relacionados a colunas, na realidade traduzem associações estatísticas de interesse entre indivíduos ou entre variáveis, respectivamente. Dessa forma, na matriz de dados a maior parte das técnicas multivariadas tem como alvo linhas, colunas ou ambas. Desse modo, trabalhar nas linhas da matriz $\boldsymbol{X}$ significa trabalhar no espaço das variáveis, ou seja, em $\mathbb{R}^p$. Da mesma forma, as técnicas estatísticas que trabalham nas colunas da matriz $\boldsymbol{X}$ estão no espaço dos indivíduos, no $\mathbb{R}^n$. Uma observação a ser feita é que na Figura 2.1 a quantidade de pontos no $\mathbb{R}^p$ é superior a quantidade de pontos no $\mathbb{R}^n$ e isso não ocorreu por mero acaso. Ao longo de todo o trabalho será considerado o caso em que número de observações n é superior ao número de variáveis p.

Figura 2.1 – Representação geométrica da matriz de observações $\boldsymbol{X}$.



Fonte: Adaptado de Lebart, Morineau e Piron (1995).

Em se tratando da análise de regressão, a matriz de dados $\boldsymbol{X}$ possui uma alteração em suas dimensões, pelo acréscimo de uma coluna contendo 1's. Isso se deve a presença do in-

tercepto no modelo usual de regressão, que será discutido na próxima seção. Assim, a matriz possui dimensões $(n \times (p+1))$, sendo dada por:

$$\boldsymbol{X}_{n\times(p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}. \tag{2.5}$$

Quando os valores x_{ij}'s são planejados pelo pesquisador, a matriz $\boldsymbol{X}$ é denominada de matriz de delineamento (em inglês "*design matrix*"). No presente trabalho, assumiremos que $n > (p+1)$ e que $\text{rank}\,[\boldsymbol{X}] = p+1$, ou seja, $\boldsymbol{X}$ terá posto coluna completo.

2.2 Uma abordagem geométrica à regressão linear múltipla

A regressão linear múltipla admite uma abordagem geométrica que possibilita um tratamento similar para três teorias importantes: as regressões *Ridge*, LASSO e *Elastic Net*, que serão discutidas posteriormente. Esse tipo de abordagem não é usual, sendo que cada uma das teorias anteriormente citadas aparenta demandar resultados teóricos diferentes. Neste trabalho, a abordagem geométrica é utilizada como uma ferramenta para a construção dessas teorias.

Primeiramente, será analisado o caso envolvendo a regressão linear simples. Sejam $[(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)]$ os valores observados, isto é, n pontos que não são necessariamente colineares. Dentre todas as curvas candidatas para o ajuste desses pontos, deseja-se aquela que torne os erros de estimação pequenos. Por erro de estimação entende-se a diferença entre um valor observado y_i e um valor ajustado $\hat{y}_i$. Dessa forma, o objetivo consiste em se minimizar uma função das diferenças entre os valores observados y_i e os valores ajustados $\hat{y}_i$, para $i = 1,2,\dots,n$. Como os erros podem ser positivos ou negativos, a soma total dos erros pode ser pequena mesmo quando o ajuste de uma curva aos dados não é bom. Isso fica mais evidente sob a suposição de normalidade dos erros. Nessa situação, a soma total dos erros é exatamente nula, independente se o ajuste é bom ou ruim.

Uma maneira de contornar essa dificuldade seria impor que a soma dos valores absolutos nas diferenças (erros) fosse tão pequena quanto possível. Entretanto, minimizar uma função das somas de valores absolutos não é conveniente em algumas deduções matemáticas, pois esse tipo de função (função modular) não é diferenciável em todos os números reais. Consequentemente,

a dificuldade apresentada anteriormente pode ser evitada pela imposição de que a soma de quadrados dos erros seja mínima. A imposição da soma de quadrados entre as diferenças de y_i e $\hat{y}_i$ é interessante, pois essa imposição penaliza mais fortemente os grandes desvios entre y_i e $\hat{y}_i$.

Considere então o modelo de regressão linear simples, dado por:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{2.6}$$

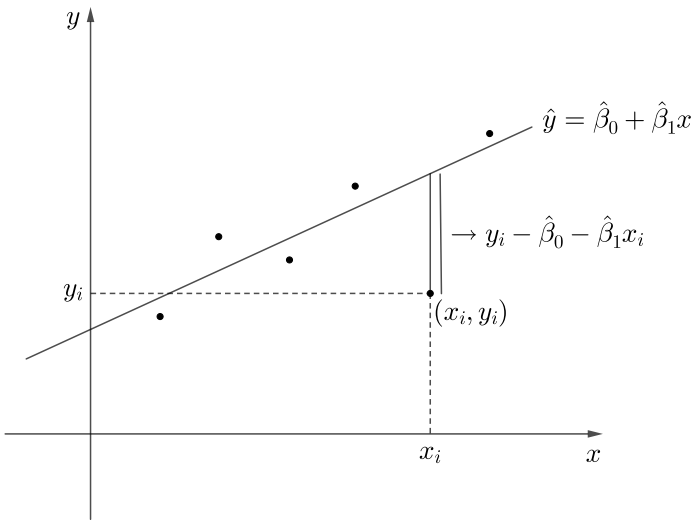
em que β_0 e β_1 são os parâmetros do modelo e ε_i é o erro associado a y_i, sendo que geralmente $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

O objetivo consiste em se minimizar a soma de quadrados dos erros ε_i, que é dada por:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2. \tag{2.7}$$

Um método usual para se obter os estimadores de β_0 e β_1 é o método dos mínimos quadrados. Utilizando esse método, os estimadores para β_0 e β_1 podem ser obtidos derivando-se (2.7) em relação a cada um desses parâmetros, igualando as respectivas derivadas a 0 e resolvendo-se o sistema obtido. A Figura 2.2 ilustra o contexto relacionado à minimização da soma dos quadrados dos desvios $y_i - (\beta_0 + \beta_1 x_i)$, com destaque a reta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ajustada por mínimos quadrados.

Figura 2.2 – Reta ajustada por mínimos quadrados, com destaque a diferença entre y_i e $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.



Fonte: Adaptado de Pereira (2017).

Essa reta ajustada é conhecida como reta de mínimos quadrados ou reta de regressão. Esse último termo foi dado por um pioneiro no campo da Estatística aplicada, Francis Galton (1822-1911). Galton era primo do biólogo, geólogo e naturalista britânico Charles Darwin (1809-1882), autor do livro *The Origin of Species*. Galton acreditava ser possível que os seres humanos herdassem de seus ancestrais tanto características físicas quanto intelectuais. Durante seis anos em seu laboratório fundado no ano de 1864 em Londres, ele coletou 9000 registros familiares, muitos deles completos. Esses registros levaram dez anos para serem analisados (CONT, 2008). Foi nesse contexto que Galton descobriu o fenômeno de regressão à média de algumas variáveis. Este novo conceito possibilitou que se explicasse, por exemplo, o controle da estatura entre pais e filhos, isto é, que a altura dos homens tende a permanecer estável, em média. Se a regressão à média não ocorresse, filhos de pais altos seriam ainda mais altos e filhos de pais baixos seriam ainda mais baixos. Isso prosseguiria de geração em geração (SALSBURG, 2009, p. 26).

Uma abordagem geométrica do modelo de regressão é obtida da seguinte forma: se $\boldsymbol{y}'_{1\times n} = (y_1,\ldots,y_n)$, $\boldsymbol{\beta}'_{1\times 2} = (\beta_0,\beta_1)$, $\boldsymbol{\varepsilon}'_{1\times n} = (\varepsilon_1,\ldots,\varepsilon_n)$ e $\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, então:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{2.8}$$

isto é, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, para $i = 1,2,\ldots,n$. Portanto, reescrevendo (2.7) tem-se que:

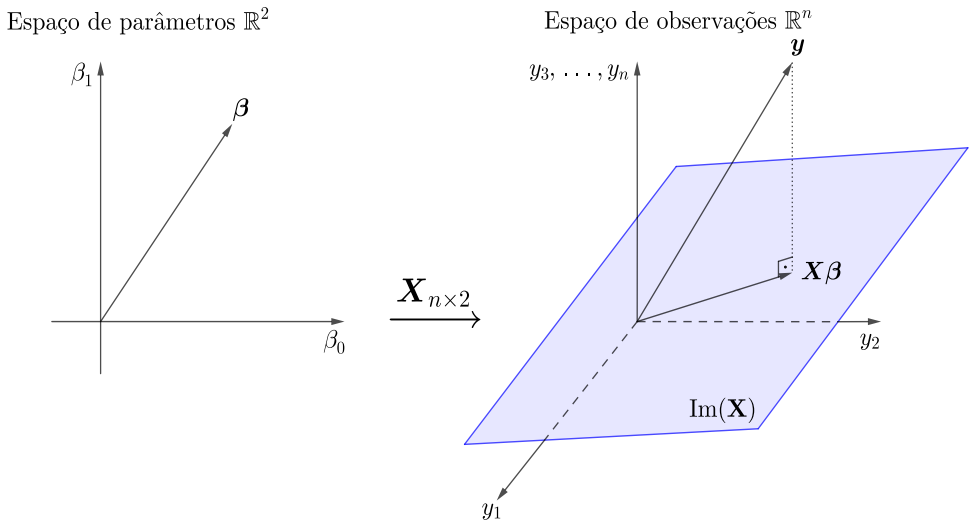
$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \left\| \boldsymbol{y}_{n\times 1} - \boldsymbol{X}_{n\times 2}\boldsymbol{\beta}_{2\times 1} \right\|^2. \tag{2.9}$$

Em termos de transformações lineares, a matriz $\boldsymbol{X}_{n\times 2}$ atua como uma transformação do espaço de parâmetros para o espaço de observações. Esse conceito está representado na Figura 2.3.

Não levando em consideração as dimensões de $\boldsymbol{y}$, $\boldsymbol{X}$ e $\boldsymbol{\beta}$, segue que (2.9) pode ser reescrita da seguinte maneira:

$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \tag{2.10}$$

Figura 2.3 – Geometria da regressão linear simples.



Fonte: Adaptado de Pereira (2017).

Dessa maneira, minimizar $\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$ equivale a minimizar $\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$. Claramente, isso ocorre quando $\boldsymbol{X\beta}$ é a projeção ortogonal do vetor de dados $\boldsymbol{y}$ no subespaço $\mathrm{Im}(\boldsymbol{X})$ (FIGURA 2.3). Esse vetor projeção será denotado por $\hat{\boldsymbol{y}} = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$, o vetor de dados ajustados. Pode-se notar que, uma vez que o posto coluna de $\boldsymbol{X}$ é 2, a dimensão da imagem de $\boldsymbol{X}$ também será 2, isto é, $\dim[\mathrm{Im}(\boldsymbol{X})] = 2$.

O modelo de regressão linear múltipla descreve $\boldsymbol{y}$ como uma soma de uma parte determinística e uma parte aleatória, sendo a parte determinística mais geral, de forma que podemos expressar o valor esperado de $\boldsymbol{y}$ como função de várias variáveis regressoras $(X_1, X_2, \ldots, X_p)$. O modelo de regressão linear múltipla pode ser expresso como:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon. \tag{2.11}$$

Para estimar os parâmetros $(\beta_0, \beta_1, \ldots, \beta_p)$ em (2.11) é necessário utilizar uma amostra de n observações em y, que são associadas as variáveis $(X_1, X_2, \ldots, X_p)$. O modelo para a i-ésima observação é:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{2.12}$$

em que x_{ij} é o valor fixo da variável regressora X_j, para $i = 1, 2, \ldots, n$.

Para todo $i, k = 1, 2, \ldots, n$, as suposições para ε_i (ou y_i) do modelo (2.12) são:

1. $E[\varepsilon_i] = 0$ ou, equivalentemente, $E[y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

2. $\text{var}[\varepsilon_i] = \sigma^2$ ou, equivalentemente, $\text{var}[y_i] = \sigma^2$.

3. $\text{cov}[\varepsilon_i, \varepsilon_k] = 0$ ou, equivalentemente, $\text{cov}[y_i, y_k] = 0$, para $i \neq k$.

A suposição 1 estabelece que o modelo (2.12) é correto, isto é, todos os x's relevantes estão incluídos no modelo de forma linear. A suposição 2 indica que a variância de y é constante e não depende dos x's. Essa suposição é conhecida como suposição de homocedasticidade, de homogeneidade de variâncias ou de variâncias constantes. A suposição 3 estabelece que os ε's ou os y's são não correlacionados entre si, o que geralmente acontece em amostras aleatórias. A inclusão da suposição de normalidade estabelece ainda que os ε's ou os y's serão independentes (RENCHER; SCHAALJE, 2008, p. 138).

Matricialmente, o modelo (2.12) pode ser escrito como $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ ou

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1} . \tag{2.13}$$

As suposições para $\boldsymbol{\varepsilon}$ ou $\boldsymbol{y}$, considerando o modelo (2.13), são dadas por:

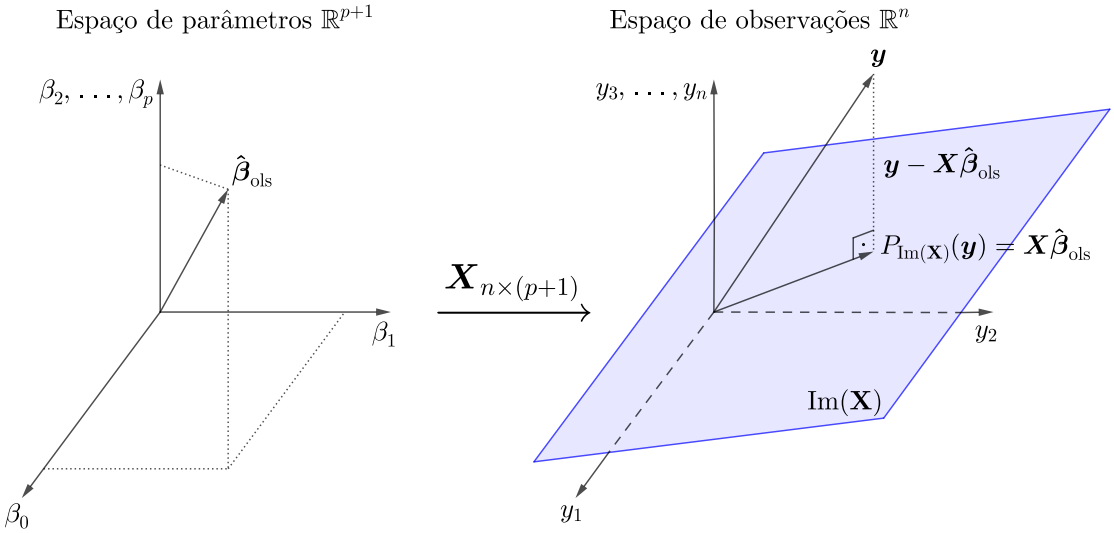
1. $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ ou $E[\boldsymbol{y}] = \boldsymbol{X\beta}$.

2. $\text{cov}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$ ou $\text{cov}[\boldsymbol{y}] = \sigma^2\boldsymbol{I}$.

A suposição 2 inclui as suposições $\text{var}[\varepsilon_i] = \sigma^2$ e $\text{cov}[\varepsilon_i, \varepsilon_k] = 0$. Cada coluna de $\boldsymbol{X}$, exceto a primeira, é um vetor de dimensão n formado pelos valores observados da variável regressora X_j, que são multiplicados pelo respectivo elemento β_j de $\boldsymbol{\beta}$, para $j = 1, 2, \ldots, p$. Pode-se observar, multiplicando $\boldsymbol{\beta}$ à esquerda pela i-ésima linha de $\boldsymbol{X}$, que estamos obtendo a parte determinística ou sistêmica do modelo. O parâmetro β_0 pode ser interpretado como a resposta média numa situação hipotética em que as variáveis regressoras assumam valores nulos. Isso justifica a inclusão da coluna de 1's em $\boldsymbol{X}$ e, nesse caso, o parâmetro β_0 representa uma parte constante para cada valor y_i. No $\mathbb{R}^2$ e $\mathbb{R}^3$, o parâmetro β_0 corresponderá, respectivamente, ao intercepto da reta com o eixo y (ordenada) e do plano com o eixo z (cota). Por sua vez, cada β_j indica uma mudança esperada na resposta média y_i, devido ao aumento (ou diminuição) de uma unidade em X_j, quando as demais variáveis são mantidas fixas.

Como mencionado anteriormente, deseja-se minimizar a soma de quadrados das diferenças entre os valores observados e ajustados, dada por $\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$. Observe que geometricamente, o modelo para o vetor de médias do vetor $\boldsymbol{y}$ é simplesmente a definição do subespaço vetorial $\mathrm{Im}(\boldsymbol{X})$. O valor mínimo é obtido pela projeção ortogonal do vetor de dados $\boldsymbol{y}$ no subespaço $\mathrm{Im}(\boldsymbol{X})$, que possui dimensão $p+1$. A Figura 2.4 ilustra esta situação.

Figura 2.4 – Geometria da regressão linear múltipla.



Fonte: Adaptado de Pereira (2017).

Geometricamente, essa representação é similar à representação apresentada na Figura 2.3, diferindo apenas na dimensão do espaço dos parâmetros, que agora passa a ser de dimensão $\mathbb{R}^{p+1}$ (FIGURA 2.4). Considerando o posto de $\boldsymbol{X}$ igual a $p+1$ (posto coluna completo), tem-se que $\dim[\mathrm{Im}(\boldsymbol{X})]=p+1$. Para se obter uma expressão para $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$ é necessária a teoria das matrizes de projeção.

Definição 2.2.1 (Matriz de projeção): Uma matriz quadrada $\boldsymbol{A}$ é dita uma *matriz de projeção* ou *projetor* se é idempotente, isto é, $\boldsymbol{A}^2=\boldsymbol{A}$.

Pela definição tem-se que um projetor $\boldsymbol{A}$, restrito a $\mathrm{Im}(\boldsymbol{A})$, é a identidade. De fato, para $\boldsymbol{A}\boldsymbol{z}\in\mathrm{Im}(\boldsymbol{A})$ segue que:

$$\boldsymbol{A}(\boldsymbol{A}\boldsymbol{z})=\boldsymbol{A}^2\boldsymbol{z}=\boldsymbol{A}\boldsymbol{z}. \tag{2.14}$$

Observe que se $\boldsymbol{I}$ é a matriz identidade, então $\boldsymbol{I}-\boldsymbol{A}$ também é um projetor, pois:

$$(\boldsymbol{I}-\boldsymbol{A})^2 = \boldsymbol{I}-2\boldsymbol{A}+\boldsymbol{A}^2 = \boldsymbol{I}-2\boldsymbol{A}+\boldsymbol{A} = \boldsymbol{I}-\boldsymbol{A}. \tag{2.15}$$

Pode ser mostrado que $\mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A}) = \ker(\boldsymbol{A})$. De fato, seja $(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{z} \in \mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A})$. Então:

$$\boldsymbol{A}\left[(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{z}\right] = \boldsymbol{A}\,(\boldsymbol{z}-\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z}-\boldsymbol{A}^2\boldsymbol{z} = \boldsymbol{A}\boldsymbol{z}-\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}. \tag{2.16}$$

Dessa maneira, $(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{z} \in \ker(\boldsymbol{A})$. Em decorrência, $\mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A}) \subset \ker(\boldsymbol{A})$. Por sua vez, se $\boldsymbol{z} \in \ker(\boldsymbol{A})$ tem-se que $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$. Assim,

$$(\boldsymbol{I}-\boldsymbol{A})\,\boldsymbol{z} = \boldsymbol{z}-\boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}-\boldsymbol{0} = \boldsymbol{z}. \tag{2.17}$$

Logo, $\boldsymbol{z} \in \mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A})$ e, consequentemente, $\ker(\boldsymbol{A}) \subset \mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A})$. Portanto, $\mathrm{Im}\,(\boldsymbol{I}-\boldsymbol{A}) = \ker(\boldsymbol{A})$. Pode-se mostrar também que $\mathrm{Im}\,(\boldsymbol{A}) = \ker(\boldsymbol{I}-\boldsymbol{A})$.

Definição 2.2.2 (Projetor ortogonal): Uma matriz de projeção $\boldsymbol{A}$ é dita um *projetor ortogonal* se, para um dado vetor $\boldsymbol{v}$, $\boldsymbol{A}\boldsymbol{v}-\boldsymbol{v}$ é perpendicular ao subespaço $\mathrm{Im}\,(\boldsymbol{A})$.

Proposição 2.2.1: Uma matriz de *projeção* $\boldsymbol{A}$ é *simétrica* se e somente se é um *projetor ortogonal*.

Prova:

Seja $\boldsymbol{A}$ uma matriz de projeção simétrica e suponha que $\boldsymbol{v}$ e $\boldsymbol{w}$ são dois vetores quaisquer. Escrevendo o produto interno na forma matricial, segue da hipótese de $\boldsymbol{A}$ ser uma matriz simétrica que:

$$\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{w}\rangle = \boldsymbol{w}'\boldsymbol{A}\boldsymbol{v} = \left(\boldsymbol{A}'\boldsymbol{w}\right)'\boldsymbol{v} = (\boldsymbol{A}\boldsymbol{w})'\boldsymbol{v} = \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle. \tag{2.18}$$

Diante disso, de (2.18) resulta que $\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle = \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{A}\boldsymbol{w}\rangle$. Dessa maneira, utilizando as propriedades de produto interno e sendo $\boldsymbol{A}^2 = \boldsymbol{A}$, tem-se que:

$$\begin{aligned}
\langle \boldsymbol{A}\boldsymbol{v}-\boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle &= \langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle - \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle \\
&= \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{A}\boldsymbol{w}\rangle - \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle \\
&= \left\langle \boldsymbol{v}, \boldsymbol{A}^2\boldsymbol{w}\right\rangle - \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle \\
&= \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle - \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle \\
&= 0.
\end{aligned} \tag{2.19}$$

Uma vez que $\boldsymbol{Aw} \in \mathrm{Im}(\boldsymbol{A})$, tem-se que $\boldsymbol{Av} - \boldsymbol{v}$ é perpendicular a $\mathrm{Im}(\boldsymbol{A})$. Portanto, $\boldsymbol{A}$ é um projetor ortogonal. Reciprocamente, com base na definição de projetor ortogonal, seja $\boldsymbol{Av} - \boldsymbol{v}$ perpendicular a $\mathrm{Im}(\boldsymbol{A})$. Para um vetor $\boldsymbol{w}$ qualquer, tem-se então que:

$$0 = \langle \boldsymbol{Av} - \boldsymbol{v}, \boldsymbol{Aw} \rangle = \langle \boldsymbol{Av}, \boldsymbol{Aw} \rangle - \langle \boldsymbol{v}, \boldsymbol{Aw} \rangle . \tag{2.20}$$

Assim,

$$\langle \boldsymbol{Av}, \boldsymbol{Aw} \rangle = \langle \boldsymbol{v}, \boldsymbol{Aw} \rangle . \tag{2.21}$$

Para $\boldsymbol{Aw} - \boldsymbol{w}$ perpendicular a $\boldsymbol{Av}$, tem-se também que:

$$0 = \langle \boldsymbol{Aw} - \boldsymbol{w}, \boldsymbol{Av} \rangle = \langle \boldsymbol{Aw}, \boldsymbol{Av} \rangle - \langle \boldsymbol{w}, \boldsymbol{Av} \rangle . \tag{2.22}$$

Dessa forma,

$$\langle \boldsymbol{Aw}, \boldsymbol{Av} \rangle = \langle \boldsymbol{w}, \boldsymbol{Av} \rangle . \tag{2.23}$$

Da comutatividade do produto interno, segue de (2.23) que:

$$\langle \boldsymbol{Av}, \boldsymbol{Aw} \rangle = \langle \boldsymbol{Av}, \boldsymbol{w} \rangle . \tag{2.24}$$

Consequentemente, de (2.21) e (2.24) resulta que:

$$\langle \boldsymbol{Av}, \boldsymbol{w} \rangle = \langle \boldsymbol{v}, \boldsymbol{Aw} \rangle .$$

Portanto, $\boldsymbol{A}$ é uma matriz simétrica.

■

Para se obter uma expressão para a projeção $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$, será apresentada na Proposição 2.2.2 uma dedução baseada apenas em argumentos geométricos.

Proposição 2.2.2: A projeção ortogonal do vetor de dados $\boldsymbol{y}$ no subespaço $\mathrm{Im}(\boldsymbol{X})$ é:

$$P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.25}$$

Prova:

Seja $\boldsymbol{y}' = (y_1, y_2, \ldots, y_n)$. Em razão da aleatoriedade do vetor $\boldsymbol{y}$, devida a sua relação com os erros $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, a probabilidade de $\boldsymbol{y} \in \text{Im}(\boldsymbol{X})$ é nula. Como queremos minimizar a função $L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ e o vetor $\boldsymbol{X}\boldsymbol{\beta}$ pertence à imagem de $\boldsymbol{X}$, segue que $L(\boldsymbol{\beta})$ é minimizada quando $\boldsymbol{X}\boldsymbol{\beta}$ é a projeção ortogonal de $\boldsymbol{y}$ na $\text{Im}(\boldsymbol{X})$.

Logo, existe um vetor $\hat{\boldsymbol{\beta}}$ tal que $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, já que a matriz $\boldsymbol{X}$ é de posto coluna completo. Considere ainda um vetor $\tilde{\boldsymbol{\beta}} = \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$. Como $\boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ é ortogonal a qualquer vetor na imagem de $\boldsymbol{X}$, tem-se que $\boldsymbol{X}' \left[\boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right] = \boldsymbol{0}$. Dessa maneira, $\boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ pertence ao $\ker(\boldsymbol{X})$.

Uma vez que $\tilde{\boldsymbol{\beta}} = \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ e $\boldsymbol{X}' \left[\boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right] = \boldsymbol{0}$, então:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) \\
&= \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) + \boldsymbol{0} \\
&= \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) + \boldsymbol{X}' \left[\boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right] \\
&= \boldsymbol{X}' \left[P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) + \boldsymbol{y} - P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right] \\
&= \boldsymbol{X}'\boldsymbol{y}. \qquad (2.26)
\end{aligned}$$

Como $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, vem que:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{X}' P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}. \qquad (2.27)$$

De (2.26) e (2.27) obtém-se o sistema de equações normais:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}, \qquad (2.28)$$

que foi obtido a partir de propriedades geométricas.

De acordo com Rencher e Shaalje (2008, p. 21-27), para qualquer matriz $\boldsymbol{X}$ segue que $\text{rank}[\boldsymbol{X}'\boldsymbol{X}] = \text{rank}[\boldsymbol{X}]$ e como $\boldsymbol{X}$ é uma matriz $[n \times (p+1)]$ de posto $p+1$, então $\boldsymbol{X}'\boldsymbol{X}$ é positiva definida. Consequentemente, $\boldsymbol{X}'\boldsymbol{X}$ é não singular e possui inversa única. Dessa forma,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \qquad (2.29)$$

Portanto, a projeção ortogonal do vetor de dados $\boldsymbol{y}$ no subespaço $\text{Im}(\boldsymbol{X})$ é:

$$P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad (2.30)$$

conforme representado na Figura 2.4

■

Da Proposição 2.2.2 segue que a estimativa do vetor de parâmetros $\boldsymbol{\beta}$, que nos fornece o ajuste de mínimos quadrados, é $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Cabe destacar que a notação que será utilizada para esse estimador no decorrer do presente trabalho é $\hat{\boldsymbol{\beta}}_{\text{ols}}$, ou seja, $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é o estimador de mínimos quadrados ordinários (ols, do inglês "*Ordinary Least Squares*"). O estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ possui propriedades ótimas. Esse estimador é não viesado, pois:

$$E\left[\hat{\boldsymbol{\beta}}_{\text{ols}}\right] = E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{2.31}$$

Se $\boldsymbol{A}$ é uma matriz de constantes e $\boldsymbol{y}$ é um vetor aleatório, por definição $\text{cov}[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}\text{cov}[\boldsymbol{y}]\boldsymbol{A}'$. Como $\text{cov}[\boldsymbol{y}] = \sigma^2\boldsymbol{I}$, a matriz de covariâncias do vetor $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é dada por:

$$\begin{aligned}
\text{cov}\left[\hat{\boldsymbol{\beta}}_{\text{ols}}\right] &= \text{cov}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\text{cov}[\boldsymbol{y}]\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned} \tag{2.32}$$

A variância total, definida como o traço da matriz de covariâncias, pode ser utilizada como uma medida indicativa da dispersão global das variáveis. Diante disso, a variância total de $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é:

$$\text{var}_{\text{total}}\left[\text{cov}\left[\hat{\boldsymbol{\beta}}_{\text{ols}}\right]\right] = \text{tr}\left[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] = \sigma^2\text{tr}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] = \sigma^2\sum_{i=1}^{p}\frac{1}{\gamma_i},$$

em que γ_i são os autovalores de $\boldsymbol{X}'\boldsymbol{X}$, para $i = 1, 2, \ldots, p$.

Se $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$ e $\text{cov}[\boldsymbol{y}] = \sigma^2\boldsymbol{I}$, segue pelo teorema de Gauss Markov que $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é o melhor estimador linear não viesado (*BLUE*, do inglês "*Best Linear Unbiased Estimator*"), dentre todos os estimadores lineares e não viesados de $\boldsymbol{\beta}$.

2.3 Matriz centrada e estimador de mínimos quadrados

Um aspecto muito importante para a utilização de estimadores como *Ridge*, LASSO e *Elastic Net*, está relacionado ao fato de que as variáveis preditoras devam estar padronizadas e

a variável resposta centralizada. O processo de padronização de variáveis consiste em reduzir todas as variáveis sob estudo para a mesma locação (ou posição) e mesma escala. Reduzir as variáveis para a mesma locação significa que os dados estarão centralizados na origem. Em relação à escala, em muitas situações é necessário que medições sejam realizadas em diferentes unidades. Em todas as situações comuns em experimentos é razoável exigir que a inferência estatística a ser realizada seja independente das unidades de medida envolvidas. Dessa forma, reduzir as variáveis para a mesma escala consiste em garantir que as observações de cada variável estejam numa escala unitária.

Para reduzir uma variável X_j a sua forma padronizada Z_j, basta subtrairmos de X_j a sua média e dividir esse resultado pelo respectivo desvio padrão de X_j, como se segue:

$$Z_j = \frac{X_j - \mu_j}{\sigma_j}, \tag{2.33}$$

para $j = 1, 2, \ldots, p$, em que μ_j e σ_j são a média e o desvio padrão de X_j, respectivamente.

Dada uma matriz $\boldsymbol{X}_{n\times p}$ de observações, como em (2.1), contendo n repetições para cada uma das p variáveis, como todas as variáveis X_j podem ser padronizadas? A resposta para essa pergunta pode ser dividida em duas etapas. A primeira diz respeito a centralização das observações de cada variável. A matriz em sua forma centrada $\boldsymbol{X}_{\mathrm{c}}$ é dada por:

$$\boldsymbol{X}_{\mathrm{c}} = \left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \ldots & x_{np} - \bar{x}_p \end{bmatrix}, \tag{2.34}$$

em que $\boldsymbol{J}$ é uma matriz quadrada de 1's e $\boldsymbol{X}$ é dada em (2.1). A matriz $\boldsymbol{I} - \left(1/n\right)\boldsymbol{J}$ é chamada de matriz de *centering*.

Para completar a padronização de $\boldsymbol{X}$, na segunda etapa é necessário reduzir cada vetor coluna $\boldsymbol{x}'_{(j)} = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ à sua escala unitária, com $j = 1, 2, \ldots, p$. Para isso, recorre-se a matriz diagonal $\boldsymbol{D}$, dada por:

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{s_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{s_p} \end{bmatrix}, \tag{2.35}$$

em que s_j corresponde ao desvio padrão amostral observado para cada vetor coluna $\boldsymbol{x}'_{(j)} = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ de $\boldsymbol{X}$.

A matriz de observações em sua forma padronizada $\boldsymbol{X}_{\mathrm{p}}$ é obtida pós multiplicando-se $\boldsymbol{X}_{\mathrm{c}}$ por $\boldsymbol{D}$, pois as colunas de $\boldsymbol{X}_{\mathrm{c}}$ são multiplicadas pelo inverso de cada desvio padrão correspondente da diagonal de $\boldsymbol{D}$. Logo,

$$\begin{aligned}\boldsymbol{X}_{\mathrm{p}} = \boldsymbol{X}_{\mathrm{c}}\boldsymbol{D} &= \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \ldots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \ldots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \ldots & x_{np}-\bar{x}_p \end{bmatrix} \begin{bmatrix} \frac{1}{s_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{s_p} \end{bmatrix} \\ &= \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \cdots & \frac{x_{1p}-\bar{x}_p}{s_p} \\ \frac{x_{21}-\bar{x}_1}{s_1} & \frac{x_{22}-\bar{x}_2}{s_2} & \cdots & \frac{x_{2p}-\bar{x}_p}{s_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \frac{x_{n2}-\bar{x}_2}{s_2} & \cdots & \frac{x_{np}-\bar{x}_p}{s_p} \end{bmatrix}. \end{aligned} \tag{2.36}$$

Considere agora a matriz de delineamento $\boldsymbol{X}_{n\times(p+1)}$, dada em (2.5). Um resultado interessante diz respeito a relação entre as matrizes $\boldsymbol{X}$ e $\boldsymbol{X}_{\mathrm{c}}$ com o estimador $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. Esse resultado indica que o estimador de mínimos quadrados é invariante para a mudança de locação das observações, a menos de uma mudança que ocorre no intercepto β_0 do modelo. Neste sentido, queremos mostrar que as estimativas de mínimos quadrados para a matriz não centrada $\boldsymbol{X}$ ou a matriz com os dados centrados $\boldsymbol{X}_{\mathrm{c}}$ são as mesmas, a menos dessa mudança em β_0. Vamos apresentar duas demonstrações algébricas desse fato, no intuito de enfatizar o quanto a abordagem algébrica é laboriosa. Posteriormente, duas demonstrações geométricas, que acreditamos que sejam muito mais simples, serão apresentadas.

Considere inicialmente a matriz de delineamento $\boldsymbol{X}_{n\times(p+1)}$ e o vetor $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ em suas formas particionadas,

$$\boldsymbol{X} = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right] \text{ e } \hat{\boldsymbol{\beta}}_{\mathrm{ols}} = \left[\begin{array}{c} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right], \tag{2.37}$$

em que $\boldsymbol{j}$ é um vetor coluna $(n\times 1)$ de 1's e $\boldsymbol{X}_1$ é a matriz de observações $\boldsymbol{X}_{n\times p}$, dada em (2.1).

Note que $\boldsymbol{X}' = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right]' = \left[\begin{array}{c} \boldsymbol{j} \\ \hline \boldsymbol{X}_1 \end{array}\right]' = \left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_1' \end{array}\right]$. Com base nas equações normais $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = \boldsymbol{X}'\boldsymbol{y}$, tem-se que:

$$\boldsymbol{X}'\boldsymbol{X} = \left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_1' \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{j}'\boldsymbol{j} & \boldsymbol{j}'\boldsymbol{X}_1 \\ \hline \boldsymbol{X}_1'\boldsymbol{j} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right] = \left[\begin{array}{c|c} n & \boldsymbol{j}'\boldsymbol{X}_1 \\ \hline \boldsymbol{X}_1'\boldsymbol{j} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right]. \tag{2.38}$$

$$\boldsymbol{X}'\boldsymbol{y} = \left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_1' \end{array}\right] \boldsymbol{y} = \left[\begin{array}{c} \boldsymbol{j}'\boldsymbol{y} \\ \hline \boldsymbol{X}_1'\boldsymbol{y} \end{array}\right] = \left[\begin{array}{c} \sum_{i=1}^{n} y_i \\ \hline \boldsymbol{X}_1'\boldsymbol{y} \end{array}\right] = \left[\begin{array}{c} n\bar{y} \\ \hline \boldsymbol{X}_1'\boldsymbol{y} \end{array}\right]. \tag{2.39}$$

Diante desses resultados, as equações normais $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} = \boldsymbol{X}'\boldsymbol{y}$ são dadas por:

$$\left[\begin{array}{c|c} n & \boldsymbol{j}'\boldsymbol{X}_1 \\ \hline \boldsymbol{X}_1'\boldsymbol{j} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right] \left[\begin{array}{c} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] = \left[\begin{array}{c} n\bar{y} \\ \hline \boldsymbol{X}_1'\boldsymbol{y} \end{array}\right]. \tag{2.40}$$

Segue então o seguinte sistema de equações,

$$n\hat{\beta}_0 + \boldsymbol{j}'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = n\bar{y}. \tag{2.41}$$

$$\boldsymbol{X}_1'\boldsymbol{j}\hat{\beta}_0 + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y}. \tag{2.42}$$

Em relação a $\boldsymbol{j}'\boldsymbol{X}_1$ vem que:

$$\begin{aligned} \boldsymbol{j}'\boldsymbol{X}_1 &= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{ip} \end{bmatrix} \\ &= \begin{bmatrix} n\bar{x}_1 & n\bar{x}_2 & \dots & n\bar{x}_p \end{bmatrix} \\ &= n \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix} \\ &= n\bar{\boldsymbol{x}}'. \end{aligned} \tag{2.43}$$

Utilizando (2.43), vem para (2.41) que:

$$\begin{aligned} n\hat{\beta}_0 + \boldsymbol{j}'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = n\bar{y} &\Rightarrow n\hat{\beta}_0 + n\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = n\bar{y} \\ &\Rightarrow \hat{\beta}_0 + \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = \bar{y} \\ &\Rightarrow \hat{\alpha} = \bar{y}, \end{aligned} \tag{2.44}$$

em que $\hat{\alpha} = \hat{\beta}_0 + \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1$.

Como $\boldsymbol{X}_1'\boldsymbol{j} = n\bar{\boldsymbol{x}}$, de (2.42) segue que:

$$\boldsymbol{X}_1'\boldsymbol{j}\hat{\beta}_0 + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y} \Rightarrow n\bar{\boldsymbol{x}}\hat{\beta}_0 + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y}. \tag{2.45}$$

Agora precisamos resolver o sistema de equações normais para o modelo centrado, que é dado por:

$$\left[\begin{array}{c} \boldsymbol{j} \\ \hline \boldsymbol{X}_{\mathrm{c}} \end{array}\right]' \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right] \left[\begin{array}{c} \hat{\alpha} \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] = \left[\begin{array}{c} \boldsymbol{j} \\ \hline \boldsymbol{X}_{\mathrm{c}} \end{array}\right]' \boldsymbol{y}. \tag{2.46}$$

Para $\left[\begin{array}{c} \boldsymbol{j} \\ \hline \boldsymbol{X}_{\mathrm{c}} \end{array}\right]' \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right]$ observa-se que:

$$\left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_{\mathrm{c}}' \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{j}'\boldsymbol{j} & \boldsymbol{j}'\boldsymbol{X}_{\mathrm{c}} \\ \hline \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} \end{array}\right] = \left[\begin{array}{c|c} n & \boldsymbol{0}' \\ \hline \boldsymbol{0} & \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} \end{array}\right], \tag{2.47}$$

em que $\boldsymbol{j}'\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{0}'$ pelo fato de cada coluna de $\boldsymbol{X}_{\mathrm{c}}$ somar zero.

Por sua vez, para $\left[\begin{array}{c} \boldsymbol{j} \\ \hline \boldsymbol{X}_{\mathrm{c}} \end{array}\right]' \boldsymbol{y}$ tem-se que:

$$\left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_{\mathrm{c}}' \end{array}\right] \boldsymbol{y} = \left[\begin{array}{c} \boldsymbol{j}'\boldsymbol{y} \\ \hline \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \end{array}\right] = \left[\begin{array}{c} \sum_{i=1}^{n} y_i \\ \hline \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \end{array}\right] = \left[\begin{array}{c} n\bar{y} \\ \hline \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \end{array}\right]. \tag{2.48}$$

Dessa maneira, substituindo (2.47) e (2.48) em (2.46), segue que o sistema de equações normais para o modelo centrado fica da seguinte forma:

$$\left[\begin{array}{c|c} n & \boldsymbol{0}' \\ \hline \boldsymbol{0} & \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} \end{array}\right] \left[\begin{array}{c} \hat{\alpha} \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] = \left[\begin{array}{c} n\bar{y} \\ \hline \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \end{array}\right]. \tag{2.49}$$

Assim, tem-se o seguinte sistema de equações:

$$n\hat{\alpha} = n\bar{y}. \tag{2.50}$$

$$\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y}. \tag{2.51}$$

Pode-se observar que a igualdade em (2.50) é a mesma igualdade resultante em (2.44), ou seja, $\hat{\alpha} = \bar{y}$. Como $\hat{\alpha} = \hat{\beta}_0 + \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1$, tem-se que $\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = \hat{\alpha} - \hat{\beta}_0$. Queremos mostrar que (2.51)

é igual a (2.45), mas antes precisamos de alguns resultados. Sabemos que $\boldsymbol{X}_{\mathrm{c}} = \left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1$. Como a matriz $\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]$ é uma matriz de projeção simétrica, isto é, essa matriz é simétrica e idempotente, resulta que:

$$\begin{aligned}
\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} &= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]^2\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{J}\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{j}\boldsymbol{j}'\boldsymbol{X}_1 \\
&= \boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\left(\boldsymbol{X}_1'\boldsymbol{j}\right)\left(\boldsymbol{X}_1'\boldsymbol{j}\right)' \\
&= \boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\left(n\bar{\boldsymbol{x}}\right)\left(n\bar{\boldsymbol{x}}\right)' \\
&= \boldsymbol{X}_1'\boldsymbol{X}_1 - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'. \qquad (2.52)
\end{aligned}$$

Temos também que:

$$\begin{aligned}
\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} &= \left[\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1\right]'\boldsymbol{y} \\
&= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]'\boldsymbol{y} \\
&= \boldsymbol{X}_1'\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{y} \\
&= \boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{J}\boldsymbol{y} \\
&= \boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{j}\boldsymbol{j}'\boldsymbol{y} \\
&= \boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}\left(\boldsymbol{X}_1'\boldsymbol{j}\right)\left(\boldsymbol{j}'\boldsymbol{y}\right) \\
&= \boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}n\bar{\boldsymbol{x}}n\bar{y} \\
&= \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y}. \qquad (2.53)
\end{aligned}$$

Utilizando os resultados de (2.52) e (2.53) em (2.51) vem que:

$$\begin{aligned}
\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} &\Rightarrow \left(\boldsymbol{X}_1'\boldsymbol{X}_1 - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'\right)\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y} \\
&\Rightarrow \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y} \\
&\Rightarrow \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - n\bar{\boldsymbol{x}}\left(\hat{\alpha} - \hat{\beta}_0\right) = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y} \\
&\Rightarrow \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - n\bar{\boldsymbol{x}}\hat{\alpha} + n\bar{\boldsymbol{x}}\hat{\beta}_0 = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y} \\
&\Rightarrow \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - n\bar{\boldsymbol{x}}\bar{y} + n\bar{\boldsymbol{x}}\hat{\beta}_0 = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y} \\
&\Rightarrow n\bar{\boldsymbol{x}}\hat{\beta}_0 + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y}, \qquad (2.54)
\end{aligned}$$

o que conclui a igualdade entre (2.51) e (2.45).

Portanto, podemos concluir que os estimadores $\hat{\alpha} = \bar{y}$ e $\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_c\right)^{-1}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y}$ obtidos a partir de $\boldsymbol{X}_{\mathrm{c}}$ são os mesmos de $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, que são obtidos de $\boldsymbol{X}$.

■

A seguir será apresentada uma segunda forma algébrica de se demonstrar esse resultado. Dado $\boldsymbol{X} = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right]$, ou seja, $\boldsymbol{X}$ em sua forma particionada, precisamos encontrar a inversa de $\boldsymbol{X}'\boldsymbol{X}$. Primeiramente, temos que:

$$\boldsymbol{X}'\boldsymbol{X} = \left[\begin{array}{c} \boldsymbol{j}' \\ \hline \boldsymbol{X}_1' \end{array}\right]\left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{j}'\boldsymbol{j} & \boldsymbol{j}'\boldsymbol{X}_1 \\ \hline \boldsymbol{X}_1'\boldsymbol{j} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right] = \left[\begin{array}{c|c} n & n\bar{\boldsymbol{x}}' \\ \hline n\bar{\boldsymbol{x}} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right]. \qquad (2.55)$$

Dessa maneira, estamos interessados em obter $\left[\begin{array}{c|c} n & n\bar{\boldsymbol{x}}' \\ \hline n\bar{\boldsymbol{x}} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right]^{-1}$, ou seja, precisamos da expressão para a inversa de uma matriz particionada (RENCHER; SHAALJE, 2008, p. 23). Se uma matriz $\boldsymbol{A}$ é simétrica, não-singular e particionada como $\boldsymbol{A} = \left[\begin{array}{c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \hline \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array}\right]$ e se $\boldsymbol{B} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$, então supondo que $\boldsymbol{A}_{11}^{-1}$ e $\boldsymbol{B}^{-1}$ existem, a inversa de $\boldsymbol{A}$ é dada por:

$$\boldsymbol{A}^{-1} = \left[\begin{array}{c|c} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}^{-1} \\ \hline -\boldsymbol{B}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{B}^{-1} \end{array}\right]. \qquad (2.56)$$

Seja $\boldsymbol{A}_{11} = n$, $\boldsymbol{A}_{12} = n\bar{\boldsymbol{x}}'$, $\boldsymbol{A}_{21} = n\bar{\boldsymbol{x}}$ e $\boldsymbol{A}_{22} = \boldsymbol{X}_1'\boldsymbol{X}_1$. Como $\boldsymbol{X}_c'\boldsymbol{X}_c = \boldsymbol{X}_1'\boldsymbol{X}_1 - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'$, então:

$$\boldsymbol{B} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = \boldsymbol{X}_1'\boldsymbol{X}_1 - n\bar{\boldsymbol{x}}n^{-1}n\bar{\boldsymbol{x}}' = \boldsymbol{X}_1'\boldsymbol{X}_1 - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' = \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}. \tag{2.57}$$

Segue então que a inversa de $\boldsymbol{X}'\boldsymbol{X}$ é dada por:

$$\begin{aligned}
\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} &= \left[\begin{array}{c|c} n & n\bar{\boldsymbol{x}}' \\ \hline n\bar{\boldsymbol{x}} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{array}\right]^{-1} \\
&= \left[\begin{array}{c|c} n^{-1} + n^{-1}n\bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}n\bar{\boldsymbol{x}}n^{-1} & -n^{-1}n\bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \\ \hline -\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}n\bar{\boldsymbol{x}}n^{-1} & \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \end{array}\right] \\
&= \left[\begin{array}{c|c} n^{-1} + \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \\ \hline -\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}} & \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \end{array}\right].
\end{aligned} \tag{2.58}$$

Uma vez que $\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} = \boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y}$ e utilizando $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ em sua forma particionada, segue que o estimador de mínimos quadrados é dado por:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{ols}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left[\begin{array}{c|c} n^{-1} + \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \\ \hline -\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}} & \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1} \end{array}\right]\left[\begin{array}{c} n\bar{y} \\ \hline \boldsymbol{X}_1'\boldsymbol{y} \end{array}\right] \\
&= \left[\begin{array}{c} \bar{y} + \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}}n\bar{y} - \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \\ \hline -\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\bar{\boldsymbol{x}}n\bar{y} + \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \end{array}\right] \\
&= \left[\begin{array}{c} \bar{y} - \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y}\right) \\ \hline \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y} - n\bar{\boldsymbol{x}}\bar{y}\right) \end{array}\right] \\
&= \left[\begin{array}{c} \bar{y} - \bar{\boldsymbol{x}}'\left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \\ \hline \left(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\right)^{-1}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y} \end{array}\right] \\
&= \left[\begin{array}{c} \bar{y} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] \\
&= \left[\begin{array}{c} \hat{\alpha} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] \\
&= \left[\begin{array}{c} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right].
\end{aligned} \tag{2.59}$$

Logo, o estimador de mínimos quadrados obtido a partir da matriz $\boldsymbol{X}$ é igual ao estimador obtido utilizando-se a matriz centrada $\boldsymbol{X}_{\mathrm{c}}$, com o ajuste:

$$\hat{\beta}_0 = \hat{\alpha} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = \bar{y} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_1 = \bar{y} - \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}} = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \ldots - \hat{\beta}_p\bar{x}_p, \tag{2.60}$$

o que conclui a demonstração.

■

Vamos agora apresentar duas demonstrações utilizando a abordagem geométrica, considerando duas situações: uma com e outra sem o intercepto β_0. Vejamos inicialmente o caso em que se tem intercepto. O parâmetro β_0 pode ser pensado como um parâmetro relativo a uma covariável que permanece constante em todas as observações. Nesse sentido, será utilizada a seguinte notação para o vetor de parâmetros $\boldsymbol{\beta}_{(p+1)\times 1}$:

$$\boldsymbol{\beta} = \left[\begin{array}{c} \beta_0 \\ \hline \boldsymbol{\beta}_{p\times 1} \end{array}\right] = \left[\begin{array}{c} \beta_0 \\ \hline \boldsymbol{\beta}_1 \end{array}\right]. \tag{2.61}$$

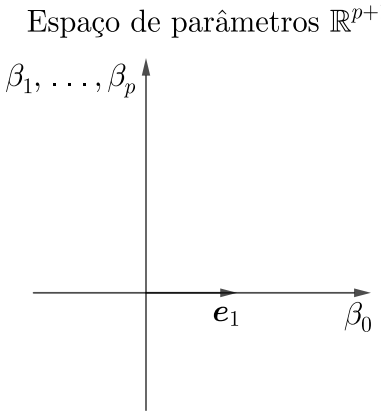
Novamente, a matriz $\boldsymbol{X}_{n\times(p+1)}$ é expressa em sua forma particionada $\boldsymbol{X} = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right]$, sendo $\boldsymbol{X}_1$ dada em (2.1). Considerando o espaço de parâmetros $\mathbb{R}^{p+1}$ e o eixo definido pelo vetor canônico $\boldsymbol{e}_1 = (1,0,\ldots,0)$ como relativo a β_0 (FIGURA 2.5), tem-se que:

$$\boldsymbol{X}\boldsymbol{e}_1 = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right]\boldsymbol{e}_1 = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \boldsymbol{j}. \tag{2.62}$$

Aqui fica clara a interpretação dada ao parâmetro β_0 do modelo de regressão. Ao aplicar a matriz $\boldsymbol{X}$ no vetor $\boldsymbol{e}_1$, relativo a β_0, o resultado é o vetor de 1's. O parâmetro β_0 representa uma parte constante e fixa para cada valor y_i no modelo de regressão.

De (2.62) assume-se que $\boldsymbol{j} \in \mathrm{Im}(\boldsymbol{X})$, ou seja, a hipótese do modelo linear possuir intercepto. Como $\boldsymbol{j} \in \mathrm{Im}(\boldsymbol{X})$, a $\mathrm{Im}(\boldsymbol{X})$ contém o subespaço gerado pelo vetor de 1's, $\boldsymbol{j}$. O subespaço p-dimensional da $\mathrm{Im}(\boldsymbol{X})$, definido pelos vetores ortogonais a $\boldsymbol{j}$, será denotado por $\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$. O

Figura 2.5 – Espaço de parâmetros $\mathbb{R}^{p+1}$, com o eixo definido pelo vetor canônico $\boldsymbol{e}_1 = (1,0,\ldots,0)$ como relativo a β_0.



Fonte: Do autor (2020).

subespaço de $\mathbb{R}^n$, complemento ortogonal da $\mathrm{Im}(\boldsymbol{X})$, será denotado por $\mathrm{Im}(\boldsymbol{X})^{\perp}$ e o subespaço unidimensional gerado pelo vetor $\boldsymbol{j}$ por $span(\boldsymbol{j})$.

Logo, o espaço de dados $\mathbb{R}^n$ se decompõe como a seguinte soma direta:

$$\mathbb{R}^n = \mathrm{Im}(\boldsymbol{X}) \oplus \mathrm{Im}(\boldsymbol{X})^{\perp} = span(\boldsymbol{j}) \oplus \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X}) \oplus \mathrm{Im}(\boldsymbol{X})^{\perp}. \tag{2.63}$$

Dessa forma, todo vetor $\boldsymbol{y} \in \mathbb{R}^n$ é escrito de forma única da seguinte maneira,

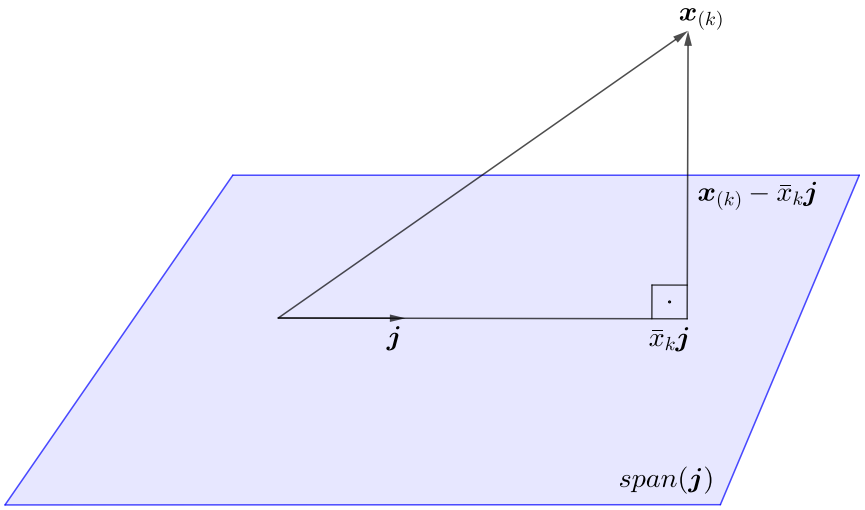
$$\boldsymbol{y} = \boldsymbol{y}_0 + \boldsymbol{y}_1 + \boldsymbol{y}_2, \tag{2.64}$$

em que $\boldsymbol{y}_0 \in span(\boldsymbol{j})$, $\boldsymbol{y}_1 \in \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$ e $\boldsymbol{y}_2 \in \mathrm{Im}(\boldsymbol{X})^{\perp}$.

Note que cada $\boldsymbol{y}_i$ é a projeção ortogonal de $\boldsymbol{y}$ no respectivo subespaço, para $i = 0,1,2$. O subespaço $\mathrm{Im}(\boldsymbol{X})$ é gerado por $\boldsymbol{j}$ e por cada $\boldsymbol{x}_{(k)}$, que representa uma coluna da matriz $\boldsymbol{X}_1$, com $k = 1,2,\ldots,p$. Projetando-se cada coluna $\boldsymbol{x}_{(k)}$ no subespaço $span(\boldsymbol{j})$ (FIGURA 2.6), resulta que os vetores obtidos por meio dessa projeção, $\left(\boldsymbol{x}_{(1)} - \bar{x}_1\boldsymbol{j}, \boldsymbol{x}_{(2)} - \bar{x}_2\boldsymbol{j}, \ldots, \boldsymbol{x}_{(p)} - \bar{x}_p\boldsymbol{j}\right)$, definem a matriz $\boldsymbol{X}_{\mathrm{c}}$ que é dada por:

$$\boldsymbol{X}_{\mathrm{c}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}. \tag{2.65}$$

Figura 2.6 – Projeção ortogonal do vetor $\boldsymbol{x}_{(k)}$ na direção do vetor $\boldsymbol{j}$, ou seja, no subespaço $span\,(\boldsymbol{j})$.



Fonte: Do autor (2020).

Tomando-se $\boldsymbol{X} = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right]$, segue que essa matriz pode ser reescrita da seguinte forma:

$$
\begin{aligned}
\boldsymbol{X} &= \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_1 \end{array}\right] \\
&= \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} + \begin{bmatrix} 0 & x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ 0 & x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}.
\end{aligned} \tag{2.66}
$$

Considere que o vetor aleatório $\boldsymbol{y}$ foi observado e que β_0 e $\boldsymbol{\beta}_1$ foram estimados. Sabe-se que as equações normais na forma geométrica são dadas por $P_{\mathrm{Im}(\boldsymbol{X})}\,(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Dessa maneira, as equações normais para o modelo não centrado são:

$$
\begin{aligned}
P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}\begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} \\
&= \left(\begin{bmatrix} 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} + \begin{bmatrix} 0 & x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ 0 & x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \right) \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} + \begin{bmatrix} 0 & x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ 0 & x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} \\
&= \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{j}\bar{\boldsymbol{x}}' \end{array}\right] \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} + \left[\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right] \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\boldsymbol{\beta}}_1 \end{bmatrix} \\
&= \left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} + \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1, \qquad (2.67)
\end{aligned}
$$

em que $\bar{\boldsymbol{x}}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$, $\left(\beta_0 + \boldsymbol{\beta}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} \in span(\boldsymbol{j})$ e $\boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta}_1 \in \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$.

Vejamos agora o caso relacionado ao modelo centrado. Levando em consideração somente a parte determinística de um modelo de regressão, $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, o modelo centrado pode ser derivado da seguinte forma:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\
&= \beta_0 + \beta_1 x_{i1} + \beta_1\bar{x}_1 - \beta_1\bar{x}_1 + \ldots + \beta_p x_{ip} + \beta_p\bar{x}_p - \beta_p\bar{x}_p \\
&= \beta_0 + \beta_1\bar{x}_1 + \ldots + \beta_p\bar{x}_p + \beta_1 x_{i1} - \beta_1\bar{x}_1 + \ldots + \beta_p x_{ip} - \beta_p\bar{x}_p \\
&= \beta_0 + \beta_1\bar{x}_1 + \ldots + \beta_p\bar{x}_p + \beta_1(x_{i1} - \bar{x}_1) + \ldots + \beta_p(x_{ip} - \bar{x}_p) \\
&= \beta_0 + \boldsymbol{\beta}_1'\bar{\boldsymbol{x}} + \boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta}_1 \\
&= \alpha + \boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta}_1, \qquad (2.68)
\end{aligned}
$$

em que $\alpha = \beta_0 + \boldsymbol{\beta}_1'\bar{\boldsymbol{x}}$, para $i = 1, 2, \ldots, n$.

Novamente, suponha que o vetor $\boldsymbol{y}$ foi observado e que β_0 e $\boldsymbol{\beta}_1$ foram estimados. As equações normais do modelo centrado são dadas por $P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}$. Dessa maneira,

$$P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}} = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right] \left[\begin{array}{c} \hat{\alpha} \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{j} & \boldsymbol{X}_{\mathrm{c}} \end{array}\right] \left[\begin{array}{c} \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}} \\ \hline \hat{\boldsymbol{\beta}}_1 \end{array}\right] = \left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} + \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1, \tag{2.69}$$

em que $\left(\beta_0 + \boldsymbol{\beta}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} \in span(\boldsymbol{j})$ e $\boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta}_1 \in \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$.

A partir dos resultados obtidos em (2.67) e (2.69) pode-se verificar que os modelos centrado e não centrado possuem o mesmo sistema de equações normais. Além disso, como $\boldsymbol{y}_0 \in span(\boldsymbol{j})$ e $\boldsymbol{y}_1 \in \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$, segue que as equações normais para ambos os modelos podem ser expressas como:

$$\left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} = \boldsymbol{y}_0 = \bar{y}\boldsymbol{j}. \tag{2.70}$$

$$\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 = \boldsymbol{y}_1. \tag{2.71}$$

Novamente fica provado que os estimadores dos modelos centrado e não centrado são os mesmos, com um ajuste na estimativa no intercepto $\hat{\beta}_0$ do modelo centrado:

$$\hat{\beta}_0 = \hat{\alpha} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 - \ldots - \beta_p\bar{x}_p = \bar{y} - \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}, \tag{2.72}$$

em que a estimativa $\hat{\alpha}$ é obtida de um estimador de α em (2.68).

■

Foi possível verificar ao se aplicarem as matrizes de transformações $\boldsymbol{X}$ e $\boldsymbol{X}_{\mathrm{c}}$ no estimador $\hat{\boldsymbol{\beta}}$, que o vetor observado $\boldsymbol{y}$ foi projetado ortogonalmente na imagem de $\boldsymbol{X}$ como uma combinação linear de vetores dos subespaços $span(\boldsymbol{j})$ e $\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$. Mas o que ocorreria se a matriz de transformação $\boldsymbol{X}_{\mathrm{c}}$ para o modelo centrado for aplicada no vetor $\hat{\boldsymbol{\beta}}_1$, que não considera o intercepto $\hat{\beta}_0$.

Como o espaço de dados $\mathbb{R}^n$ se decompõe como $\mathbb{R}^n = \mathrm{Im}(\boldsymbol{X}) \oplus \mathrm{Im}(\boldsymbol{X})^{\perp} = span(\boldsymbol{j}) \oplus \mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X}) \oplus \mathrm{Im}(\boldsymbol{X})^{\perp}$, segue que o vetor de dados $\boldsymbol{y}$ pode ser reescrito como:

$$\boldsymbol{y} = P_{span(\boldsymbol{j})}(\boldsymbol{y}) + P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}) + P_{\mathrm{Im}(\boldsymbol{X})^{\perp}}(\boldsymbol{y}) = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) + P_{\mathrm{Im}(\boldsymbol{X})^{\perp}}(\boldsymbol{y}). \tag{2.73}$$

De (2.73) decorre que:

$$P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}) = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) + P_{\mathrm{Im}(\boldsymbol{X})^{\perp}}(\boldsymbol{y}) - P_{span(\boldsymbol{j})}(\boldsymbol{y}) - P_{\mathrm{Im}(\boldsymbol{X})^{\perp}}(\boldsymbol{y})$$
$$= P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) - P_{span(\boldsymbol{j})}(\boldsymbol{y}). \tag{2.74}$$

De (2.67), (2.70) e (2.74) vem que:

$$\begin{aligned}\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 &= \bar{y}\boldsymbol{j} + \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 - \bar{y}\boldsymbol{j} \\ &= \left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}\right)\boldsymbol{j} + \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_1 - \bar{y}\boldsymbol{j} \\ &= P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) - P_{span(\boldsymbol{j})}(\boldsymbol{y}) \\ &= P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}).\end{aligned} \tag{2.75}$$

Sobre a questão suscitada, segue que a transformação $\boldsymbol{X}_{\mathrm{c}}$ quando aplicada ao vetor $\hat{\boldsymbol{\beta}}_1$ faz com que o vetor de dados $\boldsymbol{y}$ no espaço de observações seja projetado ortogonalmente no subespaço $\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})$. Diante disso, a segunda demonstração geométrica, para o caso em que não se tem o intercepto é ainda mais simples. Considere então os dois modelos lineares da forma $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ e $\boldsymbol{y} = \boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, em que ambos não possuem o intercepto β_0. As equações normais dos modelos não centrado $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ e centrado $\boldsymbol{y} = \boldsymbol{X}_{\mathrm{c}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ são dadas, respectivamente, por:

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}). \tag{2.76}$$

$$\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}^{\mathrm{c}}_{\mathrm{ols}} = P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}), \tag{2.77}$$

em que $\hat{\boldsymbol{\beta}}^{\mathrm{c}}_{\mathrm{ols}}$ é o estimador de mínimos quadrados para o modelo centrado.

Para o modelo ajustado $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ observe que:

$$\frac{1}{n}\boldsymbol{J}\hat{\boldsymbol{y}} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}\hat{y}_i \\ \sum_{i=1}^{n}\hat{y}_i \\ \vdots \\ \sum_{i=1}^{n}\hat{y}_i \end{bmatrix} = \frac{1}{n}\begin{bmatrix} n\bar{y} \\ n\bar{y} \\ \vdots \\ n\bar{y} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} = \bar{y}\boldsymbol{j}. \tag{2.78}$$

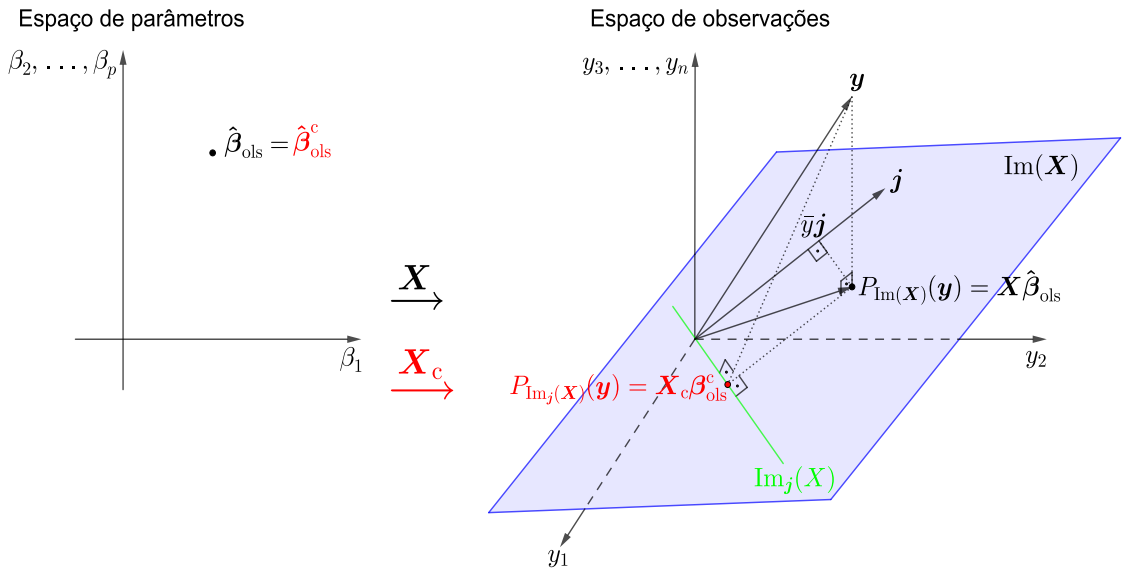
Como $\boldsymbol{X}_1$ é a matriz $\boldsymbol{X}$ dada em (2.1), utilizando os resultados em (2.75), (2.76), (2.77) e (2.78) segue que:

$$
\begin{aligned}
\boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} &= \left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{\mathrm{ols}} \\
&= \left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} \\
&= \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}-\frac{1}{n}\boldsymbol{J}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} \\
&= \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}-\frac{1}{n}\boldsymbol{J}\hat{\boldsymbol{y}} \\
&= \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}-\bar{y}\boldsymbol{j} \\
&= P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})-P_{span(\boldsymbol{j})}(\boldsymbol{y}) \\
&= P_{\mathrm{Im}_{\boldsymbol{j}}(\boldsymbol{X})}(\boldsymbol{y}) \\
&= \boldsymbol{X}_{\mathrm{c}}\hat{\boldsymbol{\beta}}^{\mathrm{c}}_{\mathrm{ols}}, \qquad (2.79)
\end{aligned}
$$

o que implica que $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = \hat{\boldsymbol{\beta}}^{\mathrm{c}}_{\mathrm{ols}}$.

Segue mais uma vez o resultado desejado para os modelos centrado e não centrado. Na Figura 2.7 apresenta-se uma representação geométrica das transformações $\boldsymbol{X}$ e $\boldsymbol{X}_{\mathrm{c}}$ sobre a imagem, ou seja, $\boldsymbol{X}$ e $\boldsymbol{X}_{\mathrm{c}}$ são isometrias entre o espaço de parâmetros e o espaço de observações. Destaca-se também nessa figura o fato dos estimadores $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ e $\hat{\boldsymbol{\beta}}^{\mathrm{c}}_{\mathrm{ols}}$ serem iguais na ausência do intercepto β_0.

Figura 2.7 – Representação das transformações $\boldsymbol{X}$ e $\boldsymbol{X}_{\mathrm{c}}$ entre o espaço de parâmetros e o espaço de observações para o caso em que não se tem o intercepto β_0.



Fonte: Do autor (2020).

■

O resultado de que o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é invariante para mudanças de locação das observações, ou seja, de que as estimativas de mínimos quadrados para a matriz original $\boldsymbol{X}$ (não centrada) e a matriz quando os dados estão centrados $\boldsymbol{X}_{\text{c}}$ são as mesmas, a menos de uma mudança no intercepto β_0, encontra-se como um exercício em Rencher e Shaalje (2008, p. 179). Neste ponto, uma pergunta pertinente a ser abordada é: qual a relação desse resultado com os estimadores *Ridge*, LASSO e *Elastic Net*?

No primeiro parágrafo desta subseção foi dito que em análises utilizando esses estimadores é importante que as variáveis preditoras estejam em sua forma padronizada. Em algumas situações em que o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ apresenta deficiências, como as que serão mencionadas na próxima seção, os estimadores *Ridge*, LASSO e *Elastic Net* surgiram como alternativas, como procedimentos que penalizam de diferentes formas as estimativas de mínimos quadrados. A padronização das variáveis para a utilização desses estimadores não é só importante para o tratamento de variáveis que possam ter locação e escalas diferentes, em razão de suas unidades de medidas, mas se relaciona também com o fato de o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ ser invariante (a menos do intercepto) quando os dados estão em sua forma centrada. Logo, a padronização para utilização desses métodos pode ser justificada pelo fato das variáveis poderem ser tratadas independentemente de suas unidades de medida, ou seja, para que possam receber um tratamento estatístico em uma mesma escala unitária e pelo fato das estimativas de mínimos quadrados serem invariantes por locação, a menos de uma mudança que ocorre no parâmetro β_0.

2.4 O problema da multicolinearidade

Segundo Montgomery, Peck e Vining (2012, p. 285), o uso e a interpretação de um modelo de regressão múltipla geralmente dependem, explícita ou implicitamente, das estimativas dos coeficientes de regressão individualmente. Os exemplos de inferências realizados frequentemente incluem:

1. Identificação dos efeitos relativos das variáveis regressoras.

2. Predição e/ou cálculo de estimativas.

3. Seleção de um conjunto de variáveis para o modelo.

Ainda segundo os autores, quando não existe uma relação linear entre as variáveis regressoras, elas são ortogonais. Quando os regressores são ortogonais, inferências como as que foram citadas anteriormente podem ser feitas com relativa facilidade. Mas com frequência, na

maioria das aplicações envolvendo os modelos de regressão, as variáveis regressoras não são ortogonais. Em outras situações as variáveis regressoras têm uma relação linear quase perfeita e, nesses casos, as inferências baseadas no modelo de regressão podem ser enganosas ou mesmo erradas. Quando existem dependências quase lineares entre as variáveis regressoras, diz-se que existe o problema da multicolinearidade.

Conforme Saleh, Arashi e Kibria (2019, p. 3), a multicolinearidade pode ser entendida como a existência de relações quase lineares entre as variáveis preditoras. A multicolinearidade ocorre quando a correlação entre as variáveis regressoras é muito alta. Isso se torna um inconveniente por causar problemas no ajuste do modelo e na interpretação dos resultados. Sobre essa perspectiva, a multicolinearidade entre as variáveis preditoras pode ser abordada sobre dois aspectos. O primeiro sendo concernente aos tipos de multicolinearidade e o segundo relacionado aos eventuais problemas que decorrem da multicolinearidade.

Sobre o primeiro aspecto, existem dois tipos básicos de multicolinearidade:

Multicolinearidade dos dados: esse tipo de multicolinearidade está presente nos próprios dados e não está intrinsecamente relacionada ao modelo teórico. Geralmente ocorre em experimentos mal planejados pelo pesquisador.

Multicolinearidade estrutural: esse tipo de multicolinearidade ocorre quando um termo inserido no modelo é criado a partir de um ou mais termos que já existem. Por exemplo, a partir de uma variável X podem ser criadas as variáveis X^2 e $\sqrt{X}$ e, nesse caso, a correlação entre as novas variáveis e a original é clara. Em resumo, as novas variáveis criadas e inseridas são um subproduto do modelo formulado inicialmente, em vez de estar presente nos próprios dados.

Montgomery, Peck e Vining (2012, p. 286-288) destacam cinco fontes para a ocorrência da multicolinearidade: *i*) coleta de dados, *ii*) restrições físicas, *iii*) modelo superparametrizado ($p \gg n$), *iv*) escolha ou especificação do modelo e *v*) *outliers*. Dessa maneira, pode-se observar que as fontes *i*), *ii*) e *v*) estão relacionadas a multicolinearidade dos dados, enquanto que *iii*) e *iv*) estão relacionadas a multicolinearidade estrutural.

Com relação ao segundo aspecto, antes será comentado brevemente sobre um importante objetivo em uma análise de regressão. Sabe-se que o principal objetivo em uma análise de regressão é isolar o relacionamento entre cada uma das variáveis regressoras $(X_1, X_2, \ldots, X_p)$ e a variável resposta Y. Cada um dos coeficientes $(\beta_0, \beta_1, \ldots, \beta_p)$, que estabelecem a relação

funcional, indica a mudança média na variável resposta para cada mudança de uma unidade na variável regressora, quando as demais variáveis são mantidas constantes. Todavia, em um cenário envolvendo multicolinearidade com alta correlação entre as variáveis preditoras, não é complicado de se conjecturar que a mudança de uma variável estará associada a mudanças de uma ou mais variáveis. Nesse caso, quanto mais forte for a correlação, mais difícil é isolar os efeitos de uma variável em relação a outra sobre a variável resposta. Logo, modelar a relação entre cada variável regressora e a variável resposta de forma independente torna-se mais difícil, pois os preditores podem atuar conjuntamente em relação a variável resposta de interesse, em razão da forte associação entre elas.

Diante disso, os maiores entraves devidos à presença de multicolinearidade na modelagem estão:

1. As estimativas dos coeficientes $(\beta_0, \beta_1, \ldots, \beta_p)$ podem variar muito com relação a presença ou não de algumas variáveis regressoras no modelo. Os coeficientes tornam-se muito sensíveis a pequenas mudanças do modelo.

2. A multicolinearidade reduz a precisão dos coeficientes estimados, o que enfraquece o poder estatístico do modelo de regressão. Nesse caso, os valores-p tornam-se não confiáveis para identificar as variáveis regressoras que são estatisticamente significativas.

Em acréscimo, Saleh, Arashi e Kibria (2019, p. 4) afirmam que a multicolinearidade pode tornar as estimativas dos coeficientes de regressão imprecisas, pode inflar os erros padrão dos coeficientes de regressão, esvaziar os testes t parciais para os coeficientes de regressão, fornecer valores-p falsos (e não significativos) e degradar a previsibilidade do modelo. A multicolinearidade também pode provocar mudanças nos sinais das estimativas dos coeficientes.

2.4.1 A multicolinearidade e o estimador de mínimos quadrados

O método de mínimos quadrados é frequentemente utilizado para estimar os parâmetros de um modelo de regressão linear, consistindo basicamente em se minimizar a soma dos quadrados das diferenças entre os valores preditos e os valores observados das variáveis dependentes. Um ponto interessante sobre a abordagem de mínimos quadrados para estimar os β's é que nenhuma suposição sobre a distribuição de $\boldsymbol{y}$ é necessária na obtenção dos estimadores.

Apesar do estimador de mínimos quadrados ser o estimador linear não viesado de variância mínima, essa abordagem apresenta problemas em pelo menos duas situações. Uma delas

é quando ocorre colinearidade. Nesse caso, pode-se ter a ocorrência de duas variáveis altamente correlacionadas. Esse fato acarreta que pelo menos um dos autovalores da matriz $\boldsymbol{X}'\boldsymbol{X}$ seja próximo de zero. Como a matriz de covariâncias do estimador de mínimos quadrados $\hat{\boldsymbol{\beta}}_{\text{ols}}$, dada por $\text{cov}\left[\hat{\boldsymbol{\beta}}_{\text{ols}}\right] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, depende dos inversos dos autovalores, isso pode implicar que a variância de algum dos dos β's seja grande, inviabilizando seu uso. A outra situação ocorre quando $p > n$, ou seja, quando o número de variáveis é maior do que o número de dados, situação comum em estudos genéticos. Nesse caso, a matriz $\boldsymbol{X}'\boldsymbol{X}$ é singular, o que leva à necessidade da utilização de inversas generalizadas.

Em situações como as que foram mencionadas surge a necessidade de métodos alternativos para a estimação dos parâmetros de um modelo de regressão linear. Uma possibilidade é permitir a utilização de uma distância que não seja a mínima. Esse procedimento recebe o nome genérico de método de mínimos quadrados penalizados. Isso significa que o estimador de mínimos quadrados continua a ser utilizado, mas agora é imposta uma restrição para os coeficientes de regressão. No sentido de se explicitar algumas penalizações usualmente utilizadas e conhecidas na literatura, mais uma vez será abordada a geometria da regressão linear.

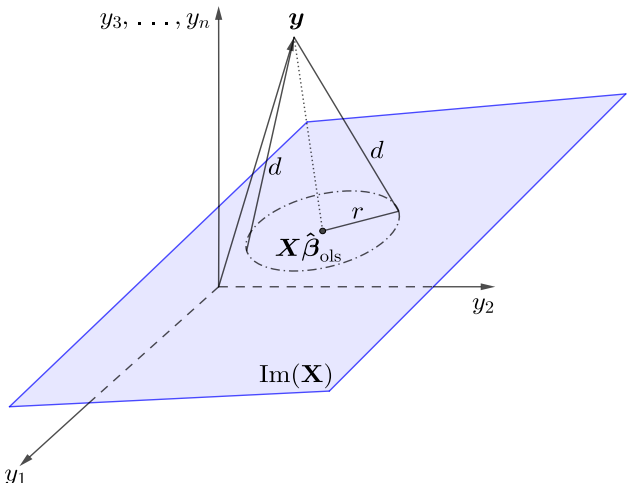
Uma vez observado o vetor de dados $\boldsymbol{y}$, considere no subespaço $\text{Im}(\boldsymbol{X})$ todos os vetores $\boldsymbol{X}\boldsymbol{\beta}$ que estão a uma distância d desse vetor. Dessa forma, observando que $\|\boldsymbol{v}\|^2 = \boldsymbol{v}'\boldsymbol{v}$ (para um vetor $\boldsymbol{v}$ qualquer), então:

$$
\begin{aligned}
d^2 &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= \left\| \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right) \right\|^2 \\
&= \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right) \right]' \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right) \right] \\
&= \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right)' + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right)' \right] \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right) \right] \\
&= \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) + \boldsymbol{0} + \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right)' \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta}\right) \\
&= \left\| \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} \right\|^2 + \left\| \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{X}\boldsymbol{\beta} \right\|^2, \qquad (2.80)
\end{aligned}
$$

em que o vetor nulo na quinta igualdade se justifica pelo fato do vetor $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}$ ser perpendicular ao subespaço $\text{Im}(\boldsymbol{X})$, ou seja, $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}$ é ortogonal a qualquer vetor que pertença a $\text{Im}(\boldsymbol{X})$.

Como $\left\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right\|$ é um valor fixo, os vetores $\boldsymbol{X}\boldsymbol{\beta}$ satisfazem a equação $\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}-\boldsymbol{X}\boldsymbol{\beta}\right\| = r$ (com r constante) e formam uma hiperesfera de raio r no subespaço $\text{Im}(\boldsymbol{X})$, como descrito na Figura 2.8. Os vetores nessa hiperesfera, do ponto de vista de distância aos dados, são indistintos. Portanto, o processo de estimação, isto é, a escolha de um estimador particular para o vetor de parâmetros $\boldsymbol{\beta}$ deve seguir algum tipo de restrição. Novamente, a chave para se entender o processo de estimação das teorias *Ridge*, LASSO e *Elastic Net* é a geometria. Que propriedades possuem os vetores $\boldsymbol{\beta}$'s que são levados pela transformação linear $\boldsymbol{X}$ nessa hiperesfera?

Figura 2.8 – Vetores equidistantes a $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}$.



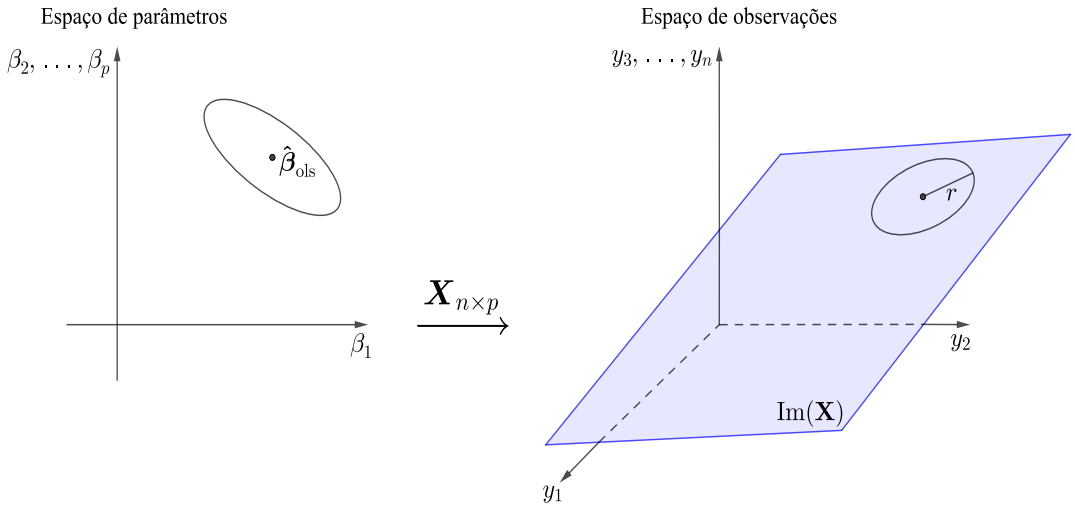
Fonte: Adaptado de Pereira (2017).

Observe, em relação ao raio r, que:

$$\begin{aligned} r^2 &= \left\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right\|^2 \\ &= \left(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\left(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}\right) \\ &= \left[\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)\right]'\left[\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)\right] \\ &= \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right). \end{aligned} \tag{2.81}$$

Portanto, os vetores $\boldsymbol{\beta}$'s que são levados na hiperesfera formam um elipsoide centrado em $\hat{\boldsymbol{\beta}}_{\text{ols}}$ no espaço de parâmetros, conforme representado na Figura 2.9.

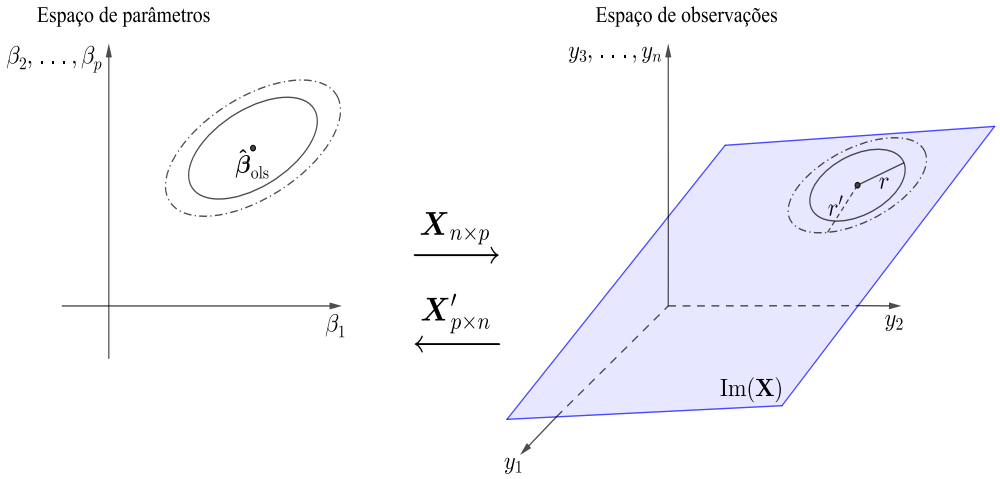
Figura 2.9 – Relação entre a hiperesfera e o elipsoide centrado em $\hat{\boldsymbol{\beta}}_{\text{ols}}$.



Fonte: Adaptado de Pereira (2017).

Variando-se o raio r da hiperesfera no espaço de observações, varia-se de forma equivalente o elipsoide no espaço de parâmetros, sem alterar no entanto as direções de seus eixos principais e sua excentricidade (FIGURA 2.10). Nesse sentido, o raio r pode ser considerado como um parâmetro de ajuste.

Figura 2.10 – Variação do raio da hiperesfera no espaço de observações.



Fonte: Adaptado de Pereira (2017).

Com o objetivo de apresentar os estimadores *Ridge*, LASSO e *Elastic Net*, antes é necessário que se aborde algumas normas de vetores que desempenham um papel fundamental na formulação desses estimadores. Logo, esse é o tema a ser abordado na próxima seção.

2.5 Normas de vetores

Nesta seção são apresentadas três importantes normas que são utilizadas na formulação de alguns métodos de regressão penalizada, abordados neste trabalho. Mas antes é conveniente formalizar, de forma muito sucinta, o conceito de comprimento de um vetor. Dados dois escalares x e y, a noção mais natural da distância entre x e y é obtida utilizando-se o valor absoluto, sendo essa distância definida como $|x-y|$. Podemos estender a definição para vetores, empregando uma função de distância que possua algumas propriedades de interesse. Neste sentido, podemos formalizar a noção de comprimento de um vetor.

Definição 2.5.1 (Norma vetorial): Dada uma função $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, chama-se *norma vetorial* de um vetor $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$, em relação ao produto interno, a seguinte relação:

$$\|\boldsymbol{u}\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2} = \sqrt{u_1 u_1 + u_2 u_2 + \ldots + u_n u_n} = \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}. \tag{2.82}$$

A definição formal da norma vetorial no $\mathbb{R}^n$ possui todas as propriedades que esperamos em relação a noção de comprimento de um vetor. De fato, dados dois vetores $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, segue que:

1. $\|\boldsymbol{u}\| \geq 0$ e $\|\boldsymbol{u}\| = 0$ se e somente se $\boldsymbol{u} = \boldsymbol{0}$.

2. $\|\alpha \boldsymbol{u}\| = |\alpha| \, \|\boldsymbol{u}\|$, para qualquer $\alpha \in \mathbb{R}$.

3. $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$.

A última propriedade é chamada de desigualdade triangular. Deve-se notar que quando $n = 1$, a norma vetorial é a função valor absoluto. As normas vetoriais que são mais comumente utilizadas pertencem à família de p-normas ou L-normas, que são definidas como:

$$\|\boldsymbol{u}\|_p = \left(\sum_{i=1}^{p} |u_i|^p \right)^{\frac{1}{p}}. \tag{2.83}$$

Pode ser mostrado, para qualquer $p \geq 1$, que a função $\|\cdot\|$ define uma norma vetorial. As seguintes p-normas que serão de particular interesse no presente trabalho são:

i) Se $p = 1$, temos a norma L_1 (norma de Manhattan):

$$\|\boldsymbol{u}\|_1 = \sum_{i=1}^{n} |u_i| = |u_1| + |u_2| + \ldots + |u_n| \, . \tag{2.84}$$

ii) Se $p = 2$, temos a norma L_2 (norma Euclidiana):

$$\|\boldsymbol{u}\|_2 = \sqrt{\sum_{i=1}^{n} u_i^2} = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2}. \tag{2.85}$$

iii) Se $p = \infty$, temos a norma L_∞ (norma de Chebyshev):

$$\|\boldsymbol{u}\|_\infty = \max_{1 \leq i \leq n} |u_i| . \tag{2.86}$$

Pode-se observar que a norma L_2 corresponde a própria norma vetorial de um vetor. Neste caso, $\|\boldsymbol{u}\|_2 = \|\boldsymbol{u}\|$. Como mencionado, essas três normas são muito importantes na formulação de alguns métodos da regressão penalizada. A importância de cada norma justifica-se pelo fato de que os métodos da regressão penalizada recebem algumas características que são devidas a cada uma das normas utilizadas. Ao se empregar a norma L_1, o método LASSO permite que algumas soluções para os parâmetros se tornem nulas. A norma L_2^2, utilizada na regressão *Ridge*, permite que as soluções obtidas por esse método apresentem valores pequenos (encolhimento). A combinação das normas citadas nesta seção também são de interesse. A técnica *Elastic Net* utiliza como termo em sua penalidade, uma combinação linear ponderada das normas L_1 e L_2^2. Por último, a norma L_∞ será empregada em outros dois métodos de regressão penalizada, que serão discutidos em seções posteriores e que são de grande importância para as propostas a serem apresentadas neste trabalho.

2.6 Regressão penalizada

Sabe-se que o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ pode se comportar mal em termos da precisão de predição, envolvendo preditores altamente correlacionados e numa situação de alta dimensão, incluindo o problema cada vez mais comum em que o número de preditores supera em muito o tamanho da amostra. Além da precisão de predição, um modelo parcimonioso é normalmente requerido em detrimento a um modelo mais complicado, devido à sua simplicidade e interpretabilidade. Diante disso, um primeiro objetivo consiste em melhorar a precisão de predição da resposta $\boldsymbol{y}$ e outro objetivo é realizar a seleção de algumas variáveis para o modelo, reduzindo a dimensão do vetor $\boldsymbol{\beta}$. Diante das dificuldades que surgem ao se utilizar o estimador de mínimos quadrados, a regressão penalizada surgiu como uma técnica altamente bem sucedida, sendo utilizada para o encolhimento de coeficientes, seleção e agrupamento de variáveis, contribuindo para a melhora da precisão de predição e da interpretabilidade do modelo de regressão.

Nesta seção serão apresentados alguns estimadores que surgiram para suprir as deficiências do estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ em situações como a de presença de multicolinearidade entre as variáveis e para casos em que $p > n$. Os estimadores a serem apresentados são os estimadores *Ridge*, LASSO e *Elastic Net*. Neste ponto, é importante destacar novamente que as normas L_1 e L_2 desempenham um importante papel na formulação desses três estimadores. Como destacado, esses métodos surgiram como uma penalização ou restrição das estimativas obtidas pelo estimador de mínimos quadrados para o modelo de regressão linear. As formas geométricas das restrições impostas estão intimamente relacionadas com as normas e isso se tornará mais claro no desenvolvimento dessa seção.

2.6.1 O estimador *Ridge*

Durante os últimos 60 a 70 anos diferentes estimadores foram propostos na literatura, como alternativas ao estimador de mínimos quadrados para a estimação dos parâmetros de um modelo de regressão linear. Alguns exemplos desses estimadores referem-se aos estimadores de encolhimento (*shirinkage* na literatura em geral), como os estimadores do tipo Stein, originalmente desenvolvidos por Bancroft (1944), Stein (1956) e James e Stein (1961), respectivamente. A principal característica desses estimadores é reduzir os valores dos coeficientes para valores alvo predeterminados.

Um exemplo de estimador de encolhimento é o estimador *Ridge*. Hoerl e Kennard (1970) introduziram a "regressão *Ridge*", que foi precursora para "estimadores penalizados", fundamentados na regularização de Tikhonov (1963). A regressão *Ridge* consiste na minimização de mínimos quadrados sujeitos a uma penalidade L_2^2. Essa metodologia é utilizada no desenvolvimento de análise de dados para casos de baixa e alta dimensão, bem como em aplicativos de redes neurais e em análises de *big data* (SALEH; ARASHI; KIBRIA, 2019). A principal característica do método *Ridge* é permitir a suavização de atributos que sejam correlacionados e que podem aumentar o ruído no modelo de regressão.

Considere o modelo usual de regressão linear com dados observados em n observações e p variáveis preditoras. Sejam $\boldsymbol{y}' = (y_1, y_2, \ldots, y_n)$ o vetor de respostas e $\boldsymbol{x}'_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ o vetor observado do j-ésimo preditor, com $j = 1, 2, \ldots, p$. Assuma que a variável resposta está centralizada e que cada preditor está em sua forma padronizada. Consequentemente,

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \ \text{ e } \ \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2 = 1, \tag{2.87}$$

para todo $j = 1, 2, \ldots, p$.

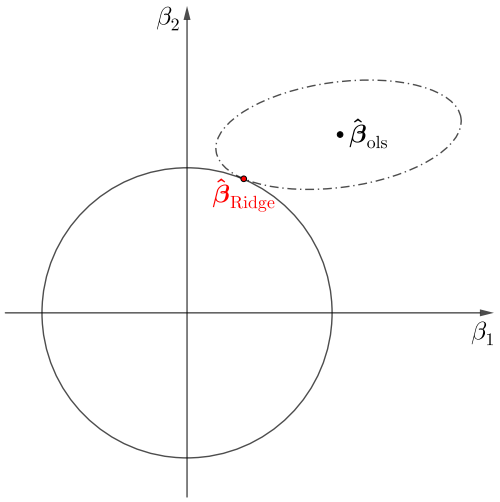
O problema de otimização de mínimos quadrados restritos para a regressão *Ridge* é dado por:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 \quad \text{restrito a} \quad \sum_{j=1}^{p} \beta_j^2 \leq t \,, \tag{2.88}$$

em que t é o parâmetro de ajuste. A norma L_2^2 favorece o encolhimento dos coeficientes de mínimos quadrados.

Embora dada matematicamente por (2.88), a forma do problema de otimização restrita é diretamente motivada mais pela interpretação geométrica da região de restrição do que pela forma algébrica da penalidade. A interpretação geométrica das soluções de mínimos quadrados restritas pela penalidade em (2.88) é útil para a compreensão de como ocorre o encolhimento das estimativas. Além de uma constante, os contornos da função de perda de soma dos quadrados, $r^2 = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)$, são elipsoides centrados na solução de mínimos quadrados, $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Considerando o $\mathbb{R}^2$, a solução *Ridge* ocorre quando a elipse intercepta a restrição imposta pela norma L_2^2 aos coeficientes, fornecendo assim as estimativas de interesse (FIGURA 2.11).

Figura 2.11 – Obtenção da solução *Ridge* no espaço paramétrico $\mathbb{R}^2$, utilizando a norma L_2^2.



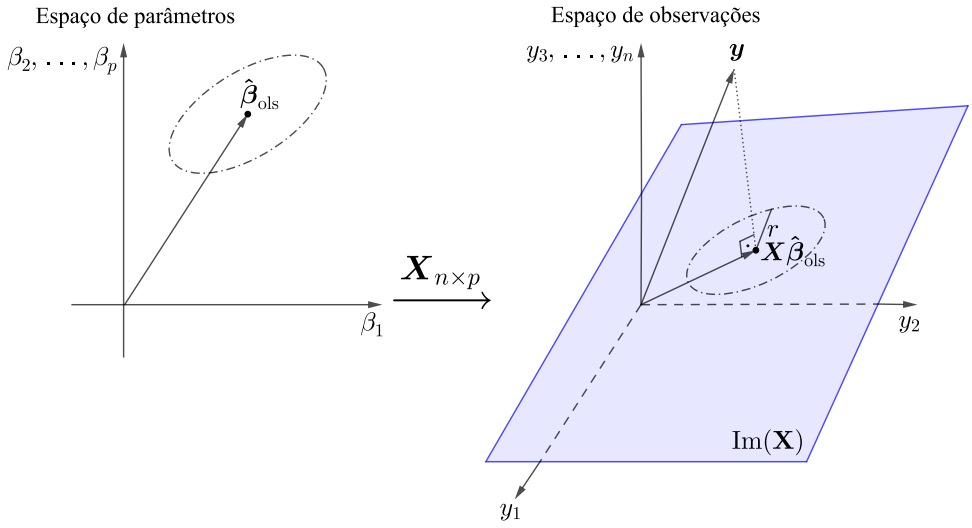
Fonte: Do autor (2020).

Sabe-se que o estimador de mínimos quadrados é o melhor estimador linear não viesado, dentre todos os estimadores lineares e não viesados. Uma pergunta natural que pode ser feita é: existe algum estimador viesado com erro quadrático menor em relação ao estimador

$\hat{\boldsymbol{\beta}}_{\text{ols}}$? A resposta para essa pergunta é sim e, em consonância com essa resposta, a regressão de encolhimento trata de estimadores que, ao permitirem viés, fornecem estimativas melhores no sentido de se diminuir o erro quadrático médio. Essa é uma das vantagens da regressão *Ridge*.

O estimador de mínimos quadrados é obtido ao se projetar ortogonalmente o vetor $\boldsymbol{y}$ na $\text{Im}(\boldsymbol{X})$. Introduz-se então uma penalização afirmando que as estimativas possíveis estão a uma distância r da projeção ortogonal, isto é, estão em uma hiperesfera de raio r contida na $\text{Im}(\boldsymbol{X})$. Essa hiperesfera está centrada na projeção ortogonal de $\boldsymbol{y}$ na imagem de $\boldsymbol{X}$, isto é, está centrada em $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}$ no espaço de observações (FIGURA 2.12). Essa penalização gera um problema, uma vez que em razão da projeção ser ortogonal, todos os pontos dessa hiperesfera estão a uma mesma distância de $\boldsymbol{y}$. Como escolher então uma estimativa? Uma ideia é utilizar a pré-imagem da hiperesfera, que é o elipsoide $r^2 = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)$ centrado em $\hat{\boldsymbol{\beta}}_{\text{ols}}$, no espaço de parâmetros $\mathbb{R}^p$.

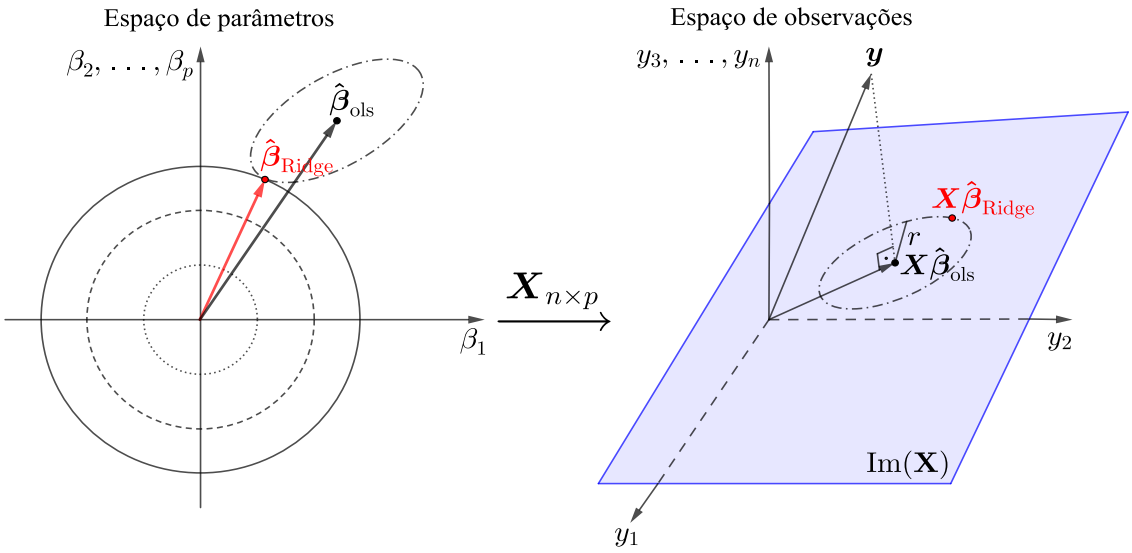
Figura 2.12 – Penalização na definição do estimador *Ridge*.



Fonte: Adaptado de Pereira (2017).

Adota-se então uma atitude conservadora. Todos os estimadores de $\boldsymbol{\beta}$ nesse elipsoide fornecem estimativas viáveis. Opta-se então por aquela de menor norma, isto é, toma-se o $\hat{\boldsymbol{\beta}}$ obtido como tangente entre o elipsoide e uma hiperesfera centrada na origem. A dedução analítica da expressão desse estimador é apresentada a seguir e é representada geometricamente na Figura 2.13. Para cada valor fixo de r, as estimativas $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ são obtidas como um problema de minimização. Analiticamente, esse problema pode ser descrito de duas formas equivalentes, tanto no espaço de parâmetros quanto no espaço de observações. No presente trabalho iremos nos restringir somente à primeira situação, ou seja, no espaço de parâmetros.

Figura 2.13 – A geometria do estimador *Ridge*.



Fonte: Adaptado de Pereira (2017).

Para a obtenção das estimativas $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$, o objetivo é minimizar a função $\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2$, sujeita à restrição $\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)=r^2$. Utilizando o método dos multiplicadores de Lagrange, a função lagrangeana para esse problema é:

$$\begin{aligned} L(\boldsymbol{\beta},\gamma) &= \|\boldsymbol{\beta}\|_2^2+\gamma\left\{\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)-r^2\right\} \\ &= \boldsymbol{\beta}'\boldsymbol{\beta}+\gamma\left\{\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)-r^2\right\}, \end{aligned} \tag{2.89}$$

em que $\gamma \geq 0$ é o parâmetro de penalização correspondente.

Como $\boldsymbol{I}$ e $\boldsymbol{X}'\boldsymbol{X}$ são matrizes simétricas de constantes, segue de Rencher e Shaalje (2008, p. 56) que:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}'\boldsymbol{I}\boldsymbol{\beta}\right)=2\boldsymbol{I}\boldsymbol{\beta}=2\boldsymbol{\beta}. \tag{2.90}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left[\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right)\right]=2\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\text{ols}}\right). \tag{2.91}$$

Diante desses últimos resultados, derivando a função L em relação aos parâmetros $\boldsymbol{\beta}$ e γ e igualando as derivadas resultantes ao vetor nulo $\mathbf{0}$, tem-se o seguinte sistema de equações:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\beta} + 2\gamma\left[\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)\right] = \mathbf{0}. \tag{2.92}$$

$$\frac{\partial L}{\partial \gamma} = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{ols}}\right) - r^2 = \mathbf{0}. \tag{2.93}$$

Utilizando a primeira igualdade segue que:

$$\boldsymbol{\beta} + \gamma\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \gamma\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} \Rightarrow \left(\boldsymbol{I} + \gamma\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{\beta} = \gamma\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}. \tag{2.94}$$

Portanto, a solução *Ridge* é dada explicitamente por:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{Ridge}}(r) &= \left(\boldsymbol{I} + \gamma\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\gamma\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} \\
&= \left[\gamma\left(\frac{1}{\gamma}\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)\right]^{-1}\gamma\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} \\
&= \gamma^{-1}\left(\frac{1}{\gamma}\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\gamma\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}} \\
&= \left(\frac{1}{\gamma}\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ols}}.
\end{aligned} \tag{2.95}$$

Uma vez que $\hat{\boldsymbol{\beta}}_{\text{ols}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, então:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}(r) = \left(\frac{1}{\gamma}\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\tau\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.96}$$

em que $\tau = \frac{1}{\gamma}$.

O valor de τ, em função de r, é obtido substituindo-se $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ na restrição

$$\left(\hat{\boldsymbol{\beta}}_{\text{Ridge}}(r) - \hat{\boldsymbol{\beta}}_{\text{ols}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_{\text{Ridge}}(r) - \hat{\boldsymbol{\beta}}_{\text{ols}}\right) = r^2. \tag{2.97}$$

Com base nas formas explícitas dos estimadores $\hat{\boldsymbol{\beta}}_{\text{ols}}$ e $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$, vem de (2.97) que:

$$\left[\left(\tau\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right]'\boldsymbol{X}'\boldsymbol{X}\left[\left(\tau I + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = r^2. \tag{2.98}$$

Como é fácil atribuir um valor para τ e obter o valor correspondente de r, geralmente o estimador *Ridge* é expresso em função de τ em detrimento de r:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\tau) = \left(\tau\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.99}$$

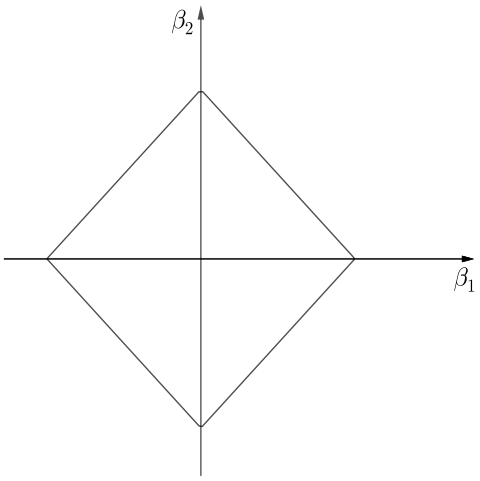
Tal substituição é adequada, porém τ não tem um significado geométrico como tem r. Cabe ressaltar novamente, como o problema que define o estimador *Ridge* é variacional, esse problema também admite uma outra forma equivalente no espaço de observações.

2.6.2 O estimador *Least Absolute Shrinkage and Selection Operator* (LASSO)

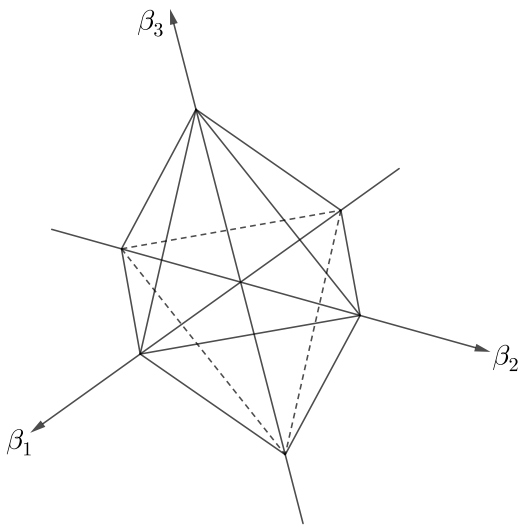
Embora a regressão *Ridge* alcance frequentemente uma melhor precisão de previsão ao encolher os coeficientes de $\hat{\boldsymbol{\beta}}_{\text{ols}}$, particularmente na situação de preditores altamente correlacionados, ela não pode produzir um modelo parcimonioso, uma vez que mantém naturalmente todos os preditores. Uma alternativa para se obterem estimativas que, além de apresentarem pouca variabilidade, forneçam estimativas de $\hat{\boldsymbol{\beta}}_{\text{ols}}$ com algumas de suas componentes nulas é o método de regressão LASSO. O método LASSO, acrônimo de *Least Absolute Shrinkage and Selection Operator*, foi proposto por Tibshirani (1996). Esse método consiste de um processo automático e supervisionado para a seleção de variáveis. Isso é alcançado devido ao fato do LASSO utilizar a norma L_1 como penalização às estimativas de mínimos quadrados. Nas Figuras 2.14 a) e b) ilustra-se a geometria da restrição LASSO, imposta com a utilização da norma L_1, considerando o $\mathbb{R}^2$ e o $\mathbb{R}^3$.

Figura 2.14 – Geometria da restrição LASSO imposta pela norma L_1.

(a) Restrição LASSO no $\mathbb{R}^2$.



(b) Restrição LASSO no $\mathbb{R}^3$.



Fonte: Do autor (2020).

Considerando o modelo usual de regressão linear, em que a variável resposta e as variáveis preditoras estão, respectivamente, em suas formas centralizada e padronizada, segue que o problema de otimização de mínimos quadrados restritos para a regressão LASSO é dado por:

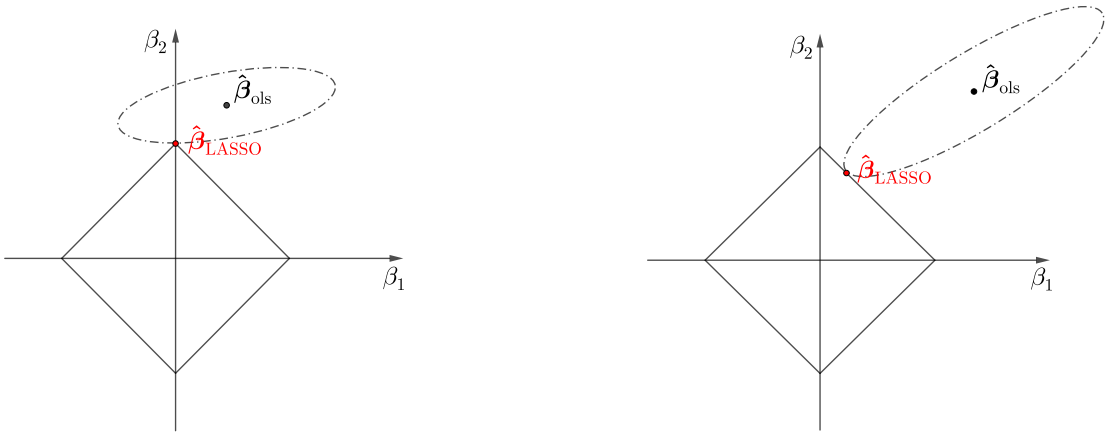
$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 \quad \text{restrito a} \quad \sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.100}$$

em que t é o parâmetro de ajuste. A norma L_1 favorece tanto o encolhimento quanto a esparsidade dos coeficientes de mínimos quadrados.

Mais uma vez a interpretação geométrica é empregada, agora para ilustrar como a restrição imposta pela utilização da norma L_1 atua sobre os coeficientes de mínimos quadrados ao se considerar a regressão LASSO. Pode-se observar na Figura 2.15 que devido à natureza da região de restrição, essa técnica permite o encolhimento de coeficientes e, eventualmente, a seleção de variáveis. Nesse último caso alguns coeficientes tornam-se nulos, o que naturalmente elimina algumas variáveis do modelo e reduz a dimensionalidade do problema. As estimativas nulas para os parâmetros de regressão são obtidas quando o elipsoide centrado em $\hat{\boldsymbol{\beta}}_{\text{ols}}$ tangencia a região da restrição LASSO nos eixos coordenados. Considerando o $\mathbb{R}^2$, na Figura 2.15a) pode-se observar o caso em que $\hat{\beta}_1 = 0$.

Figura 2.15 – Descrição geométrica da obtenção da estimativa LASSO no espaço paramétrico $\mathbb{R}^2$.

(a) Solução esparsa. (b) Solução não esparsa.



Fonte: Do autor (2020).

Embora seja uma técnica altamente bem sucedida, o método LASSO apresenta algumas limitações como um método de seleção de variáveis:

1. No caso $p > n$, o LASSO pode selecionar no máximo n variáveis. Em muitas situações, incluindo aquelas envolvendo dados *microarray* em que $p \gg n$, isso se torna uma característica limitante para um método de seleção de variáveis.

2. Se houver um conjunto de variáveis altamente correlacionadas, o LASSO tende a selecionar arbitrariamente apenas uma variável desse conjunto.

Zou e Hastie (2005) abordaram essas limitações do LASSO, ilustrando com um problema de seleção de genes em dados de análise *microarray*. Um conjunto de dados de *microarray* típico possui milhares de indicadores (genes) e frequentemente menos de 100 amostras. Para aqueles genes que compartilham o mesmo "caminho" biológico, as correlações entre eles pode ser alta (SEGAL; DAHLQUIST; CONKLIN, 2003). Esses genes podem ser pensados como um grupo. O método ideal de seleção de genes deve ser capaz de fazer duas coisas: eliminar genes triviais e automaticamente incluir grupos inteiros no modelo, uma vez que um gene entre eles é selecionado ("seleção agrupada"). Para $p \gg n$ e na situação de variáveis agrupadas, o LASSO não é o método ideal, porque ele só pode selecionar no máximo n variáveis de um conjunto de p variáveis candidatas (EFRON et al., 2004), ou seja, falta-lhe a capacidade de revelar as informações do agrupamento. Em relação a segunda limitação, Zou e Hastie (2005) afirmam que o LASSO tende a selecionar apenas uma variável do conjunto, não se "importando" qual seja a variável selecionada.

2.6.3 O estimador *Elastic Net*

Zou e Hastie (2005) introduziram o estimador *Elastic Net*, usando uma combinação ponderada das normas L_1 e L_2^2 para contornar os problemas relacionados a multicolinearidade das variáveis e para os casos em que $p > n$. Logo, a ideia na formulação desse método é simples, consistindo em minimizar a soma de quadrados dos resíduos restrito a uma combinação linear das restrições dos métodos *Ridge* e LASSO.

Mais uma vez, considere o modelo usual de regressão linear múltipla, assumindo que a variável resposta está centralizada e que cada preditor está em sua forma padronizada. Zou e Hastie (2005) definiram o novo estimador como um problema de otimização com mínimos quadrados restritos a combinação ponderada das normas L_1 e L_2^2, sendo dado por:

$$\hat{\boldsymbol{\beta}}_{\text{nen}} = \arg\min_{\boldsymbol{\beta}} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 \quad \text{restrito a} \quad \gamma_1 \sum_{j=1}^{p} \left|\beta_j\right| + \gamma_2 \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{2.101}$$

em que t é o parâmetro de ajuste e $\gamma_1, \gamma_2 \geq 0$.

O estimador obtido é denominado estimador *Naïve Elastic Net* (nen), denotado por $\hat{\boldsymbol{\beta}}_{\text{nen}}$. Uma vez que $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$ e $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$, pode-se reescrever a penalização da soma de quadrados para esse método. De fato,

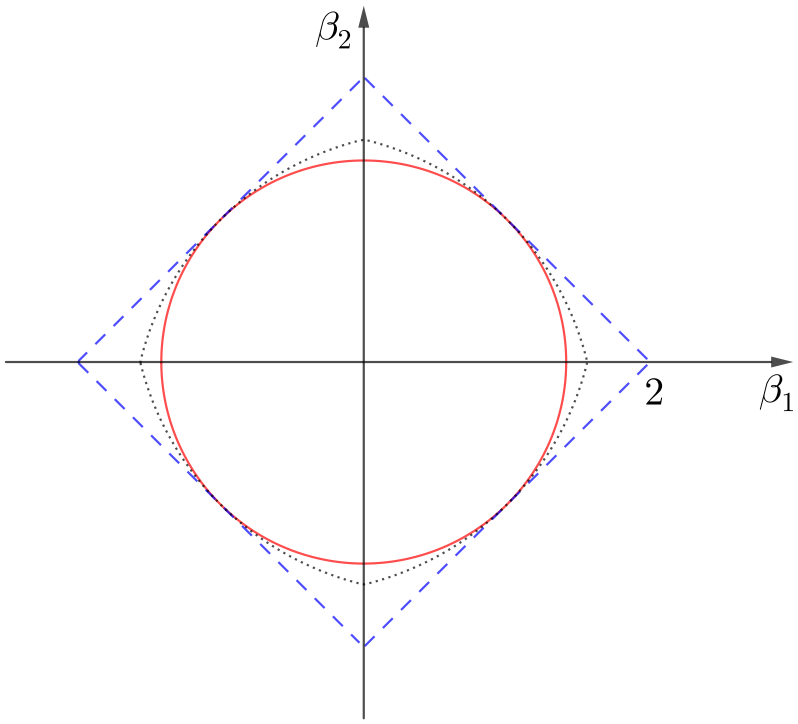
$$\begin{aligned} \gamma_1 \|\boldsymbol{\beta}\|_1 + \gamma_2 \|\boldsymbol{\beta}\|_2^2 \leq t &\Rightarrow \frac{\gamma_1}{\gamma_1+\gamma_2}\|\boldsymbol{\beta}\|_1 + \frac{\gamma_2}{\gamma_1+\gamma_2}\|\boldsymbol{\beta}\|_2^2 \leq \frac{t}{\gamma_1+\gamma_2} \\ &\Rightarrow \frac{\gamma_1}{\gamma_1+\gamma_2}\|\boldsymbol{\beta}\|_1 + \left(1 - \frac{\gamma_1}{\gamma_1+\gamma_2}\right)\|\boldsymbol{\beta}\|_2^2 \leq t' \\ &\Rightarrow \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 \leq t', \end{aligned} \tag{2.102}$$

com $0 \leq \alpha \leq 1$, em que para $\alpha = 0$ obtêm-se a estimação *Ridge* e para $\alpha = 1$ a estimação LASSO.

Decorre do fato das penalidades de estimação *Ridge* e LASSO serem estritamente convexa e convexa, respectivamente, que a penalidade $\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$ do método de regressão *Elastic Net* é estritamente convexa. Na situação em que $p > n$, o *Elastic Net* tem a capacidade de escolher todas as p variáveis, se necessário, tendendo a escolher as variáveis correlacionadas como um grupo. No entanto, essa seleção de agrupamento agora cria um modelo menos parcimonioso em comparação ao método LASSO.

Na Figura 2.16 ilustram-se três formas diferentes da restrição *Elastic Net*, para $\alpha = 0$ (*Ridge*), $\alpha = 1$ (LASSO) e $\alpha = 0,5$. Para analisar a forma do conjunto $\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$, basta analisar o seu bordo, ou seja, $\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 = t'$. Para isso, considere $t' = 2$ e $p = 2$. No plano (β_1, β_2), a solução obtida pelo método *Elastic Net* corresponde à primeira vez que os contornos da função de perda de soma dos quadrados atinge a região de restrição. Note que o contorno da restrição *Ridge* ($\alpha = 0$) é um círculo centrado na origem. Como os contornos são mais propensos a atingir um vértice, a não diferenciabilidade dos métodos LASSO e *Elastic Net* nos eixos favorece a esparsidade, com o método LASSO fazendo isso em maior grau devido a forma de sua restrição.

Figura 2.16 – Representação gráfica da região de restrição *Elastic Net* no plano (β_1, β_2), considerando $\alpha = 0,5$ (pontilhado) e os casos particulares *Ridge* ($\alpha = 0$, contínuo) e LASSO ($\alpha = 1$, tracejado).



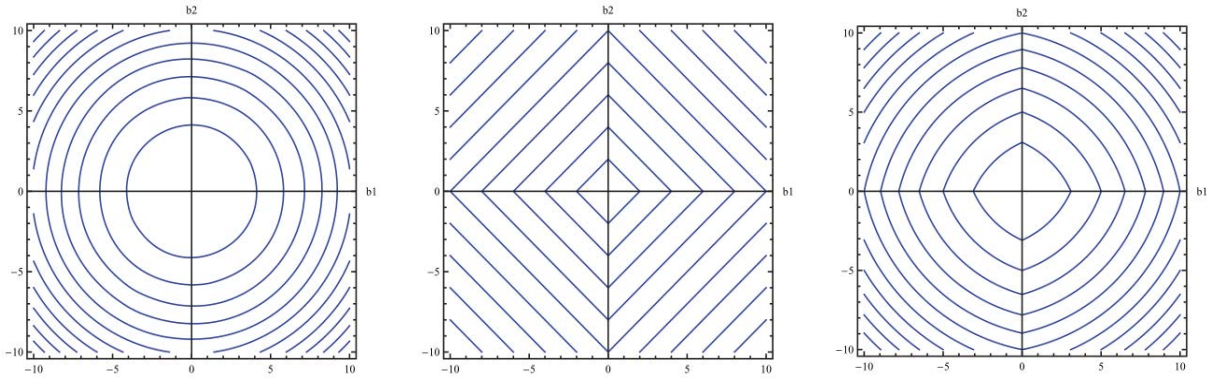
Fonte: Do autor (2020).

Na Figura 2.17 são apresentados os gráficos de contorno da penalidade *Elastic Net* (FIGURA 2.17c)), considerando novamente os casos particulares desse método, ou seja, as penalidades *Ridge* (FIGURA 2.17a)) e LASSO (FIGURA 2.17b)). Um aspecto interessante relacionado as penalidades desses métodos é que os seus contornos não dependem dos dados, ou seja, as suas formas não dependem da estrutura de correlação da matriz de delineamento $\boldsymbol{X}$. Um método que leva em consideração a estrutura de correlação dos dados na formação dos contornos da penalização é o método *Cluster Elastic Net*. Esse método foi proposto por Witten, Shojaie e Zhang (2014), sendo utilizado na teoria da regressão linear para a formação de *clusters* de variáveis que sejam altamente correlacionadas e associadas a resposta.

De acordo com Pereira (2017), o processo do *Naïve Elastic Net* pode ser visto como um processo de duas etapas: uma regressão *Ridge* e uma regressão tipo LASSO. Esse fato é muito importante na construção de algoritmos eficientes para o cálculo de estimativas *Elastic Net*. A ideia é que a regressão *Ridge* pode ser obtida com o emprego de estimadores mistos (GRUBER, 1998; COSTA, 2015), isto é, utilizando um modelo linear aumentado a partir do modelo original. Este procedimento é descrito a seguir.

Figura 2.17 – Gráficos de contorno considerando os estimadores *Ridge*, LASSO e *Elastic Net*.

(a) *Ridge*. (b) LASSO. (c) *Elastic Net*.



Fonte: Witten, Shojaie e Zhang (2014).

Primeiramente é feita uma reparametrização, em que os novos parâmetros são da forma $\boldsymbol{\beta}^* = \sqrt{1+\gamma_2}\boldsymbol{\beta}$. A partir do conjunto de dados $(\boldsymbol{y}_{n\times 1}, \boldsymbol{X}_{n\times p})$ define-se um novo conjunto de dados (dados aumentados) que pode, por exemplo, ser proveniente de um outro experimento anteriormente realizado. Com os dados aumentados, a regressão fica da forma $\left(\boldsymbol{y}^*_{(n+p)\times 1}, \boldsymbol{X}^*_{(n+p)\times p}\right)$, em que:

$$\boldsymbol{y}^*_{(n+p)\times 1} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \text{ e } \boldsymbol{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\gamma_2}} \begin{pmatrix} \boldsymbol{X}_{n\times p} \\ \sqrt{\gamma_2}\boldsymbol{I}_{p\times p} \end{pmatrix}. \tag{2.103}$$

Assim, tem-se então uma nova regressão em relação aos novos parâmetros $\boldsymbol{\beta}^*$, dada por $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$. O estimador de mínimos quadrados dessa regressão é então dado por:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^*_{\text{ols}} &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}^* \\
&= \left[\frac{1}{\sqrt{1+\gamma_2}}\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\frac{1}{\sqrt{1+\gamma_2}}\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\gamma_2}\boldsymbol{I} \end{pmatrix}\right]^{-1}\frac{1}{\sqrt{1+\gamma_2}}\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\
&= \left[\frac{1}{1+\gamma_2}\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\gamma_2}\boldsymbol{I} \end{pmatrix}\right]^{-1}\frac{1}{\sqrt{1+\gamma_2}}\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\
&= \frac{1+\gamma_2}{\sqrt{1+\gamma_2}}\left[\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\gamma_2}\boldsymbol{I} \end{pmatrix}\right]^{-1}\left(\begin{array}{c|c} \boldsymbol{X}' & \sqrt{\gamma_2}\boldsymbol{I} \end{array}\right)\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\
&= \sqrt{1+\gamma_2}\left(\boldsymbol{X}'\boldsymbol{X}+\gamma_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.
\end{aligned} \tag{2.104}$$

Como $\hat{\boldsymbol{\beta}}^*_{\text{ols}} = \sqrt{1+\gamma_2}\hat{\boldsymbol{\beta}}$, então $\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\gamma_2}}\hat{\boldsymbol{\beta}}^*_{\text{ols}}$. Dessa maneira, utilizando (2.104) tem-se:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \frac{1}{\sqrt{1+\gamma_2}}\hat{\boldsymbol{\beta}}^*_{\text{ols}} \\
&= \frac{1}{\sqrt{1+\gamma_2}}\sqrt{1+\gamma_2}\left(\boldsymbol{X}'\boldsymbol{X}+\gamma_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}+\gamma_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \hat{\boldsymbol{\beta}}_{\text{Ridge}}(\gamma_2). \qquad (2.105)
\end{aligned}$$

Portanto, o estimador de mínimos quadrados $\hat{\boldsymbol{\beta}}^*_{\text{ols}}$ satisfaz $\hat{\boldsymbol{\beta}}^*_{\text{ols}} = \sqrt{1+\gamma_2}\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\gamma_2)$, ou seja, é exatamente o valor obtido pela reparametrização da estimativa *Ridge* da regressão original $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

A estimativa LASSO para a regressão $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ é obtida minimizando a função lagrangeana, que é dada por:

$$L\left(\delta, \boldsymbol{\beta}^*\right) = \left\|\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right\|^2 + \delta\left\|\boldsymbol{\beta}^*\right\|_1, \qquad (2.106)$$

para algum $\delta \geq 0$.

Consequentemente, $\boldsymbol{\beta}^*_{\text{LASSO}} = \underset{\boldsymbol{\beta}^*}{\arg\min}\, L\left(\delta, \boldsymbol{\beta}^*\right)$, em que $\delta = \frac{\gamma_1}{\sqrt{1+\gamma_2}}$. Como

$$\begin{aligned}
\left\|\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\right\|^2 + \delta\left\|\boldsymbol{\beta}^*\right\|_1 &= \left\| \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} - \frac{1}{\sqrt{1+\gamma_2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\gamma_2}\boldsymbol{I} \end{pmatrix} \sqrt{1+\gamma_2}\boldsymbol{\beta} \right\|^2 + \delta\left\| \sqrt{1+\gamma_2}\boldsymbol{\beta} \right\|_1 \\
&= \left\| \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \sqrt{\gamma_2}\boldsymbol{\beta} \end{pmatrix} \right\|^2 + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= \left\| \frac{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}}{-\sqrt{\gamma_2}\boldsymbol{\beta}} \right\|^2 + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= \begin{pmatrix} \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ -\sqrt{\gamma_2}\boldsymbol{\beta} \end{pmatrix}' \begin{pmatrix} \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ -\sqrt{\gamma_2}\boldsymbol{\beta} \end{pmatrix} + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \gamma_2\boldsymbol{\beta}'\boldsymbol{\beta} + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma_2\|\boldsymbol{\beta}\|^2 + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma_2\|\boldsymbol{\beta}\|_2^2 + \sqrt{1+\gamma_2}\delta\|\boldsymbol{\beta}\|_1 \\
&= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma_1\|\boldsymbol{\beta}\|_1 + \gamma_2\|\boldsymbol{\beta}\|_2^2, \qquad (2.107)
\end{aligned}$$

pois $\|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{\beta}\|^2$ e $\gamma_1 = \sqrt{1+\gamma_2}\delta$.

Tem-se então que o estimador $\hat{\boldsymbol{\beta}}^*_{\text{LASSO}}$ é exatamente o estimador $\hat{\boldsymbol{\beta}}_{\text{nen}}$, pois ambos são definidos pela minimização da mesma função. Em termos dos parâmetros originais, $\hat{\boldsymbol{\beta}}_{\text{nen}} = \frac{1}{\sqrt{1+\gamma_2}}\hat{\boldsymbol{\beta}}^*_{\text{LASSO}}$. Dessa forma, o cálculo do estimador *Naïve Elastic Net* pode ser obtido como o estimador LASSO de um sistema aumentado.

Uma vez que o estimador *Naïve Elastic Net* é obtido por um procedimento em dois estágios, um encolhimento *Ridge* e em seguida um encolhimento do tipo LASSO, resulta que esse estimador se origina de um duplo encolhimento. Zou e Hastie (2005) afirmaram que esse duplo encolhimento não é muito eficiente para reduzir as variâncias e ainda introduz um viés desnecessário em comparação com os encolhimentos LASSO ou *Ridge*, quando considerados isoladamente. Os autores melhoraram o desempenho de predição do estimador $\hat{\boldsymbol{\beta}}_{\text{nen}}$ corrigindo o seu duplo encolhimento. Para corrigir os problemas apontados, o estimador *Elastic Net* foi definido como um reescalonamento do estimador *Naïve Elastic Net*, $\hat{\boldsymbol{\beta}}_{\text{en}} = (1+\gamma_2)\,\hat{\boldsymbol{\beta}}_{\text{nen}}$. Logo,

$$\hat{\boldsymbol{\beta}}_{\text{en}} = (1+\gamma_2)\left\{\arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma_1\|\boldsymbol{\beta}\|_1 + \gamma_2\|\boldsymbol{\beta}\|_2^2\right)\right\}. \tag{2.108}$$

2.7 Análise de componentes principais (PCA)

Nas seções subsequentes é abordada a PCA. Inicialmente, será apresentado um panorama do surgimento desse método com destaque ao contexto histórico, bem como as contribuições de alguns nomes importantes para o seu desenvolvimento. Seguem algumas definições e propriedades básicas do método. Entre as seções 2.7.4 e 2.7.6 é explorada a PCA utilizando-se a abordagem geométrica.

2.7.1 Breve histórico do desenvolvimento da análise de componentes principais

A presente seção, intitulada "breve histórico do desenvolvimento da análise de componentes principais", emerge do ponto de vista do autor no que se refere ao historicismo como ferramenta para o estudo da realidade histórica de pessoas ou eventos, para se compreender o atual estado de um determinado assunto. Por meio dessa ferramenta busca-se compreender em que estágio a PCA desenvolveu-se dentro de um determinado contexto histórico, a partir de sua formulação inicial. Mas o que vem a ser estágio quanto ao desenvolvimento em um contexto histórico? Essa questão pode ser respondida não somente em face a como ocorreu o desenvolvimento de uma técnica em termos estritamente matemáticos. A questão pode ser respondida

também com alguns panoramas culturais e históricos do período, que são importantes como instrumento para contextualizar como a matemática se situou numa sociedade em uma dada época, o que torna ainda mais relevante as contribuições de alguns cientistas. O "breve histórico" do título dessa seção é fornecido por ser uma tarefa muito ambiciosa a questão de se apresentar em poucas páginas os detalhes do período histórico, os cientistas e suas contribuições as origens da PCA. Diante disso e em conformidade ao que salienta Jolliffe (2002, p. 6), traçar as origens de algumas técnicas estatísticas é uma tarefa difícil. Todavia, como ponto de partida destacam-se dois trabalhos quando o assunto em voga diz respeito as origens da PCA. Apesar de haver entre eles um intervalo de 32 anos, esses são os artigos de Karl Pearson (1901) e Harold Hotelling (1933).

Se o esforço para se traçar as origens de uma técnica científica por si só já é uma tarefa difícil, pode-se conjecturar como também o é quando essa busca diz respeito a uma revolução do pensamento humano. A Estatística, como uma grande área do conhecimento humano, ocasionou uma revolução na ciência no século XX. Contudo, a origem dessa revolução iniciou-se no século anterior. Conforme Salsburg (2009, p. 25), é difícil definir o momento exato em que a ideia de um modelo estatístico tornou-se parte da ciência. Contribuições específicas a estatística podem ser encontradas com algum indício até nos trabalhos do grande astrônomo Johannes Kepler no século XVII, nos trabalhos de Daniel Bernoulli (1700-1782), Pierre-Simon Laplace (1749-1827), de Carl Friedrich Gauss (1777-1855), entre outros. Não obstante, Salsburg (2009) defende que a revolução que a Estatística iria ocasionar na ciência durante o século XX tenha se iniciado ainda no século XIX, mais precisamente na década de 1890 com os trabalhos do inglês Karl Pearson (1857-1936). Por quase todo o século XIX ainda reinou a visão determinista da ciência, sendo os trabalhos de Charles Robert Darwin (1809-1882) acerca de sua teoria da sobrevivência do mais apto e principalmente os trabalhos de Karl Pearson acerca dos modelos estatísticos, aqueles que ofereciam uma visão diferente da que até então era difundida e mais aceita no meio científico da época.

Nascido em 27 de março de 1857, "Carl" Pearson era o segundo de três filhos de William Pearson (1822-1907) e Fanny Smith Pearson (1827-1903) (FIGURA 2.18). Nos primeiros anos de vida de Pearson, sua família residiu em vários bairros do norte de Londres até 1866, quando se mudaram para uma casa na Mecklenburgh Square, em Bloomsbury, onde eles residiram até 1875. William Pearson viu o conhecimento como a chave do sucesso e dizia a seus filhos que isso exigiria muito esforço deles. William enfatizava regularmente a importância do trabalho

duro, especialmente quando seus filhos estavam na Universidade de Cambridge. Porter (2004, p. 16) destaca que o contexto familiar do jovem "Carl" não foi isento de tensões. Em Londres, seu pai estudou direito no *Inner Temple*, tornando-se advogado. A capacidade intelectual e a dedicação de William Pearson o levaram ao topo de sua profissão, junto ao *Queen's Counsel* (Conselho da Rainha) em 1875. Um dos preços desse sucesso foi o distanciamento de William Pearson a seus pais e isso pôde ser constatado pelo de fato de Pearson nunca ter conhecido sua avó paterna, embora ele tivesse 25 anos quando ela faleceu. Na época da morte de William em 1907, Pearson lembrou-se dele nos seguintes termos: "Eu aprendi muitas coisas com ele e sei que devo muito a ele, física e mentalmente. Mas éramos parecidos demais para ser totalmente solidários. Ele achou minha ciência loucura e eu achei sua lei restrita".

Figura 2.18 – Fotos de época dos pais de Karl Pearson.

(a) William Pearson.

(b) Fanny Smith Pearson.

Fonte: Porter (2004).

Em abril de 1934 em um jantar realizado em sua homenagem, Pearson descreveu seu pai como frio e remoto. De acordo com o seu relato, durante os termos legais seu pai acordava às 4h da manhã para ler seus resumos e preparar seus discursos para o tribunal, retornando para casa somente às 19h para jantar e seguia para a cama às 21h. Somente nas férias é que as crianças o viam. Todavia, Pearson reconhecia a influência de William sobre a sua forma de trabalhar. Em suas palavras nesse mesmo jantar ele disse que: "Eu herdei uma fração de seu poder pelo trabalho duro". Pearson parece ter exagerado acerca da frieza de seu pai. William Pearson podia ser franco e ele tinha um temperamento forte, mas uma preocupação inconfundível e até ter-

nura aparecem em suas cartas aos jovens. As dificuldades de suas relações foram grandemente exacerbadas por uma brecha entre “o pai” e “Mère” (Fanny Smith Pearson, quando ela assinava suas cartas). Fanny Pearson apresentou-se às crianças como vulnerável e maltratada, praticamente obrigando os meninos a tomarem partido. Eles a escolheram. Ela era especialmente próxima e dependente de seu filho mais sensível, “Carl” (PORTER, 2004, p. 16-19).

Na década de 1870, o jovem “Carl” Pearson mudou-se da Inglaterra para a Alemanha para estudar ciência política durante sua graduação. Nesse novo país, as obras de Karl Marx chamaram a sua atenção. O impacto das obras de Marx sobre ele foram tão grandes que Pearson alterou a grafia de seu nome, passando a se chamar Karl Person (FIGURA 2.19b). Apesar de ter feito o seu doutorado também em ciência política, o seu interesse residia na natureza dos modelos matemáticos. Em 1892 Pearson publicou *The grammar of science*. No período anterior à primeira guerra mundial, o seu livro era considerado um dos grandes livros sobre a natureza da ciência e da matemática (SALSBURG, 2009, p. 26). Nesse livro, ele argumentou que o método científico é essencialmente descritivo e não explicativo. O impacto desse livro pode ser constatado na obra daquele que é considerando um dos maiores cientistas de todos os tempos, o físico alemão Albert Einstein (1879-1955), que desenvolveu a teoria da relatividade geral, um dos pilares da física moderna ao lado da mecânica quântica. Com 26 anos de idade e no grupo de estudo de filosofia e física, a Academia Olímpia, Einstein juntamente com seus amigos Maurice Solovine (1875-1958) e Conrad Habicht (1876-1958) decidiram que o primeiro livro que eles deveriam ler seria *The grammar of science* de Pearson. Esse livro abordava vários temas que mais tarde se converteriam em parte das teorias de Einstein e de outros cientistas famosos. Pearson assegurou que as leis da natureza são relativas a habilidade perceptiva do observador. A irreversibilidade dos processos naturais, dizia Pearson, é um conceito puramente relativo (PEARSON, 1911, p. 600).

Heyde e Seneta (2001, p. 248-249) trazem um resumo da grandiosa obra de Karl Pearson, o considerando um dos principais arquitetos da moderna teoria da estatística matemática. Como alguns grandes cientistas, os seus interesses foram de ampla abrangência, desde astronomia, mecânica, meteorologia, física e ciências biológicas, incluindo em particular antropologia, eugenia, biologia evolutiva, hereditariedade e medicina. Além disso, ele direcionou os seus estudos sobre folclore e literatura alemã, a história de reformas humanistas ocorridas na Alemanha como a promovida por Martinho Lutero. Os autores consideram Pearson um escritor prodigioso: ele publicou mais de 650 artigos em sua vida, dos quais 400 foram relacionados

Figura 2.19 – Fotos de Karl Pearson em duas diferentes fases de sua vida.

(a) O menino “Carl” Pearson com seu taco de críquete.

(b) O jovem Karl Pearson.

Fonte: Porter (2004).

a Estatística. Em um período de 28 anos ele fundou e editou 6 jornais, sendo o co-fundador com o zoólogo darwiniano Walter Frank Raphael Weldon (1860-1906) e com Francis Galton (1822-1911) do jornal *Biometrika*. Na Estatística, as principais contribuições de Pearson foram o método da regressão linear, o coeficiente de correlação de Pearson, o teste Qui-quadrado de Pearson e dois coeficientes de assimetria. No final do século XIX, Pearson introduziu um novo vernáculo para estatísticas, incluindo termos como histograma, desvio padrão, moda, homoscedasticidade, heterocedasticidade, curtose e coeficiente de correlação produto-momento. Muitos desses termos foram cunhados em conferências ministradas por ele a partir de 1891. Embora o material dessas palestras não tenha conteúdo original, a abordagem de Pearson para o ensino foi altamente inovadora. Em uma de suas palestras, ele espalhou 10000 centavos pelo chão da sala de aula e pediu aos alunos que contassem o número de caras e coroas. O resultado foi aproximadamente metade de caras e coroas, elucidando de forma prática um importante teorema da probabilidade, a lei dos grandes números.

Essas palestras foram ministradas enquanto Pearson era simultaneamente professor do colégio universitário de Londres (UCL, do inglês “*University College London*”) e professor de geometria em Gresham. O primeiro cargo foi alcançado por Pearson em 1883, enquanto o segundo foi obtido em 1891. Em junho de 1890, seis meses antes de ocupar o seu segundo cargo

em Gresham, Pearson casou-se com Maria Sharpe (1853-1928), com quem teve três filhos, Sigrid Letitia Sharpe Pearson (1891-1971), Egon Sharpe Pearson (1895-1980) e Helga Sharpe Pearson (1899-1975) (FIGURA 2.20). As palestras que Pearson ministrou à partir de 1891 em Gresham, em particular as últimas 12 palestras, representaram um momento decisivo em sua carreira, devido especialmente a seu relacionamento com Weldon. Weldon foi o primeiro biólogo que Pearson conheceu e que estava interessado em usar uma abordagem estatística para problemas da evolução darwiniana. Por volta desse mesmo período, Pearson liderou com Galton, fundador da eugenia, um grupo de evolucionistas conhecidos como biometristas. O objetivo desse grupo era encontrar regularidades estatísticas que pudessem descrever a passagem de variações contínuas de uma população parental a sua prole. Sua ênfase na população darwiniana das espécies não apenas implicava a necessidade de medir sistematicamente a variação, mas também instigava a uma nova conceitualização de populações estatísticas. Além disso, foi essa matematização na teoria de Darwin que levou Pearson a uma mudança de paradigma do essencialismo aristotélico que sustentou o uso e os desenvolvimentos anteriores de estatísticas sociais e vitais. As perguntas de Weldon não apenas deram o impulso para o trabalho estatístico seminal de Pearson, mas também levaram à criação da escola de biometria na UCL (HEYDE; SENETA, 2001, p. 251; MAGNELLO, 2009).

Figura 2.20 – Karl e Maria Pearson, com o seu filho Egon e sua filha Sigrid.

Fonte: Porter (2004).

Na Figura 2.21 apresenta-se Pearson trabalhando, já em uma idade um pouco mais avançada, com destaque a fotos de alguns crânios que certamente foram utilizados para algumas de

suas medições biométricas. Sabe-se que Pearson era um computador entusiasmado e passava grande parte de seu tempo reduzindo vastos conjuntos de medições. Logo, ele sentiu que tinha que ter acesso a uma máquina de computação. Após muitos problemas na obtenção das finanças necessárias para essa finalidade, ele conseguiu comprar um aparelho *Brunsviga*. Até o final de sua vida em 1936 ele iria utilizar a velha calculadora *Brunsviga*, que era do começo do século.

Figura 2.21 – O velho estatístico no trabalho, com fotos de alguns crânios e sua calculadora *Brunsviga*.

Fonte: Porter (2004).

Em 1901, Karl Pearson publicou um trabalho intitulado "*On lines and planes of closest fit to systems of points in space*", sobre o ajuste de um conjunto de pontos no espaço p-dimensional a uma linha ou um plano. De acordo com Jolliffe (2002, p. 7), os problemas de otimização geométrica que ele considerava também o levaram aos componentes principais. Cabe ressaltar que ele não usou em nenhuma parte de seu artigo o termo "componente principal", mas afirmou que "é desejável representar um sistema de pontos no plano, em um espaço com três ou mais dimensões, por uma reta ou plano de melhor ajuste". O ajuste do sistema de pontos é "o melhor", uma vez que a soma das distâncias de cada ponto ao plano de ajuste é mínima. Na época e de forma mais geral acerca de seus trabalhos, Pearson teceu um interessante comentário sobre os seus novos métodos. Ele afirmou, 50 anos antes da ampla disponibilidade de computadores, que "os novos métodos poderiam ser facilmente aplicados a problemas numéricos" e embora também afirmasse que os cálculos se tornassem "pesados" para quatro ou mais variáveis, ele sugeriu que os métodos ainda seriam bastante viáveis. Mais recentemente e no atual cenário envolvendo *big data*, as afirmações de Pearson são corroboradas pela necessidade

de utilização de computação intensiva para conjuntos de dados em que o número de variáveis e de observações são cada vez maiores. Quase 120 anos após seu artigo, a PCA é comumente utilizada como uma ferramenta de análise exploratória de dados com essas características.

Outro personagem importante no desenvolvimento inicial da PCA foi Addison Harold Hotelling (1895-1973). Hotelling nasceu no dia 29 de setembro de 1895 em Fulda, Minnesota, Estados Unidos. Seus pais se chamavam Clair Alberta Hotelling (1869-1944) e Lucy Amelia Rawson Hotelling (1875-1959). Harold era o mais velho entre os seis filhos, criado de acordo com rígidos princípios metodistas. Seu pai tinha um negócio que se baseava na venda de feno. Ainda que pareça estranho inicialmente, nessa época os cavalos eram o principal meio de transporte e o feno era o principal combustível. O ano em que Harold nasceu marca o início da produção de automóveis nos Estados Unidos e logo ficou claro que os automóveis substituiriam os cavalos como meio de transporte. A *Ford Motor Company* começou a vender carros em 1903 e o pai de Harold logo percebeu que vender feno para cavalos se tornaria em breve algo do passado. Em Seattle, para onde sua família havia se mudado quando ele tinha 9 anos, Hotelling fez um ótimo uso da biblioteca pública, tornando-se um ávido leitor. Durante seus anos no ensino médio, estudou matemática, ciências e clássicos, mas estava particularmente interessado em eletricidade, lendo todos os livros que pôde encontrar sobre o assunto.

Ele se formou em jornalismo na Universidade de Washington, mas fez alguns cursos de matemática ministrados por Eric Temple Bell (1883-1960), que descobriu seu talento. No entanto, antes de concluir o curso ele foi convocado para o serviço de guerra durante a primeira guerra mundial. Apesar de seus talentos acadêmicos óbvios, o exército decidiu que ele era mais adequado para cuidar de mulas. Isso acabou sendo uma bênção disfarçada, pois uma mula temperamental chamada Dinamite o chutou e quebrou sua perna. Embora isso não pareça uma bênção em um primeiro momento, ele acabou sendo dispensado do serviço militar por causa disso, enquanto os outros membros de sua divisão foram enviados para a França, onde a maior parte foi morta. Ele foi dispensado do exército em 4 de fevereiro de 1919 e retomou seus estudos na Universidade de Washington. Convencido por Eric Temple Bell, Hotelling retornou à Universidade de Washington em janeiro de 1920 para estudar em um mestrado em matemática. Mais tarde naquele ano, ele se casou com Floy Tracy (1890-1932). Eles tiveram dois filhos, o primeiro nascido em 1923 com o nome de Eric Bell Hotelling (1923-1991), mostrando o seu apreço por Eric Temple Bell. Após a conclusão de seu mestrado em 1921, ele foi para a Universidade de Princeton com uma bolsa de doutorado para realizar pesquisas em matemática, que

foi concluído em 1924. No decorrer de sua carreira, Hotelling faria bom uso dos conhecimentos adquiridos em Princeton sobre topologia, geometria diferencial, análise e física matemática.

Em 1927, Hotelling foi nomeado professor associado no departamento de matemática da Universidade de Stanford. Esse também foi o ano em que foi publicada sua resenha do livro "*Statistical methods for research workers*", de Ronald Aylmer Fisher (1890-1962). Nele, ele elogiou Fisher por suas "brilhantes contribuições ao assunto", terminando sua crítica com as seguintes palavras: "O trabalho do autor é de importância revolucionária e deve ser muito mais conhecido neste país". Hotelling decidiu conhecer o trabalho de Fisher por experiência própria, uma vez que ele passaria seis meses no ano de 1929 trabalhando com Fisher na estação de pesquisa de experiências agrícolas de Rothamsted em Harpenden, na Inglaterra. Apesar do temperamento e das polêmicas do cientista inglês, Hotelling foi capaz de manter boas relações profissionais com Fisher no decorrer de sua carreira. Na Figura 2.22 são apresentadas duas fotos de Hotelling.

Figura 2.22 – Duas fotos de Harold Hotelling (1895-1973).

(a) Primeira foto de Harold Hotelling.

(b) Segunda foto de Harold Hotelling.

Fonte: Heyde e Seneta (2001).

Talvez sua contribuição mais significativa para a Estatística Matemática tenha sido a generalização da razão de *Student*, que ele publicou em 1931. Esse artigo sobre teste de hipóteses contém a ideia agora padrão de "intervalos de confiança". O ano de 1931 foi significativo para Hotelling de outra maneira, pois naquele ano ele deixou a universidade de Stanford para assumir um cargo de professor no departamento de economia da universidade Columbia. Lá,

ele ministrou cursos de economia, mas a maior parte de sua energia para pesquisa foi dedicada ao desenvolvimento de estatísticas. No entanto, a tragédia atingiu sua vida pessoal logo após a família chegar a Nova York. Sua esposa Floy ficou doente e morreu em outubro de 1932. Deixado com dois filhos pequenos para criar, ele foi ajudado pela irmã de Floy, que os levou para sua própria família e cuidou deles como seus próprios filhos. Hotelling conheceu Susanna Porter Edmonson (1909-1989), estudante de um curso de estatística que ele estava ministrando. Eles se casaram em junho de 1934 e se mudaram da cidade de Nova York para uma casa em Mountain Lakes, em Nova Jersey. Com sua nova esposa Hotelling teve mais 6 filhos.

Hotelling era "apaixonadamente" contra tudo o que Hitler defendia e argumentou fortemente que os Estados Unidos entrasse na guerra, muito antes dos ataques japoneses a base americana de Pearl Harbor no oceano Pacífico. No entanto, uma vez que os Estados Unidos se envolveram na segunda guerra mundial, ele convenceu os militares a criar um grupo de pesquisa estatística na universidade Columbia, que trabalhou em vários problemas estatísticos associados ao esforço de guerra, como questões de controle de qualidade. Hotelling pressionou a Universidade Columbia a criar um departamento independente de estatística com seu próprio pessoal permanente. Hotelling recebeu uma oferta da Universidade da Carolina do Norte para iniciar um programa de estatística nessa universidade. Ele já havia tentado persuadir a universidade Columbia a oferecer-lhe a mesma oportunidade, mas eles não foram persuadidos. Diante disso, Hotelling deixou a universidade Columbia em 1946 para iniciar um departamento de Estatística Matemática na universidade da Carolina do Norte em Chapel Hill. Ele foi presidente do departamento, bem como diretor associado do instituto de Estatística. Em 1966 ele se aposentou, tornando-se professor emérito.

Hotelling é conhecido pelos estatísticos devido à distribuição T-quadrado de Hotelling e seu uso nos testes de hipóteses e intervalos de confiança. Ele também introduziu a análise de correlação canônica. Como resultado de sua brilhante carreira nos Estados Unidos, Harold Hotelling ficou conhecido por sua liderança na profissão de estatístico, em particular por sua visão em relação a necessidade de um departamento de estatística em uma universidade, o que levou muitas universidades a criarem esses departamentos devido a sua influência. Hotelling foi reconhecido por sua liderança nos departamentos da Universidade Columbia e também na Universidade da Carolina do Norte.

A contribuição de Hotelling ao desenvolvimento da PCA foi fornecida em seu artigo de 1933, intitulado "*Analysis of a complex of statistical variables into principal components*". A

abordagem de Hotelling também parte das ideias da análise fatorial, mas a PCA que Hotelling define, possui um caráter realmente diferente da análise fatorial. A formulação da PCA proposta em 1933 por Hotelling ficou consagrada, sendo a formulação utilizada até hoje. A motivação de Hotelling é que pode haver um conjunto menor e fundamental de variáveis que determina uma quantidade de interesse em relação a p variáveis originais. Ele observou que essas variáveis foram chamadas de "fatores" na literatura psicológica e em decorrência disso introduziu o termo alternativo "componentes" para evitar confusão com outros usos da palavra "fator" na matemática. Hotelling escolhia os componentes que maximizassem com contribuições sucessivas a variância total das variáveis originais e chamou essas novas variáveis derivadas de componentes principais. O adjetivo "principal" talvez tenha sido dado por essa razão. A análise que se baseia em encontrar esses componentes ficou denominada de "método dos componentes principais".

Conforme Jolliffe (2002, p. 7), nos 32 anos entre os artigos de Pearson e Hotelling, muito pouco material relevante parece ter sido publicado, embora Rao (1964) indique que Frisch (1929) tenha adotado uma abordagem semelhante à de Pearson. Além disso, uma nota de rodapé em Hotelling (1933) sugere que Thurstone (1931) trabalhava de maneira semelhante ao seu trabalho, mas o artigo citado por ele se preocupa com a análise fatorial ao invés da PCA. Desde a sua origem, a PCA tem sido aplicada em uma ampla variedade de situações: na psicologia, medicina, meteorologia, geografia, ecologia, agronomia, entre outros.

Além das referências citadas é importante ressaltar que para as partes relacionadas ao desenvolvimento da PCA utilizou-se Jolliffe (2002, p. 6-9), enquanto que em relação aos detalhes biográficos de Harold Hotelling foram utilizados os artigos de Levene (1974), Smith (1978) e Darnell (1988).

2.7.2 Definições e propriedades básicas

Seja $\boldsymbol{X}' = (X_1, X_2 \ldots, X_p)$ um vetor aleatório com vetor de médias $\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$ e matriz de covariâncias $\boldsymbol{\Sigma}_{p\times p}$. Sejam $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ os autovalores da matriz $\boldsymbol{\Sigma}$, com autovetores unitários $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ que satisfazem:

1. $\boldsymbol{v}_i'\boldsymbol{v}_j = 0$.

2. $\boldsymbol{v}_i'\boldsymbol{v}_i = 1$.

3. $\boldsymbol{\Sigma}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$,

para todo $i, j = 1, 2, \ldots, p$, com $i \neq j$.

Como $\boldsymbol{\Sigma}$ é uma matriz positiva definida, os seus autovalores são todos positivos. Além disso, sendo $\boldsymbol{\Sigma}$ uma matriz positiva definida é muito comum listar seus autovalores em ordem decrescente: $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Os autovetores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ são listados na mesma ordem, com $\boldsymbol{v}_1$ correspondendo a λ_1, $\boldsymbol{v}_2$ correspondendo a λ_2 e, assim por diante. Os elementos v_{ij} dos autovetores são denominados na literatura estatística por coeficientes ou por *loadings*. A partir dos autovetores normalizados da matriz $\boldsymbol{\Sigma}$ pode-se definir uma matriz ortogonal $\boldsymbol{V}$ de dimensões $(p \times p)$, da seguinte forma:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_{(1)} & \boldsymbol{v}_{(2)} & \ldots & \boldsymbol{v}_{(p)} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \ldots & v_{pp} \end{bmatrix}, \tag{2.109}$$

em que $\left(\boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}, \ldots, \boldsymbol{v}_{(p)}\right)$ são vetores coluna de $\boldsymbol{V}$.

Considere então $Z_i = \boldsymbol{v}_i' \boldsymbol{X}$. Cada Z_i é a combinação linear das variáveis aleatórias do vetor $\boldsymbol{X}$, possuindo média igual a $\boldsymbol{v}_i' \boldsymbol{\mu}$ e variância igual a λ_i. De fato, a média e variância de cada Z_i são obtidas, respectivamente, da seguinte maneira:

$$\begin{aligned} E[Z_i] &= E\left[\boldsymbol{v}_i' \boldsymbol{X}\right] \\ &= E\left[v_{1i}X_1 + v_{2i}X_2 + \ldots + v_{pi}X_p\right] \\ &= E\left[v_{1i}X_1\right] + E\left[v_{2i}X_2\right] + \ldots + E\left[v_{pi}X_p\right] \\ &= v_{1i}E[X_1] + v_{2i}E[X_2] + \ldots + v_{pi}E[X_p] \\ &= v_{1i}\mu_1 + v_{2i}\mu_2 + \ldots + v_{pi}\mu_p \\ &= \boldsymbol{v}_i' \boldsymbol{\mu}. \end{aligned} \tag{2.110}$$

$$\text{var}[Z_i] = \text{var}\left[\boldsymbol{v}_i' \boldsymbol{X}\right] = \boldsymbol{v}_i' \text{cov}[\boldsymbol{X}]\, \boldsymbol{v}_i = \boldsymbol{v}_i' \boldsymbol{\Sigma} \boldsymbol{v}_i = \boldsymbol{v}_i' \lambda_i \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i' \boldsymbol{v}_i = \lambda_i. \tag{2.111}$$

De forma mais geral, o vetor $\boldsymbol{Z}' = (Z_1, Z_2, \ldots, Z_p)$ é composto de p combinações lineares das variáveis aleatórias do vetor $\boldsymbol{X}$, possuindo vetor de médias igual a $\boldsymbol{V}'\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{D}_{p \times p}$. A matriz $\boldsymbol{D}$ é uma matriz diagonal, cujos elementos são os autovalores $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ da matriz $\boldsymbol{\Sigma}$. Os elementos que não se encontram na diagonal principal são todos nulos. De fato, para todo $i \neq j$ com $i, j = 1, 2, \ldots, p$ resulta que:

$$\text{cov}\left[Z_i, Z_j\right] = \text{cov}\left[\boldsymbol{v}_i' \boldsymbol{X}, \boldsymbol{v}_j' \boldsymbol{X}\right] = \boldsymbol{v}_i' \text{cov}[\boldsymbol{X}]\, \boldsymbol{v}_j = \boldsymbol{v}_i' \boldsymbol{\Sigma} \boldsymbol{v}_j = \boldsymbol{v}_i' \lambda_j \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_i' \boldsymbol{v}_j = 0. \tag{2.112}$$

Logo,

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}. \tag{2.113}$$

Com base na matriz $\boldsymbol{D}$ pode-se observar que $\boldsymbol{Z}$ é composto de variáveis aleatórias que são não correlacionadas entre si. A PCA parte do estudo de um fenômeno com p variáveis correlacionadas para p variáveis que devem ser mutuamente não correlacionadas. A ideia central presente em sua metodologia é estudar a estrutura de variâncias e covariâncias do vetor original $\boldsymbol{X}$ por meio das p combinações lineares construídas em $\boldsymbol{Z}$, denominadas de componentes principais. Como consequência desse contexto, os objetivos mais importantes ao utilizar-se a PCA consistem em:

1. Obter novas variáveis que possam expressar a máxima informação possível, em termos de variabilidade, contida no conjunto de variáveis originais.

2. Reduzir a dimensão do problema que está sendo estudado, como passo prévio para trabalhos futuros.

3. Eliminar, quando possível, algumas das variáveis originais se elas apresentam pouca importância para o problema em análise.

Em específico, o objetivo ii) consiste em encontrar um número reduzido k ($k < p$) de novas variáveis $(Z_1, Z_2, \dots, Z_k)$, sendo dessa forma o objetivo ii) intrinsecamente relacionado ao objetivo i). Essa relação deve-se ao fato de que ao novo conjunto de variáveis $(Z_1, Z_2, \dots, Z_k)$ espera-se que o mesmo forneça uma proporção de variância maior possível em relação a variabilidade presente nas variáveis originais $(X_1, X_2, \dots, X_p)$. Todavia, ainda que o número de componentes a ser utilizado seja satisfatório, no sentido de ser reduzido, um número p muito elevado de variáveis pode ocasionar outro problema. Esse problema ocorre em razão da dificuldade inerente a interpretação de cada componente que seja derivado eventualmente de um número muito elevado de variáveis. Dessa maneira, o objetivo iii) também é almejado na PCA. Um dos avanços recentes na literatura para contornar o problema de interpretação em componentes principais contendo um número muito elevado de variáveis são os métodos que fornecem componentes esparsos, dentre os quais se destacam o método SPCA proposto por Zou, Hastie e Tibshirani (2006).

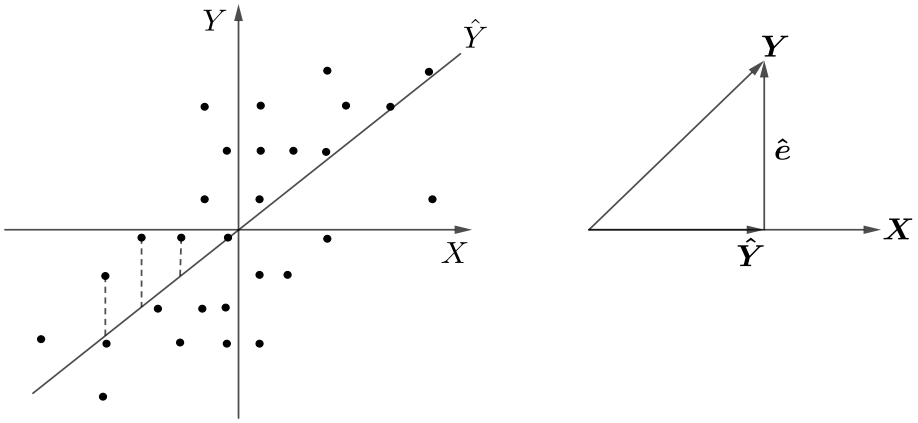
Monroy e Rivera (2012, p. 272) fazem um interessante paralelo entre a obtenção dos componentes na PCA e uma fotografia retirada de um grupos de pessoas. Considere que um grupo de pessoas deseja ser fotografado. Segundo os autores, a cabeça de cada pessoa pode ser comparada a pontos no $\mathbb{R}^4$ (três coordenadas para o espaço e uma para a hora ou a data). Não é difícil conjecturar como a localização da câmera em relação ao grupo produzirá fotografias diferentes do mesmo grupo, de tal maneira que indivíduos próximos em uma fotografia podem parecer muito distantes em outra. A PCA procura "tirar a melhor fotografia" sobre um conjunto de dados. Nesse caso, uma boa fotografia corresponderia ao menor subespaço, k por exemplo, sendo aquele que forneça um bom ajuste para as observações e variáveis, de modo que as distâncias entre pontos no novo subespaço forneçam uma boa representação das distâncias originais. Para concluir esse paralelo, uma fotografia nada mais seria do que a representação bidimensional de um evento que ocorre em quatro dimensões (espaço-tempo).

Neste ponto é importante destacar alguns aspectos que diferenciam os modelos de regressão e de componentes principais, uma vez que esses modelos são frequentemente confundidos. Em um modelo de regressão estabelece-se uma relação funcional entre uma variável resposta Y e variáveis regressoras $(X_1, X_2, \ldots, X_p)$. O objetivo consiste então em se obter estimativas para parâmetros $(\beta_0, \beta_1, \ldots, \beta_p)$ que determinam a relação funcional. Por sua vez, em um modelo de componentes principais cada uma das variáveis $(Z_1, Z_2, \ldots, Z_p)$ é determinada a partir de combinações lineares das variáveis $(X_1, X_2, \ldots, X_p)$. Uma primeira diferença entre os modelos é que no método PCA as variáveis $(Z_1, Z_2, \ldots, Z_p)$ não são mensuradas com base no experimento ou levantamento amostral, como o são as variáveis $(Y, X_1, X_2, \ldots, X_p)$ para do modelo de regressão. Em decorrência desse fato, os componentes também são denominados de variáveis latentes (JOHNSON; WICHERN, 2007).

Além do que foi exposto anteriormente, mais uma diferença pode ser fornecida em relação aos modelos de regressão e dos componentes principais. O aspecto que norteia essa diferença é concernente à forma como cada modelo é obtido. Para isso, considere o modelo de regressão linear entre as variáveis X e Y. A reta de regressão para a variável resposta Y é obtida pela minimização das somas das distâncias quadráticas dos pontos para a reta de regressão na direção paralela ao eixo Y, como é mostrado pelas linhas tracejados de alguns pontos na Figura 2.23. Por sua vez, o componente principal Z_1 pode ser obtido com base na combinação linear entre as variáveis X_1 e X_2. Na Figura 2.24 pode-se observar que o eixo do primeiro componente é definido no sentido de maior variabilidade verificada entre as variáveis X_1 e X_2. Neste

caso, esse eixo pertence a elipse que majoritariamente envolve a maior parte dos pontos. O componente Z_1 é determinado pela minimização dos desvios quadrados dos pontos na direção perpendicular ao eixo, o que é evidenciado pelos segmentos de reta tracejados dos pontos ao componente Z_1.

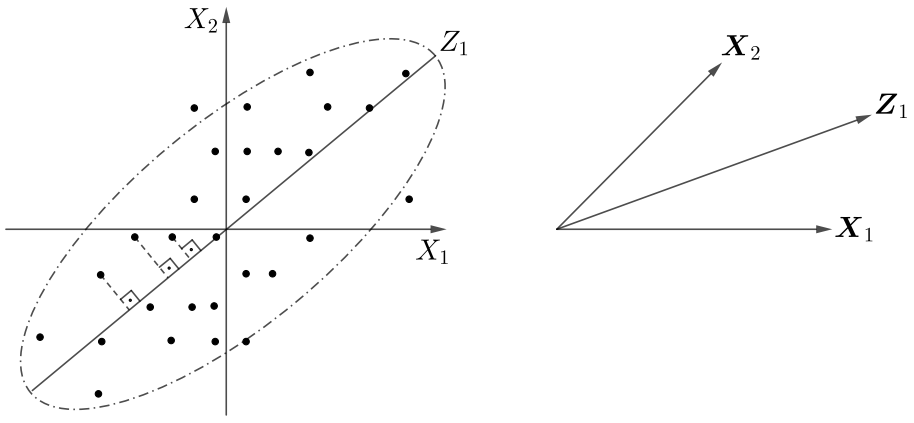
Figura 2.23 – Obtenção da reta de regressão enre as variáveis X e Y.



Fonte: Adaptado de Wickens (1995).

Logo, a obtenção da reta de regressão $\hat{Y}$ ocorre pela projeção vertical (paralela a Y) de cada ponto em $\hat{Y}$, de tal forma que essa reta seja, dentre todas as retas possíveis, aquela que minimize uma função das diferenças entre os valores observados e os valores ajustados. Por sua vez, o primeiro componente é obtido pela projeção ortogonal de cada ponto em Z_1, assegurando-se que esse componente contemple a máxima variabilidade possível presente entre as variáveis originais X_1 e X_2. Com isso torna-se claro a diferença entre os métodos com relação a forma como cada modelo é construído.

Figura 2.24 – Obtenção do componente Z_1 com base na combinação linear entre as variáveis X_1 e X_2.



Fonte: Adaptado de Wickens (1995).

Uma maneira de construir os componentes principais é utilizar a decomposição espectral da matriz de covariâncias $\boldsymbol{\Sigma}$. Todavia, no presente trabalho esse resultado será fornecido com o objetivo de se apresentar algumas propriedades das variáveis $(Z_1, Z_2, \ldots, Z_p)$.

Teorema 2.7.1 (Decomposição Espectral): Para a matriz quadrada e simétrica $\boldsymbol{\Sigma}_{p\times p}$, existe uma matriz ortogonal $\boldsymbol{V}_{p\times p}$ e uma matriz diagonal $\boldsymbol{D}_{p\times p}$ tais que:

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}', \tag{2.114}$$

em que $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_{(1)} & \boldsymbol{v}_{(2)} & \ldots & \boldsymbol{v}_{(p)} \end{bmatrix}$ e $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, sendo que $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ e $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ são, respectivamente, os autovetores e os autovalores da matriz $\boldsymbol{\Sigma}$.

Prova:

Como $\boldsymbol{V}$ é uma matriz quadrada $(p \times p)$ com colunas ortonormais tem-se que $\boldsymbol{I} = \boldsymbol{V}\boldsymbol{V}'$ (RENCHER; SHAALJE, 2008, p. 42-43). Multiplicando ambos os membros de $\boldsymbol{I} = \boldsymbol{V}\boldsymbol{V}'$ por $\boldsymbol{\Sigma}$ à esquerda resulta que:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{V}'. \tag{2.115}$$

Logo, substituindo $\boldsymbol{V}$ em (2.115) segue que:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{V}' \\
&= \boldsymbol{\Sigma}\begin{bmatrix} \boldsymbol{v}_{(1)} & \boldsymbol{v}_{(2)} & \ldots & \boldsymbol{v}_{(p)} \end{bmatrix}\begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_p' \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}\boldsymbol{v}_{(1)} & \boldsymbol{\Sigma}\boldsymbol{v}_{(2)} & \ldots & \boldsymbol{\Sigma}\boldsymbol{v}_{(p)} \end{bmatrix}\begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_p' \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1\boldsymbol{v}_{(1)} & \lambda_2\boldsymbol{v}_{(2)} & \ldots & \lambda_p\boldsymbol{v}_{(p)} \end{bmatrix}\begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_p' \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{v}_{(1)} & \boldsymbol{v}_{(2)} & \ldots & \boldsymbol{v}_{(p)} \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}\begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_p' \end{bmatrix} \\
&= \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}',
\end{aligned} \tag{2.116}$$

como queríamos mostrar.

■

Essa demonstração não é original e pode ser vista também em Rencher e Shaalje (2008, p. 51-52). Ainda de acordo com os autores, como $\boldsymbol{\Sigma}$ é simétrica e em conformidade com as definições de $\boldsymbol{V}$ e $\boldsymbol{D}$ dadas no teorema da decomposição espectral, segue que $\boldsymbol{V}$ diagonaliza $\boldsymbol{\Sigma}$, isto é, $\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{D}$.

O primeiro resultado a ser apresentado é em relação a variância total. A variância total corresponde ao traço da matriz de covariâncias de um vetor aleatório $\boldsymbol{X}$ e é uma medida da dispersão da variabilidade das variáveis X_i, $i = 1,2,\ldots,p$.

Proposição 2.7.1: A variância total da matriz de covariâncias dos vetores $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ e $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)$ são iguais.

Prova:

A variância total da matriz de covariâncias $\boldsymbol{\Sigma}$ do vetor $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ é:

$$\text{var}_{\text{total}}[\boldsymbol{\Sigma}] = \text{tr}[\boldsymbol{\Sigma}] = \sum_{i=1}^{p} \sigma_{ii}, \tag{2.117}$$

em que $\text{var}[X_i] = \sigma_{ii}$, para $i = 1,2,\ldots,p$.

Por sua vez, a variância total da matriz de covariâncias $\boldsymbol{D}$ de $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)$ é:

$$\text{var}_{\text{total}}[\boldsymbol{D}] = \text{tr}[\boldsymbol{D}] = \sum_{i=1}^{p} \lambda_i, \tag{2.118}$$

em que $\text{var}[Z_i] = \lambda_i$, para $i = 1,2,\ldots,p$.

Diante desses resultados e utilizando o teorema da decomposição espectral resulta que:

$$\begin{aligned}
\text{var}_{\text{total}}[\boldsymbol{\Sigma}] &= \text{tr}[\boldsymbol{\Sigma}] \\
&= \text{tr}\left[\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'\right] \\
&= \text{tr}\left[(\boldsymbol{V}\boldsymbol{D})\boldsymbol{V}'\right] \\
&= \text{tr}\left[\boldsymbol{V}'(\boldsymbol{V}\boldsymbol{D})\right] \\
&= \text{tr}\left[\boldsymbol{V}'\boldsymbol{V}\boldsymbol{D}\right] \\
&= \text{tr}[\boldsymbol{I}\boldsymbol{D}] \\
&= \text{tr}[\boldsymbol{D}] \\
&= \sum_{i=1}^{p} \lambda_i.
\end{aligned} \tag{2.119}$$

Logo, $\text{var}_{\text{total}}[\boldsymbol{\Sigma}] = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$. Portanto, $\text{var}_{\text{total}}[\boldsymbol{\Sigma}] = \text{var}_{\text{total}}[\boldsymbol{D}]$.

■

Outra medida de variabilidade dos dados é a variância generalizada, definida como o determinante da matriz de covariâncias. Mais uma vez, o teorema da decomposição espectral pode ser empregado para provar que essa medida de variabilidade é igual para as matrizes de covariâncias dos vetores $\boldsymbol{X}$ e $\boldsymbol{Z}$.

Proposição 2.7.2: A variância generalizada das matrizes de covariâncias do vetores aleatórios $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ e $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)$ são iguais.

Prova:

De fato,

$$|\boldsymbol{\Sigma}| = \left|\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'\right| = \left|(\boldsymbol{V}\boldsymbol{D})\boldsymbol{V}'\right| = \left|\boldsymbol{V}'(\boldsymbol{V}\boldsymbol{D})\right| = \left|\boldsymbol{V}'\boldsymbol{V}\boldsymbol{D}\right| = |\boldsymbol{I}\boldsymbol{D}| = |\boldsymbol{D}|. \tag{2.120}$$

Portanto,

$$|\boldsymbol{\Sigma}| = \prod_{i=1}^{p} \lambda_i = |\boldsymbol{D}|. \tag{2.121}$$

■

Por meios desses resultados pode-se verificar que a variabilidade presente nas variáveis originais é a mesma quando se utilizam todos os componentes principais. Uma vez que um dos objetivos da PCA é utilizar $k < p$ componentes, os resultados indicam que o uso de k componentes fornece uma aproximação em termos da estrutura de variâncias e covariâncias das variáveis originais.

Diante de tudo o que foi exposto, pode-se agora definir a proporção de variância de um componente principal.

Definição 2.7.1 (Proporção de variância): A proporção de variância total de $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ que é explicada pelo k-ésimo componente principal Z_k é dada por:

$$\frac{\text{var}[Z_k]}{\text{var}_{\text{total}}[\boldsymbol{X}]} = \frac{\text{var}[Z_k]}{\text{tr}[\boldsymbol{\Sigma}]} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}. \tag{2.122}$$

Uma vez que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, os componentes já são concebidos hierarquicamente em termos da proporção de variância, ou seja, o primeiro componente apresenta maior valor dessa quantidade em relação aos demais componentes. O mesmo raciocínio aplica-se ao segundo, terceiro e aos demais componentes. Outra medida útil é a proporção de variância acumulada, que é definida a seguir.

Definição 2.7.2 (Proporção de variância acumulada): A proporção de variância total do vetor aleatório $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ que é explicada pelos k primeiros componentes principais $(Z_1, Z_2, \ldots, Z_k)$ é dada por:

$$\frac{\sum_{i=1}^{k} \text{var}[Z_i]}{\text{var}_{\text{total}}[\boldsymbol{X}]} = \frac{\sum_{i=1}^{k} \text{var}[Z_i]}{\text{tr}[\boldsymbol{\Sigma}]} = \frac{\sum_{i=1}^{k} \text{var}[Z_i]}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}. \tag{2.123}$$

Toda a apresentação e construção dos resultados foi realizada considerando a matriz $\boldsymbol{\Sigma}$, ou seja, os componentes principais são obtidos via matriz de covariâncias populacional. Todavia, na maior parte dos estudos de problemas reais o conhecimento da população, isto é, o conhecimento dos parâmetros que a caracterizam não é possível de ser realizado devido a fatores que podem variar desde a sua natureza física, de ordem financeira ou ainda relacionado ao tempo para coleta de informações de todos os indivíduos da população. Nesse cenário tão comum faz-se necessário a utilização de uma amostra que seja representativa para a população em análise. Logo, as inferências estatísticas para a matriz populacional $\boldsymbol{\Sigma}$ devem ser feitas utilizando-se a matriz de covariâncias amostrais.

2.7.3 Análise de componentes principais via matriz de covariâncias amostrais

Se $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ é uma amostra multivariada de uma população p-dimensional com matriz de covariâncias populacional $\boldsymbol{\Sigma}$, segue que a matriz de observações de dimensões $(n \times p)$ é dada por:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}. \tag{2.124}$$

Os componentes principais podem ser obtidos de forma análoga como apresentado na seção anterior. Os componentes serão combinações lineares não correlacionadas dos $\boldsymbol{x}_i'$s com

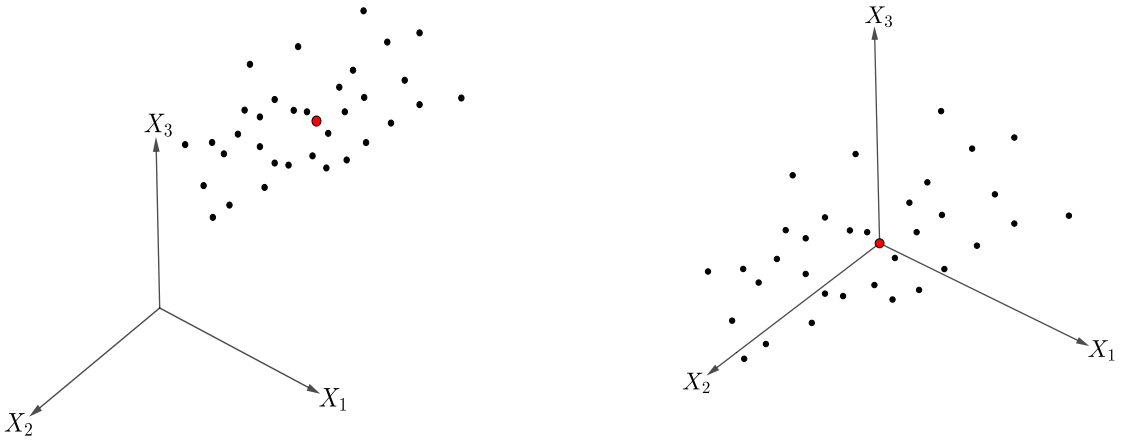
máxima variância amostral possível para um número k de componentes. Porém, como a matriz de covariâncias $\boldsymbol{\Sigma}$ não é conhecida, um estimador para essa matriz é requerido. Se $\boldsymbol{X}$ é centralizada previamente, um estimador para a matriz $\boldsymbol{\Sigma}$ é a matriz de covariâncias amostrais $\frac{1}{n-1}\boldsymbol{X}'\boldsymbol{X}$. Contudo, para se obter os autovalores e autovetores estimados de $\boldsymbol{\Sigma}$ é comum utilizar-se a matriz $\boldsymbol{X}'\boldsymbol{X}$. Jolliffe (2002, p. 31) salienta que os autovetores de $\frac{1}{n-1}\boldsymbol{X}'\boldsymbol{X}$ e $\boldsymbol{X}'\boldsymbol{X}$ são idênticos e os autovalores de $\frac{1}{n-1}\boldsymbol{X}'\boldsymbol{X}$ são simplesmente $\frac{1}{n-1}$ em relação aos autovalores de $\boldsymbol{X}'\boldsymbol{X}$. Por causa dessas relações é conveniente trabalhar em termos dos autovalores e autovetores de $\boldsymbol{X}'\boldsymbol{X}$.

Com o propósito de explorar algumas características para o cenário em que são utilizados dados amostrais, considere o caso mais simples contendo apenas três variáveis ($p = 3$). Para entender o princípio de métodos estatísticos exploratórios multidimensionais como a PCA, é útil representar geometricamente as n linhas e p colunas da matriz de observações $\boldsymbol{X}$ por pontos cujas coordenadas são precisamente os elementos dessa matriz. Essa representação também é denominada de nuvem de pontos (FIGURA 2.25a). Duas nuvens de pontos podem ser construídas, sendo uma no espaço $\mathbb{R}^p$ e outra no espaço $\mathbb{R}^n$. Como estamos lidando com n observações multivariadas, a nuvem de pontos considerada possui n indivíduos (linhas de $\boldsymbol{X}$) localizados no espaço p-dimensional $\mathbb{R}^p$ das variáveis (colunas de $\boldsymbol{X}$), uma vez que cada uma das n linhas é representada por um ponto com p coordenadas.

Um primeiro procedimento a ser feito na PCA consiste em padronizar as variáveis, pois é muito comum que as escalas das variáveis sejam diferentes. O primeiro passo então na PCA é mover os dados para o centro do sistema de coordenadas. Para cada uma das n amostras, isso é feito subtraindo-se a média amostral de cada observação. Esse procedimento é chamado de centralização à média. Por outro lado, em diversas situações as medições são realizadas em diferentes escalas. Logo, é razoável exigir que a inferência estatística a ser realizada seja independente das unidades de medida envolvidas. Nesse tipo de situação, os dados também podem ser escalonados pela divisão de cada observação pelo desvio padrão amostral. Esse procedimento faz com que cada uma das p variáveis possua variância unitária. Juntos, os procedimentos de centralização e redução a escala unitária são denominados de normalização ou padronização dos dados (FIGURA 2.25b). No procedimento de centralização à média, primeiro calculam-se as médias das variáveis. Esse vetor de médias é interpretável também como um ponto no espaço. O ponto (em vermelho) está situado no meio da nuvem de pontos, em seu centro de gravidade (FIGURA 2.25).

Figura 2.25 – Representação de uma nuvem de pontos no $\mathbb{R}^3$.

(a) Dados não padronizados. (b) Dados padronizados.



Fonte: Do autor (2020).

Uma consequência natural emerge na PCA quando o analista ou pesquisador padroniza os dados. A utilização das variáveis em sua forma padronizada faz com que os componentes não sejam mais obtidos pela matriz de covariâncias e sim pela matriz de correlações. Um fato interessante decorrente das variâncias tornarem-se unitárias é que isso evita a situação em que as variáveis com maior variabilidade dominem os componentes principais mais relevantes, o que eventualmente pode ocorrer quando se adota a matriz de covariâncias amostrais na análise. O procedimento de padronização, que permite que os dados estejam em uma escala de medida unitária, contribui para que a única fonte para a diferença entre os *loadings* em cada componente seja devida a correlação entre as variáveis.

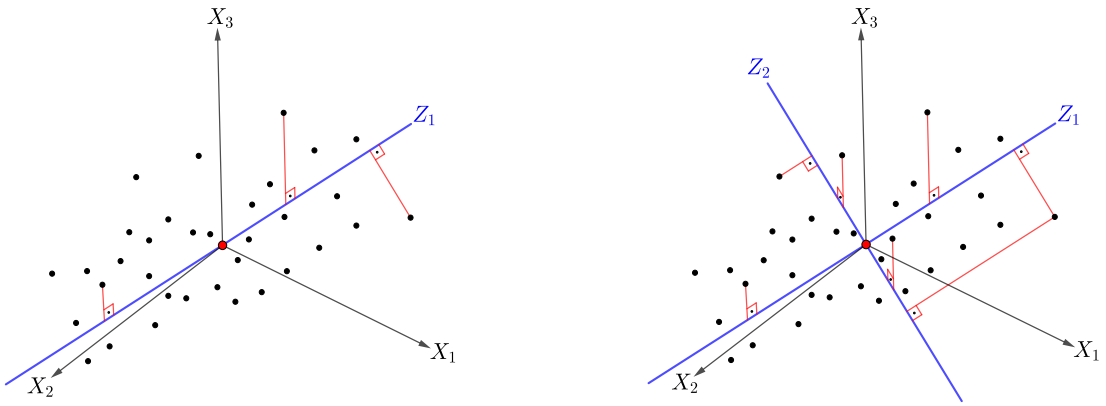
Após essa etapa, o passo seguinte consiste na obtenção do primeiro componente principal Z_1. A busca pelo componente Z_1 pode ser expressa como o componente que encontra-se na direção de variância máxima das projeções de cada ponto a Z_1. Mas qual o significado dessas projeções? Quando a direção de melhor ajuste é encontrada, pode-se marcar a localização de cada observação ao longo de tal direção, o que ocorre pela projeção ortogonal das observações ao respectivo componente (FIGURA 2.26a).

Ao primeiro componente principal é agora adicionado um segundo componente ao sistema. O segundo componente Z_2 é determinado de tal forma a ser perpendicular à direção do primeiro componente. Observe que esse vetor também passa na origem do sistema de coordenadas. A direção de Z_2 a ser encontrada deve assegurar a segunda maior variação nos valores de pontuação quando estes são projetados ortogonalmente nesse novo vetor (FIGURA 2.26b).

Figura 2.26 – Obtenção dos dois primeiros componentes principais considerando a nuvem de pontos no $\mathbb{R}^3$.

(a) Componente Z_1 que fornece a direção de maior variabilidade.

(b) Componente Z_2 que fornece a segunda direção de maior variabilidade.



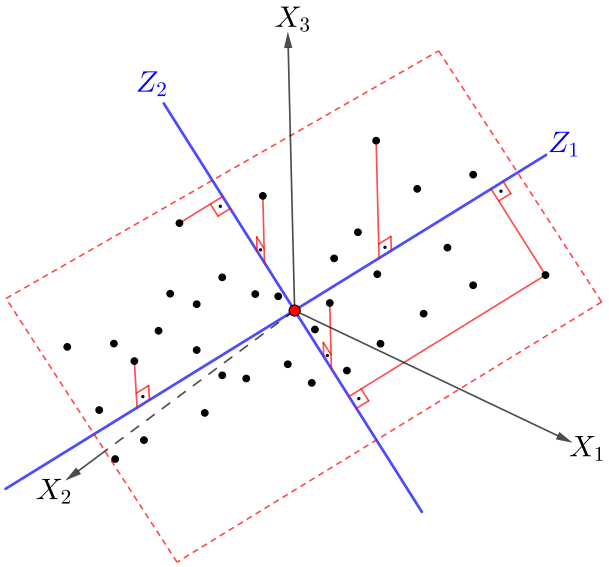
Fonte: Do autor (2020).

Em resumo, o componente principal Z_2 é calculado de modo a refletir a segunda maior fonte de variação nos dados, embora deva preservar a ortogonalidade ao primeiro componente Z_1.

Observe que os componentes principais Z_1 e Z_2 definem em conjunto um plano (FIGURA 2.27). Esse plano é o modelo de variável latente com dois componentes. Dessa forma, com um componente o modelo de variável latente é apenas uma reta, com dois componentes o modelo é um plano e com três ou mais componentes, o modelo é definido por um hiperplano. Esse hiperplano é a melhor aproximação que pode ser feita em relação aos dados originais. A distância perpendicular de cada ponto ao hiperplano é chamada de distância residual ou erro residual. O modelo de componentes principais é portanto um modelo de variáveis latentes, construído de tal forma que essas novas variáveis sejam orientadas na direção que fornece a maior variação das observações das variáveis originais.

Com base no teorema da decomposição espectral, na seção 2.7.2 foi possível expressar a matriz de covariâncias populacionais $\boldsymbol{\Sigma}$ em termos de seus autovalores e autovetores. Nesse sentido, os componentes principais amostrais também podem ser obtidos utilizando-se a decomposição espectral da matriz de covariâncias amostrais. Porém, uma vez que a matriz de covariâncias amostrais é definida como $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$ vezes uma constante multiplicativa $\frac{1}{n-1}$, nós podemos utilizar a decomposição em valores singulares de $\boldsymbol{X}$ como uma alternativa à decomposição espectral de $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$. Dada uma matriz $\boldsymbol{X}$ é possível expressar essa matriz em termos dos autovalores e autovetores de $\boldsymbol{X}'\boldsymbol{X}$ e $\boldsymbol{X}\boldsymbol{X}'$.

Figura 2.27 – Componentes principais Z_1 e Z_2 com destaque ao plano formado por eles.



Fonte: Do autor (2020).

Teorema 2.7.2 (Decomposição em valores singulares): Para uma matriz $\boldsymbol{X}_{n\times p}$ com posto k, existem matrizes $\boldsymbol{U}_{n\times k}$, $\boldsymbol{D}_{k\times k}$ e $\boldsymbol{V}_{p\times k}$ tais que:

i) $\boldsymbol{D}$ é uma matriz diagonal com $(\lambda_1,\lambda_2,\ldots,\lambda_k)$ na diagonal principal.

ii) $\boldsymbol{U}$ e $\boldsymbol{V}$ possuem colunas ortonormais.

iii) $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$.

Prova: Uma demonstração desse teorema não será fornecida aqui e pode ser vista em Ferreira (2018, p. 60-62).

Os valores $(\lambda_1,\lambda_2,\ldots,\lambda_k)$ são chamados valores singulares de $\boldsymbol{X}$. Esses elementos, presentes na matriz não singular $\boldsymbol{D}=\text{diag}\,(\lambda_1,\lambda_2,\ldots,\lambda_k)$ são raízes quadradas positivas de $\left(\lambda_1^2,\lambda_2^2,\ldots,\lambda_k^2\right)$, que são por sua vez os k primeiros autovalores não nulos de $\boldsymbol{X}\boldsymbol{X}'_{n\times n}$ ou $\boldsymbol{X}'\boldsymbol{X}_{p\times p}$. Do conhecimento das dimensões de $\boldsymbol{X}\boldsymbol{X}'$ e $\boldsymbol{X}'\boldsymbol{X}$ segue que essas matrizes possuem, respectivamente, n e p autovetores. Como $k\leq p<n$, as k colunas de $\boldsymbol{U}$ são os k primeiros autovetores normalizados de $\boldsymbol{X}\boldsymbol{X}'$, enquanto as k colunas de $\boldsymbol{V}$ são os k primeiros autovetores normalizados de $\boldsymbol{X}'\boldsymbol{X}$. Em ambos os casos, os autovetores correspondem aos autovalores $\left(\lambda_1^2,\lambda_2^2,\ldots,\lambda_k^2\right)$. Uma vez que as colunas de $\boldsymbol{U}$ e $\boldsymbol{V}$ são autovetores normalizados, então $\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}$ e $\boldsymbol{V}'\boldsymbol{V}=\boldsymbol{I}$. A igualdade em $k\leq p<n$ ocorre quando a matriz de observações $\boldsymbol{X}$ é de posto coluna completo, ou seja, $\text{rank}\,[\boldsymbol{X}]=p$. Quando isso ocorre, $\boldsymbol{X}\boldsymbol{X}'$ e $\boldsymbol{X}'\boldsymbol{X}$ passam a ter p autovalores não nulos $\left(\lambda_1^2,\lambda_2^2,\ldots,\lambda_p^2\right)$ e para a matriz quadrada $\boldsymbol{V}_{p\times p}$ tem-se que $\boldsymbol{V}\boldsymbol{V}'=\boldsymbol{I}$.

De acordo com Zou, Hastie e Tibshirani (2006), $\boldsymbol{Z} = \boldsymbol{UD}$ são os componentes principais e as colunas de $\boldsymbol{V}$ são os vetores de *loadings* correspondentes dos componentes principais. A variância do i-ésimo componente Z_i é dada por λ_i. Como $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$ é o estimador de $\boldsymbol{\Sigma}$ a menos de uma constante $\frac{1}{n-1}$, pode-se denotar os autovalores e autovetores amostrais como $\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p\right)$ e $(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \ldots, \hat{\boldsymbol{v}}_p)$, respectivamente.

As definições de proporção de variância de um componente principal e a proporção de variância acumulada até um componente principal são semelhantes ao caso populacional. Para comodidade do leitor, essas definições são novamente apresentadas com ligeiras modificações.

Definição 2.7.3 (Proporção de variância): A proporção de variância total amostral do vetor aleatório $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ que é explicada pelo k-ésimo componente principal $\hat{Z}_k$ é dada por:

$$\frac{\hat{\text{var}}\left[\hat{Z}_k\right]}{\hat{\text{var}}_{\text{total}}\left[\boldsymbol{X}\right]} = \frac{\hat{\text{var}}\left[\hat{Z}_k\right]}{\text{tr}\left[\hat{\boldsymbol{\Sigma}}\right]} = \frac{\hat{\lambda}_k}{\sum_{i=1}^{p} \hat{\lambda}_i} = \frac{\hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p}. \tag{2.125}$$

Definição 2.7.4 (Proporção de variância acumulada): A proporção de variância total do vetor aleatório $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ que é explicada pelos k primeiros componentes principais $\left(\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_k\right)$ é dada por:

$$\frac{\sum_{i=1}^{k} \hat{\text{var}}\left[\hat{Z}_i\right]}{\hat{\text{var}}_{\text{total}}\left[\boldsymbol{X}\right]} = \frac{\sum_{i=1}^{k} \hat{\text{var}}\left[\hat{Z}_i\right]}{\text{tr}\left[\hat{\boldsymbol{\Sigma}}\right]} = \frac{\sum_{i=1}^{k} \hat{\text{var}}\left[\hat{Z}_i\right]}{\sum_{i=1}^{p} \hat{\lambda}_i} = \frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p}. \tag{2.126}$$

A quantidade de proporção de variância diminuirá gradualmente do primeiro ao último componente e, consequentemente, em algum momento o interesse pelo acréscimo de mais um componente a um conjunto de k componentes não será interessante, pois apenas um pequeno número de componentes já será suficiente para se obterem as variâncias mais importantes. Em relação à notação, por não ser usual a utilização do "chapéu" para autovalores e autovetores, essa notação não será mais adotada ao longo do texto. O próprio contexto indicará se os autovalores e autovetores são amostrais ou populacionais.

2.7.4 Análise geométrica dos componentes principais

Nas seções que se seguem é apresentada a PCA utilizando uma abordagem geométrica. Na seção 2.7.5, os componentes principais são abordados no contexto populacional utilizando-se ideias geométricas como direção e projeção de vetores. Ainda que a abordagem geométrica

nessa breve seção seja predominante, “ferramentas” e conceitos do cálculo diferencial e integral como multiplicadores de Lagrange, derivadas, ponto de sela, máximo e mínimo também são empregados. Na seção 2.7.6 são apresentadas algumas propriedades dos componentes principais populacionais e amostrais que se encontram em Jolliffe (2002, p. 13-17). Em relação a essas propriedades buscou-se apresentar provas de caráter estritamente geométrico, que podem ser consideradas contribuições ao texto de Jolliffe (2002). Na seção 2.7.7 é abordado brevemente como os componentes principais podem ser obtidos no contexto da regressão linear. Essa seção é apresentada como ensejo para o desenvolvimento das seções 2.8 a 2.8.2, nas quais discuti-se como os componentes principais podem ser modificados como um problema de regressão ao se utilizar os métodos *Ridge*, LASSO e *Elastic Net*. O desenvolvimento teórico até a seção 2.8.2 é importante para a compreensão dos fundamentos que culminam com a proposta dos novos métodos apresentados neste trabalho, que baseiam-se nos métodos de regressão OSCAR e PACS.

2.7.5 Componentes principais populacionais

A teoria de componentes principais está relacionada ao seguinte problema básico: dado um vetor aleatório $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$, qual a direção em que a combinação linear das variáveis desse vetor apresenta a maior variação? Geometricamente, o problema pode ser descrito da maneira apresentada a seguir. Dada uma direção $\boldsymbol{v}_1 \in \mathbb{R}^p$, um vetor de norma unitária ($\|\boldsymbol{v}_1\| = 1$), projeta-se $\boldsymbol{X}$ em $\boldsymbol{v}_1$ obtendo-se a variável aleatória unidimensional $Z_1 = \boldsymbol{v}_1 \cdot \boldsymbol{X} = \|\boldsymbol{X}\| \cos(\theta_1)$, que é totalmente descrita pela norma de $\boldsymbol{X}$ e pelo ângulo θ_1 (FIGURA 2.28). Essa nova variável consiste no produto escalar de $\boldsymbol{X}$ e $\boldsymbol{v}_1$, ou seja, na soma dos produtos termo a termo dos componentes de $\boldsymbol{X}$ com $\boldsymbol{v}_1$.

A variância de Z_1 está relacionada ao vetor $\boldsymbol{X}$ na direção de $\boldsymbol{v}_1$. De fato,

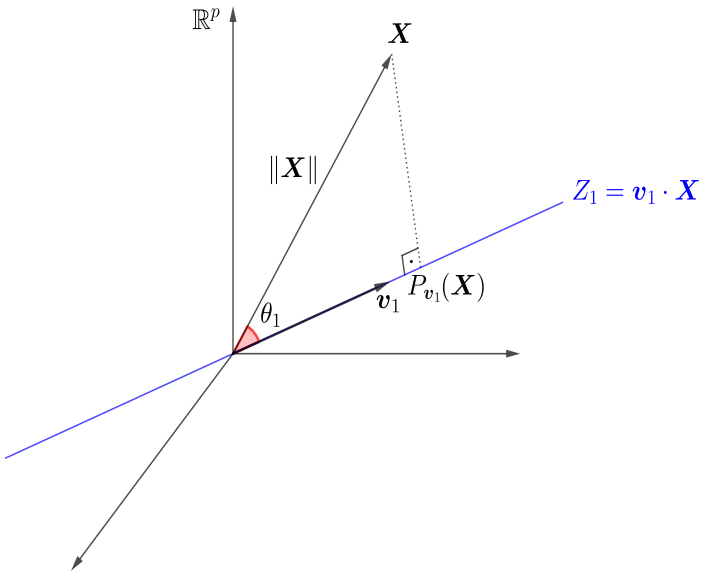
$$\operatorname{var}[Z_1] = \operatorname{var}[\boldsymbol{v}_1 \cdot \boldsymbol{X}] = \operatorname{var}\left[\boldsymbol{v}_1' \boldsymbol{X}\right] = \boldsymbol{v}_1' \operatorname{cov}[\boldsymbol{X}]\, \boldsymbol{v}_1 = \boldsymbol{v}_1' \boldsymbol{\Sigma} \boldsymbol{v}_1. \tag{2.127}$$

Dessa forma, o objetivo é maximizar $\operatorname{var}[\boldsymbol{v}_1 \cdot \boldsymbol{X}] = \boldsymbol{v}_1' \boldsymbol{\Sigma} \boldsymbol{v}_1$ sujeita a restrição $\|\boldsymbol{v}_1\| = 1$. Para atingir esse objetivo podem ser empregados dois métodos de maximização, sendo o primeiro utilizando os multiplicadores de Lagrange e outro fazendo-se uso do cálculo diferencial. Em relação ao primeiro método pode-se definir a função lagrangeana H_1, que é dada por:

$$H_1(\boldsymbol{v}_1, \lambda_1) = \boldsymbol{v}_1' \boldsymbol{\Sigma} \boldsymbol{v}_1 - \lambda_1 \left(\|\boldsymbol{v}_1\|^2 - 1 \right), \tag{2.128}$$

em que λ_1 é não negativo.

Figura 2.28 – Representação geométrica da obtenção do componente principal Z_1.



Fonte: Adaptado de Silveira (2014).

Como $\|\boldsymbol{v}_1\|^2 = \boldsymbol{v}_1'\boldsymbol{v}_1$, tem-se que $\frac{\partial}{\partial \boldsymbol{v}_1}(\boldsymbol{v}_1'\boldsymbol{I}\boldsymbol{v}_1) = 2\boldsymbol{I}\boldsymbol{v}_1 = 2\boldsymbol{v}_1$. Dessa maneira, derivando-se a função lagrangeana H_1 em relação a $\boldsymbol{v}_1$ e igualando a derivada resultante ao vetor de zeros, segue que:

$$\frac{\partial H_1}{\partial \boldsymbol{v}_1} = 2\boldsymbol{\Sigma}\boldsymbol{v}_1 - 2\lambda_1\boldsymbol{v}_1 = \boldsymbol{0} \Rightarrow \boldsymbol{\Sigma}\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1. \tag{2.129}$$

Portanto, o vetor de pontos críticos $\boldsymbol{v}_1$ é um autovetor da matriz $\boldsymbol{\Sigma}$. Por sua vez, como

$$\boldsymbol{v}_1'\boldsymbol{\Sigma}\boldsymbol{v}_1 = \boldsymbol{v}_1'(\boldsymbol{\Sigma}\boldsymbol{v}_1) = \boldsymbol{v}_1'\lambda_1\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1'\boldsymbol{v}_1 = \lambda_1, \tag{2.130}$$

mostra-se que o valor máximo é alcançado com o maior autovalor de $\boldsymbol{\Sigma}$. Desse modo, o autovalor $\boldsymbol{v}_1$ está associado ao maior autovalor λ_1 de $\boldsymbol{\Sigma}$. Com isso, concluí-se que $\boldsymbol{v}_1$ gera o subespaço Z_1, chamado de primeiro componente principal.

Caso deseje-se um subespaço bidimensional com características semelhantes de ajuste ao subespaço anterior e que eventualmente o contenha, deve-se procurar um segundo vetor $\boldsymbol{v}_2$ que maximize $\boldsymbol{v}_2'\boldsymbol{\Sigma}\boldsymbol{v}_2$, de tal forma que $\|\boldsymbol{v}_2\|^2 = 1$ e que seja ortogonal a $\boldsymbol{v}_1$. Se $Z_2 = \boldsymbol{v}_2 \cdot \boldsymbol{X}$, deseja-se maximizar $\text{var}[\boldsymbol{v}_2 \cdot \boldsymbol{X}] = \boldsymbol{v}_2'\boldsymbol{\Sigma}\boldsymbol{v}_2$ sujeita as restrições $\boldsymbol{v}_2'\boldsymbol{v}_2 = 1$ e $\boldsymbol{v}_2'\boldsymbol{v}_1 = 0$.

Seja H_2 a função lagrangeana dada por:

$$H_2(\boldsymbol{v}_2, \lambda_2, \psi) = \boldsymbol{v}_2'\boldsymbol{\Sigma}\boldsymbol{v}_2 - \lambda_2(\boldsymbol{v}_2'\boldsymbol{v}_2 - 1) - \psi\boldsymbol{v}_2'\boldsymbol{v}_1, \tag{2.131}$$

em que λ_2 e ψ são os multiplicadores de Lagrange.

Derivando a função lagrangeana H_2 em relação a $\boldsymbol{v}_2$ e igualando a derivada resultante ao vetor de zeros, resulta que:

$$\frac{\partial H_2}{\partial \boldsymbol{v}_2} = 2\boldsymbol{\Sigma}\boldsymbol{v}_2 - 2\lambda_2\boldsymbol{v}_2 - \psi\boldsymbol{v}_1 = \mathbf{0}. \tag{2.132}$$

Pré-multiplicando a igualdade anterior por $\boldsymbol{v}_1'$ vem que:

$$2\boldsymbol{v}_1'\boldsymbol{\Sigma}\boldsymbol{v}_2 - 2\lambda_2\boldsymbol{v}_1'\boldsymbol{v}_2 - \psi\boldsymbol{v}_1'\boldsymbol{v}_1 = 0. \tag{2.133}$$

De $\boldsymbol{\Sigma}\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1$ e da simetria de $\boldsymbol{\Sigma}$ segue que $\boldsymbol{v}_1'\boldsymbol{\Sigma} = \lambda_1\boldsymbol{v}_1'$. Do fato de $\boldsymbol{v}_1$ ser unitário ($\boldsymbol{v}_1'\boldsymbol{v}_1 = 1$) e da ortogonalidade entre $\boldsymbol{v}_1$ e $\boldsymbol{v}_2$ ($\boldsymbol{v}_1'\boldsymbol{v}_2 = 0$) segue para (2.133) que:

$$2\lambda_1\boldsymbol{v}_1'\boldsymbol{v}_2 - 0 - \psi = 0 \Rightarrow \psi = 0. \tag{2.134}$$

Desse modo,

$$\frac{\partial H_2}{\partial \boldsymbol{v}_2} = 2\boldsymbol{\Sigma}\boldsymbol{v}_2 - 2\lambda_2\boldsymbol{v}_2 = \mathbf{0} \Rightarrow \boldsymbol{\Sigma}\boldsymbol{v}_2 = \lambda_2\boldsymbol{v}_2. \tag{2.135}$$

Pode-se concluir que $\boldsymbol{v}_2$ é o segundo autovetor, correspondendo a λ_2, o segundo maior autovalor de $\boldsymbol{\Sigma}$. Como consequência, $\boldsymbol{v}_2$ gera uma reta ortogonal a reta gerada por $\boldsymbol{v}_1$, chamada de segundo componente principal (FIGURA 2.29). Além disso, $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ geram o subespaço de dimensão dois que "melhor" se ajusta aos dados. Com base em um procedimento análogo é possível se obter um subespaço de dimensão $k \leq p$, gerado pelos k autovetores associados aos k maiores autovalores de $\boldsymbol{\Sigma}$. Finalmente, dessa exposição decorre que a definição dos componentes principais consiste na obtenção dos autovalores e autovetores da matriz $\boldsymbol{\Sigma}$.

■

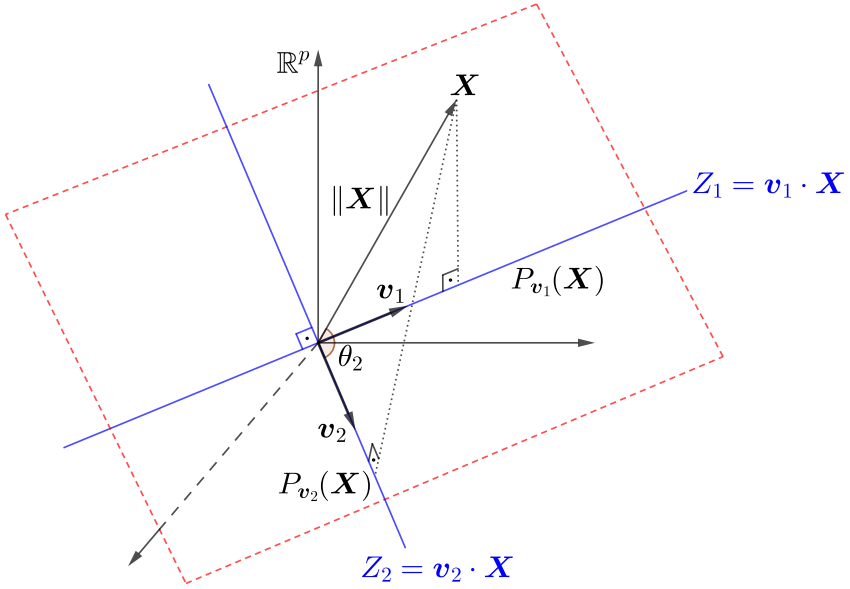
Neste ponto, torna-se relevante um comentário acerca da importância da restrição $\|\boldsymbol{v}\| = 1$ na PCA. Para isso, considere que a forma quadrática $\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v}$ é expressa da seguinte forma:

$$\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v} = \sum_{j=1}^{p}\sigma_{jj}v_j^2 + 2\sum_{j=1}^{p-1}\sum_{k=j+1}^{p}\sigma_{jk}v_jv_k = \sum_{j=1}^{p}\sum_{k=1}^{p}\sigma_{jk}v_jv_k, \tag{2.136}$$

em que σ_{jj} é a variância da variável X_j e σ_{jk} é a covariância entre as variáveis X_j e X_k.

Com base em (2.136) pode ser observado que a forma quadrática $\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v}$ não possui máximo, conforme ocorra um aumento nos valores de $\boldsymbol{v}$. Uma vez que um dos objetivos na PCA é maximizar as variâncias dos componentes principais, sem a restrição $\|\boldsymbol{v}\| = 1$ o valor máximo da variância de um componente Z não existiria, uma vez que a medida que os valores de $\boldsymbol{v}$ cres-

Figura 2.29 – Representação geométrica da obtenção do componente principal Z_2.



Fonte: Do autor (2020).

cessem sua variância λ tenderia ao infinito. Logo, a imposição da restrição da norma unitária ao vetor $\boldsymbol{v}$ justifica-se por assegurar a maximização da forma quadrática $\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v}$.

Uma outra demonstração pode ser feita com base no conceito abstrato de curvas parametrizadas, que são úteis para descrever as trajetórias e movimentos de objetos que se deslocam em um plano ou no espaço. Seja $\boldsymbol{r}'(t) = (r_1(t), r_2(t) \ldots, r_p(t))$ uma curva parametrizada e centrada na origem do espaço $\mathbb{R}^p$ com $\|\boldsymbol{r}(t)\| = 1$, isto é, uma curva cujo vetor posição $\boldsymbol{r}(t)$ tem comprimento constante igual ao raio unitário da hiperesfera (FIGURA 2.30). Vamos também supor que essa curva esteja parametrizada pelo comprimento do arco, isto é, $\left\| \frac{d}{dt}[\boldsymbol{r}(t)] \right\| = 1$. Definindo a função $f(t) = g(\boldsymbol{r}(t)) = \boldsymbol{r}'(t)\boldsymbol{\Sigma}\boldsymbol{r}(t)$, em que $\boldsymbol{r}$ é derivável em t e g é derivável em $\boldsymbol{r}$, deseja-se então os pontos críticos de f.

Com base na regra da cadeia, derivando a função f em relação a t e igualando a derivada resultante a zero vem que:

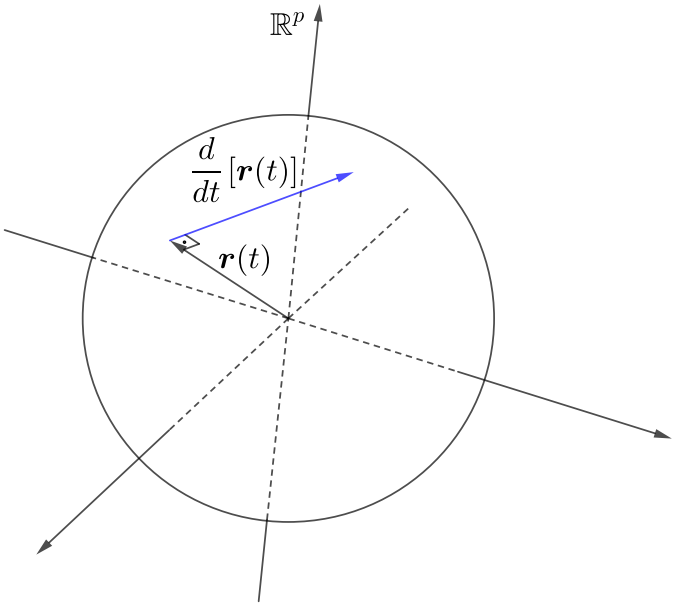
$$\frac{d}{dt}[f(t)] = \frac{d}{d\boldsymbol{r}}[g(\boldsymbol{r}(t))] \cdot \frac{d}{dt}[\boldsymbol{r}(t)] = 0. \tag{2.137}$$

Como $\frac{d}{d\boldsymbol{r}}[g(\boldsymbol{r}(t))] = 2\boldsymbol{\Sigma}\boldsymbol{r}(t)$ e utilizando a notação do transposto para o produto interno, tem-se para $t = 0$ que:

$$\frac{d}{dt}[f(0)] = 2\boldsymbol{\Sigma}[\boldsymbol{r}(0)]' \frac{d}{dt}[\boldsymbol{r}(0)] = 0. \tag{2.138}$$

Decorre portanto que o vetor $\frac{d}{dt}[\boldsymbol{r}(0)]$ é perpendicular ao vetor $\boldsymbol{\Sigma}\boldsymbol{r}(0)$. Como tal fato é verdadeiro para qualquer curva que passa em $\boldsymbol{r}(0)$, isso significa que $\boldsymbol{\Sigma}\boldsymbol{r}(0)$ é um vetor per-

Figura 2.30 – hiperesfera de raio unitário e centrada na origem do espaço $\mathbb{R}^p$, parametrizada pelo vetor posição $\boldsymbol{r}(t)$.



Fonte: Do autor (2020).

pendicular ao espaço tangente da hiperesfera em $\boldsymbol{r}(0)$. Dessa última afirmação decorre que os vetores $\boldsymbol{\Sigma r}(0)$ e $\boldsymbol{r}(0)$ são paralelos e, neste caso, $\boldsymbol{\Sigma r}(0)$ é um múltiplo de $\boldsymbol{r}(0)$. Portanto, existe uma constante real λ tal que $\boldsymbol{\Sigma r}(0) = \lambda \boldsymbol{r}(0)$, com $\lambda \neq 0$. Assim, os pontos críticos de f são justamente os autovetores de $\boldsymbol{\Sigma}$.

A natureza dos pontos críticos pode ser estudada obtendo-se a derivada segunda da função f. Em primeiro lugar, considerando um t geral, segue de (2.138) que a derivada primeira de f é:

$$\frac{d}{dt}[f(t)] = 2[\boldsymbol{\Sigma r}(t)]' \frac{d}{dt}[\boldsymbol{r}(t)] = 2\boldsymbol{r}'(t)\boldsymbol{\Sigma}\frac{d}{dt}[\boldsymbol{r}(t)]. \tag{2.139}$$

A partir de (2.139) a derivada segunda de f é dada por:

$$\frac{d^2}{dt^2}[f(t)] = 2\frac{d}{dt}[\boldsymbol{r}'(t)]\boldsymbol{\Sigma}\frac{d}{dt}[\boldsymbol{r}(t)] + 2\boldsymbol{r}'(t)\boldsymbol{\Sigma}\frac{d^2}{dt^2}[\boldsymbol{r}(t)]. \tag{2.140}$$

Agora iremos mostrar que $\frac{d^2}{dt^2}[\boldsymbol{r}(t)]$ também é um autovetor de $\boldsymbol{\Sigma}$. Isso equivale a mostrar que $\boldsymbol{r}(t)$ e $\frac{d^2}{dt^2}[\boldsymbol{r}(t)]$ são vetores paralelos. Para isso, considere a suposição inicial de que $\|\boldsymbol{r}(t)\| = 1$, ou seja, $\boldsymbol{r}'(t)\boldsymbol{r}(t) = 1$. Dessa maneira, derivando $\boldsymbol{r}'(t)\boldsymbol{r}(t) = 1$ em relação a t tem-se que:

$$\frac{d}{dt}[\boldsymbol{r}'(t)]\boldsymbol{r}(t) + \boldsymbol{r}'(t)\frac{d}{dt}[\boldsymbol{r}(t)] = 0. \tag{2.141}$$

Uma vez que $\frac{d}{dt}[\boldsymbol{r}'(t)]\boldsymbol{r}(t)$ e $\boldsymbol{r}'(t)\frac{d}{dt}[\boldsymbol{r}(t)]$ são iguais, então:

$$2\frac{d}{dt}\left[\boldsymbol{r}'(t)\right]\boldsymbol{r}(t)=0\Rightarrow\frac{d}{dt}\left[\boldsymbol{r}'(t)\right]\boldsymbol{r}(t)=0$$
$$\Rightarrow\left[\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\right]'\boldsymbol{r}(t)=0. \tag{2.142}$$

Consequentemente, $\boldsymbol{r}(t)\perp\frac{d}{dt}\left[\boldsymbol{r}(t)\right]$. O vetor velocidade $\frac{d}{dt}\left[\boldsymbol{r}(t)\right]$ é tangente à hiperesfera e, por consequência, perpendicular a $\boldsymbol{r}(t)$. Esse é sempre o caso para uma função vetorial derivável de comprimento constante: o vetor e sua primeira derivada são ortogonais (FIGURA 2.30).

Uma vez que $\left\|\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\right\|=1$ tem-se então que $\left[\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\right]'\frac{d}{dt}\left[\boldsymbol{r}(t)\right]=1$. Derivando a última expressão em relação a t resulta que:

$$\left[\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]\right]'\frac{d}{dt}\left[\boldsymbol{r}(t)\right]+\left[\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\right]'\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=0\Rightarrow 2\left[\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]\right]'\frac{d}{dt}\left[\boldsymbol{r}(t)\right]=0.$$
$$\Rightarrow\left[\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]\right]'\frac{d}{dt}\left[\boldsymbol{r}(t)\right]=0. \tag{2.143}$$

Desse modo, $\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\perp\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$. Como $\boldsymbol{r}(t)\perp\frac{d}{dt}\left[\boldsymbol{r}(t)\right]$ e $\frac{d}{dt}\left[\boldsymbol{r}(t)\right]\perp\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$, segue então que $\boldsymbol{r}(t)$ é paralelo a $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$ (FIGURA 2.31). Dessa maneira, existe uma constante real c, não nula, tal que $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=c\boldsymbol{r}(t)$. A constante c também é negativa, pois $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$ é a aceleração centrípeta , ou seja, esse vetor atua na mesma direção do vetor posição $\boldsymbol{r}(t)$, porém no sentido oposto (FIGURA 2.31).

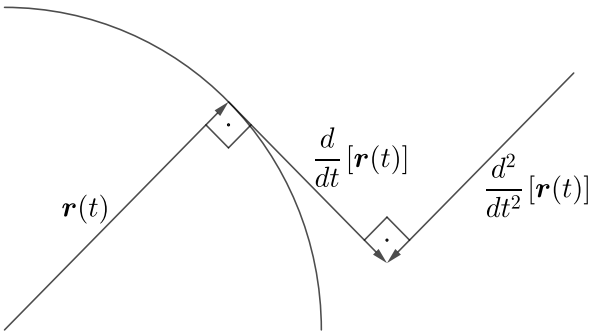
Como $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=c\boldsymbol{r}(t)$ então $\boldsymbol{r}(t)=\frac{1}{c}\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$. Por conseguinte,

$$\boldsymbol{\Sigma}\boldsymbol{r}(t)=\lambda\boldsymbol{r}(t)\Rightarrow\frac{1}{c}\boldsymbol{\Sigma}\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=\frac{1}{c}\lambda\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$$
$$\Rightarrow\boldsymbol{\Sigma}\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=\lambda\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]. \tag{2.144}$$

Portanto, $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$ também é um autovetor de $\boldsymbol{\Sigma}$, como queríamos mostrar. Além disso, $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$ está associado ao mesmo autovalor λ de $\boldsymbol{r}(t)$. De posse desse resultado, pode-se concluir o estudo da natureza dos pontos críticos da função $f(t)=\boldsymbol{r}'(t)\boldsymbol{\Sigma}\boldsymbol{r}(t)$ por meio de sua derivada segunda obtida em (2.140).

Como $\boldsymbol{\Sigma}\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=\lambda\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]$ e $\frac{d^2}{dt^2}\left[\boldsymbol{r}(t)\right]=c\boldsymbol{r}(t)$ tem-se para (2.140) que:

Figura 2.31 – Ortogonalidade dos vetores na curva parametrizada $\boldsymbol{r}(t)$.



Fonte: Do autor (2020).

$$\begin{aligned}\frac{d^2}{dt^2}[f(t)] &= 2\frac{d}{dt}\left[\boldsymbol{r}'(t)\right]\boldsymbol{\Sigma}\frac{d}{dt}[\boldsymbol{r}(t)] + 2\boldsymbol{r}'(t)\boldsymbol{\Sigma}\frac{d^2}{dt^2}[\boldsymbol{r}(t)] \\ &= 2\frac{d}{dt}\left[\boldsymbol{r}'(t)\right]\boldsymbol{\Sigma}\frac{d}{dt}[\boldsymbol{r}(t)] + 2\lambda\boldsymbol{r}'(t)\frac{d^2}{dt^2}[\boldsymbol{r}(t)] \\ &= 2\frac{d}{dt}\left[\boldsymbol{r}'(t)\right]\boldsymbol{\Sigma}\frac{d}{dt}[\boldsymbol{r}(t)] + 2\lambda c\boldsymbol{r}'(t)\boldsymbol{r}(t) \\ &= 2\left[\frac{d}{dt}[\boldsymbol{r}(t)]\right]'\boldsymbol{\Sigma}\left[\frac{d}{dt}[\boldsymbol{r}(t)]\right] + 2\lambda c, \end{aligned} \tag{2.145}$$

pois $\boldsymbol{r}'(t)\boldsymbol{r}(t) = 1$ para todo t.

Caso exista, o valor máximo da função f será:

$$f(t_0) = g(\boldsymbol{r}(t_0)) = \boldsymbol{r}'(t_0)\boldsymbol{\Sigma}\boldsymbol{r}(t_0) = \boldsymbol{r}'(t_0)\lambda\boldsymbol{r}(t_0) = \lambda, \tag{2.146}$$

para um valor t_0 qualquer.

A condição de existência desse máximo da função f, um autovalor de $\boldsymbol{\Sigma}$, para os valores críticos $\boldsymbol{r}(t_0)$ é que $\frac{d^2}{dt^2}[f(t_0)]\Big|_{t=t_0} < 0$. Como $\boldsymbol{\Sigma}$ é positiva semidefinida, resulta para todo t que $\left[\frac{d}{dt}[\boldsymbol{r}(t)]\right]'\boldsymbol{\Sigma}\left[\frac{d}{dt}[\boldsymbol{r}(t)]\right] \geq 0$. Os autovalores de $\boldsymbol{\Sigma}$ são não negativos pelo mesmo motivo.

Desse modo,

$$\begin{aligned}\frac{d^2}{dt^2}[f(t_0)]\Big|_{t=t_0} < 0 &\Rightarrow 2\left[\frac{d}{dt}[\boldsymbol{r}(t_0)]\right]'\boldsymbol{\Sigma}\left[\frac{d}{dt}[\boldsymbol{r}(t_0)]\right] + 2\lambda c < 0 \\ &\Rightarrow \left[\frac{d}{dt}[\boldsymbol{r}(t_0)]\right]'\boldsymbol{\Sigma}\left[\frac{d}{dt}[\boldsymbol{r}(t_0)]\right] < -\lambda c. \end{aligned} \tag{2.147}$$

Como exemplo, a curva $\boldsymbol{r}(t)$ pode ser pensada como $\boldsymbol{r}(t) = \cos(t)\,i + \operatorname{sen}(t)\,j$. Logo,

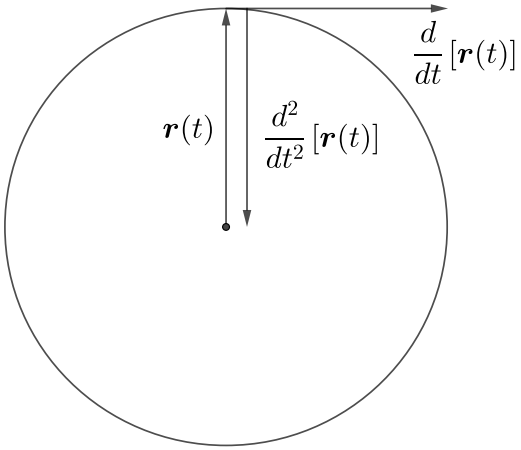
$$\frac{d^2}{dt^2}[\boldsymbol{r}(t)] = -\cos(t)\,i - \operatorname{sen}(t)\,j = -\boldsymbol{r}(t). \tag{2.148}$$

Dessa forma, $c = -1$. Considerando (2.147), para $t = 0$ o autovetor $\boldsymbol{r}(0)$ é um ponto de máximo se:

$$\frac{d^2}{dt^2}\left[f(0)\right]\Big|_{t=0} < 0 \Rightarrow \left[\frac{d}{dt}\left[\boldsymbol{r}(0)\right]\right]' \boldsymbol{\Sigma} \left[\frac{d}{dt}\left[\boldsymbol{r}(0)\right]\right] < \lambda, \tag{2.149}$$

para todo vetor $\frac{d}{dt}\left[\boldsymbol{r}(0)\right]$ no espaço tangente da hiperesfera $\mathbb{S}^p$ em $\boldsymbol{r}(0)$ (FIGURA 2.32).

Figura 2.32 – Representação geométrica da curva parametrizada pelo comprimento do arco sobre a hiperesfera de raio unitário, centrada na origem do espaço $\mathbb{R}^p$.



Fonte: Adaptado de Silveira (2014).

Com os autovetores $\boldsymbol{r}$ de $\boldsymbol{\Sigma}$, os componentes principais são obtidos como as combinações lineares $\boldsymbol{r} \cdot \boldsymbol{X}$. Pode-se concluir então que, para o maior autovalor tem-se o máximo, para o menor autovalor o mínimo e os outros pontos críticos são pontos de sela.

■

2.7.6 Algumas propriedades dos componentes principais populacionais e amostrais

A seguir são apresentadas demonstrações de caráter geométrico, alternativas às demonstrações apresentadas em Jolliffe (2002, p. 11-17), em relação a algumas propriedades dos componentes principais populacionais e amostrais.

Propriedade 2.7.1: Para qualquer inteiro q, $1 \leq q \leq p$, considere a transformação linear $\boldsymbol{B}_{p\times q}$ com colunas ortonormais tal que

$$\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x},$$

em que $\boldsymbol{y}$ é um vetor $(q \times 1)$ com componentes que são combinações lineares das componentes do vetor $\boldsymbol{x}$ de dimensão $(p \times 1)$ e $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_{q\times q}$. Seja $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{B}'\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{B}$ a matriz $(q \times q)$ de

variâncias e covariâncias para $\boldsymbol{y}$. Então, o traço de $\boldsymbol{\Sigma_y}$, denotado $\mathrm{tr}[\boldsymbol{\Sigma_y}]$, é maximizado por $\boldsymbol{B} = \boldsymbol{A}_q$, em que as colunas de $\boldsymbol{A}_q$ são os q primeiros autovetores da matriz $\boldsymbol{\Sigma_x}$ ordenados pelos autovalores.

Prova:

Vamos completar a matriz $\boldsymbol{B}$ com $p-q$ colunas para torná-la uma matriz ortogonal $\boldsymbol{C}_{p\times p}$. Considerando $\boldsymbol{C}$ como uma transformação linear $\boldsymbol{C}:\mathbb{R}^p \to \mathbb{R}^p$, tem-se que $\boldsymbol{C}$ é uma isometria. Portanto, o elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_x}\boldsymbol{x} \leq c_1$ em $\mathbb{R}^p$ é levado por uma rotação no elipsoide $(\boldsymbol{Cx})'\boldsymbol{\Sigma_x}(\boldsymbol{Cx}) \leq c_2$, o que implica que $\boldsymbol{x}'\boldsymbol{C}'\boldsymbol{\Sigma_x}\boldsymbol{Cx} \leq c_2$. Esse novo elipsoide, restrito a $\mathbb{R}^q$, é exatamente o elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_y}\boldsymbol{x} \leq c_3$. De fato,

$$\boldsymbol{x}'\boldsymbol{C}'\boldsymbol{\Sigma_x}\boldsymbol{Cx} = \boldsymbol{x}'\left(\boldsymbol{C}'\boldsymbol{\Sigma_x}\boldsymbol{C}\right)\boldsymbol{x} = \boldsymbol{x}'\left(\boldsymbol{B}'\boldsymbol{\Sigma_x}\boldsymbol{B}\right)\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{\Sigma_y}\boldsymbol{x} \leq c_3, \tag{2.150}$$

em que devido a restrição ao $\mathbb{R}^q$ as matrizes $\boldsymbol{C}$ e $\boldsymbol{\Sigma_x}$ tem dimensões $(q\times q)$ e $\boldsymbol{x}$ é um vetor $(q\times 1)$. A igualdade das matrizes $\boldsymbol{B}$ e $\boldsymbol{C}$ também é devida a essa restrição.

O novo elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_y}\boldsymbol{x} \leq c_3$ possui as mesmas dimensões que o elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_x}\boldsymbol{x} \leq c_1$. Observe que a construção inversa também pode ser feita: dada uma matriz ortogonal $\boldsymbol{C}_{p\times p}$, para se obter a matriz $\boldsymbol{B}_{p\times q}$ tomam-se as q primeiras colunas de $\boldsymbol{C}$. Dessa forma, vamos escolher uma matriz $\boldsymbol{C}_{p\times p}$ que define uma rotação de tal maneira que os q maiores eixos principais do elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_x}\boldsymbol{x}$ sejam coincidentes com os eixos coordenados de $\mathbb{R}^q$. Se $\boldsymbol{B}$ é a matriz obtida pelas q primeiras colunas de $\boldsymbol{C}$, o elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma_x}\boldsymbol{x}$ em $\mathbb{R}^q$, definido por $\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x}$, possui a maior soma para as variâncias presentes na diagonal principal de $\boldsymbol{\Sigma_y}$, isto é, tal que o traço de $\boldsymbol{\Sigma_y}$ é máximo. Ora, a matriz $\boldsymbol{C}$ que leva os eixos principais do elipsóide $\boldsymbol{x}'\boldsymbol{\Sigma_x}\boldsymbol{x} \leq c_1$ nos eixos coordenados é justamente a matriz em que as colunas são os autovetores de $\boldsymbol{\Sigma_x}$, isto é, na notação de Jolliffe (2002, p. 11), a matriz $\boldsymbol{A}$. Portanto, $\boldsymbol{B} = \boldsymbol{A}_q$ é a matriz que maximiza o traço de $\boldsymbol{\Sigma_y}$, obtida pelos q primeiros autovetores da matriz $\boldsymbol{\Sigma_x}$. Esses autovetores são ordenados pelos autovalores associados da matriz $\boldsymbol{\Sigma_x}$.

■

A construção acima também prova uma outra propriedade importante.

Propriedade 2.7.2: Considere a transformação $\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x}$. Se $\left|\boldsymbol{\Sigma_y}\right|$ denota o determinante da matriz de covariâncias de $\boldsymbol{y}$, então o $\left|\boldsymbol{\Sigma_y}\right|$ é maximizado quando $\boldsymbol{B} = \boldsymbol{A}_q$.

Prova:

Novamente, completando $\boldsymbol{B}$ para uma transformação ortogonal $\boldsymbol{C}_{p\times p}$, o elipsoide $\boldsymbol{y}'\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{y} \leq c$ é a intersecção da imagem do elipsoide $\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{x}$ com o subespaço $\mathbb{R}^q$. O determinante de $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ é proporcional ao produto dos eixos principais do elipsoide $\boldsymbol{y}'\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{y} \leq c$. Portanto, o determinante é também maximizado se os maiores eixos estiverem contidos em $\mathbb{R}^q$. Em particular, se os maiores eixos forem coincidentes com os eixos coordenados de $\mathbb{R}^q$. Tal fato é obtido novamente se $\boldsymbol{B} = \boldsymbol{A}_q$.

■

Propriedade 2.7.3: Suponha que os vetores $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ são transformados da forma $\boldsymbol{y}_i = \boldsymbol{B}'\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, em que $\boldsymbol{B}$ é uma matriz $(p \times q)$ com colunas ortonormais, então $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ são projeções de $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ em um subespaço q-dimensional. Uma medida de qualidade de ajuste desse subespaço q-dimensional para $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ pode ser definida como uma soma de distâncias perpendiculares de $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ para o subespaço. Esta medida é minimizada quando $\boldsymbol{B} = \boldsymbol{A}_q$, obtida pelos q primeiros autovetores da matriz de covariâncias amostrais.

Prova:

A questão é obter o subespaço vetorial q-dimensional o mais perto possível, no sentido de quadrado das distâncias perpendiculares dos vetores observados. Considere os subespaços q-dimensionais como imagens das transformações lineares de posto q, $\boldsymbol{B} : \mathbb{R}^q \to \mathbb{R}^p$. Nesse caso, sabemos que a projeção ortogonal no subespaço $\mathrm{Im}(\boldsymbol{B})$ é $\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$ e a soma dos quadrados das distâncias perpendiculares é dada por:

$$\begin{aligned}
\sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i \right\|^2 &= \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i\right]' \left[\boldsymbol{x}_i - \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i' - \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\right] \left[\boldsymbol{x}_i - \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i'\boldsymbol{x}_i - \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i - \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i'\boldsymbol{x}_i - \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i - \sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{x}_i, \qquad (2.151)
\end{aligned}$$

em que o vetor $\boldsymbol{x}_i$ $(p \times 1)$ corresponde a i-ésima observação multivariada.

Podemos observar que cada vetor $\boldsymbol{x}_i$ corresponde a uma linha da matriz de observações $\boldsymbol{X}$. Nesse caso,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}. \tag{2.152}$$

Como consequência,

$$\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{B} \left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1} \boldsymbol{B}' \boldsymbol{x}_i = \operatorname{tr}\left[\boldsymbol{X}\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{X}'\right]. \tag{2.153}$$

Para minimizar a soma em (2.151), basta maximizar $\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{B} (\boldsymbol{B}'\boldsymbol{B})^{-1} \boldsymbol{B}' \boldsymbol{x}_i$. Sendo $\boldsymbol{B}$ uma matriz com colunas ortonormais, sabe-se que $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}$. Dessa forma:

$$\max \operatorname{tr}\left[\boldsymbol{X}\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{X}'\right] = \max \operatorname{tr}\left[\boldsymbol{X}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{X}'\right] = \max \operatorname{tr}\left[\left(\boldsymbol{X}\boldsymbol{B}\right)\boldsymbol{B}'\boldsymbol{X}'\right] = \max \operatorname{tr}\left[\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}\right]. \tag{2.154}$$

Por suposição, a matriz $\boldsymbol{B}_{p\times q}$ possui colunas ortonormais. Todavia, para utilizar uma propriedade do traço vamos completar a matriz $\boldsymbol{B}$ com $p-q$ colunas para torná-la quadrada, de tal maneira que suas colunas continuem ortonormais. Diante disso, uma vez que $\boldsymbol{X}'\boldsymbol{X}$ é matriz $(p \times p)$ e $\boldsymbol{B}$ agora é uma matriz ortogonal $(p \times p)$, segue de Rencher e Shaalje (2008, p. 45) que:

$$\begin{aligned} \max \operatorname{tr}\left[\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}\right] &= \max \operatorname{tr}\left[\boldsymbol{B}'\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{B}\right] \\ &= \max \operatorname{tr}\left[\boldsymbol{X}'\boldsymbol{X}\right] \\ &= \max \operatorname{tr}\left[\hat{\boldsymbol{\Sigma}}\right], \end{aligned} \tag{2.155}$$

ou seja, a soma máxima dos autovalores de $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}$.

Pela propriedade 2.7.1, a maximização ocorre quando $\boldsymbol{B} = \boldsymbol{A}_q$, em que $\boldsymbol{A}_q$ é a matriz formada pelas primeiras q primeiras colunas (autovetores) de $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}$.

■

2.7.7 Regressão em componentes principais

A regressão via componentes principais (PCR, do inglês "*Principal Components Regression*") foi introduzida por Kendall (1957) e Hotelling (1957). Esse método faz uso dos procedimentos empregados na análise de componentes principais visando contornar os problemas apresentados pela regressão linear múltipla.

A construção geométrica de componentes principais pode ser vista no contexto da regressão linear considerando-se o modelo $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, em que $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ e $\text{cov}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$. Nesse caso, tem-se basicamente três vetores aleatórios, sendo os vetores $\boldsymbol{y}$, $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ e o estimador de mínimos quadrados $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Como $\text{cov}[\boldsymbol{y}] = \sigma^2\boldsymbol{I}$, a variabilidade do vetor $\boldsymbol{y}$ é constante. Nessa situação, não existem direções que forneçam maior variabilidade para a obtenção de componentes principais, ou seja, a projeção de $\boldsymbol{y}$ em qualquer direção possui a mesma variância σ^2. Esse fato pode ser visto de forma intuitiva, uma vez que a nuvem de pontos de $\boldsymbol{y}$ é esférica e a projeção ortogonal no subespaço também será esférica e de mesmo raio. Portanto, qualquer direção na $\text{Im}(\boldsymbol{X})$ definirá uma variável com variância σ^2.

Em termos de direções de máxima variabilidade, o mesmo ocorre para o vetor $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$. De fato, dado que $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, a matriz de covariâncias vetor $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ é:

$$\begin{aligned}
\text{cov}\left[P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right] &= \text{cov}\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= \left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\text{cov}[\boldsymbol{y}]\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]' \\
&= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \sigma^2\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \sigma^2\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'. \qquad (2.156)
\end{aligned}$$

Como $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é uma matriz de projeção $(n \times n)$, segue da definição 2.2.1 que essa matriz é idempotente, ou seja, $\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]^2 = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Disso resulta que 0 e 1 são os únicos autovalores de $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. De fato, suponha que λ seja um autovalor de $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Então, existe um autovetor não nulo $\boldsymbol{v}$ tal que:

$$\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\boldsymbol{v} = \lambda\boldsymbol{v}. \qquad (2.157)$$

Pré-multiplicando os dois membros na igualdade em (2.157) por $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ vem que:

$$\begin{aligned}
\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]^2\boldsymbol{v} &= \lambda\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\boldsymbol{v} \\
&= \lambda\lambda\boldsymbol{v} \\
&= \lambda^2\boldsymbol{v}. \qquad (2.158)
\end{aligned}$$

Do fato de $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ ser uma matriz idempotente segue que:

$$\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]^2\boldsymbol{v}=\lambda^2\boldsymbol{v}\Rightarrow\left[\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\boldsymbol{v}=\lambda^2\boldsymbol{v}$$
$$\Rightarrow\lambda\boldsymbol{v}=\lambda^2\boldsymbol{v}$$
$$\Rightarrow\left(\lambda^2-\lambda\right)\boldsymbol{v}=\boldsymbol{0}. \tag{2.159}$$

Uma vez que $\boldsymbol{v}$ é um vetor não nulo, segue de (2.159) que $\lambda^2-\lambda=0$, ou seja, $\lambda\left(\lambda-1\right)=0$. Portanto, $\lambda=0$ ou $\lambda=1$ são os únicos autovalores de $\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, como queríamos mostrar. O autoespaço relativo ao autovalor $\lambda=1$, ou seja, o subespaço relativo a esse autovalor é a própria $\mathrm{Im}\left(\boldsymbol{X}\right)$. Nesse caso, também não há que se considerar a obtenção dos componentes principais, pois toda a variabilidade conhecida está no subespaço $\mathrm{Im}\left(\boldsymbol{X}\right)$.

Finalmente, tem-se o vetor aleatório $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$. De (2.32) sabe-se para esse vetor que $\mathrm{cov}\left[\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right]=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Neste caso, tem-se que a variabilidade não é constante e depende da matriz $\boldsymbol{X}'\boldsymbol{X}$. Logo, pode-se construir os componentes principais do vetor $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}=\left[\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_p\right]$. Se para um vetor aleatório $\boldsymbol{X}$ os componentes são combinações lineares das variáveis aleatórias $(X_1,X_2,\ldots,X_p)$, agora a construção pode ser feita para a obtenção das combinações lineares de $\left(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_p\right)$.

Como a matriz de covariâncias de $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ depende de $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, o primeiro passo é determinar os autovalores dessa matriz. Seja $\boldsymbol{v}_i$ um autovetor unitário de $\boldsymbol{X}'\boldsymbol{X}$, relativo a um autovalor λ_i. Como $\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{v}_i=\lambda_i\boldsymbol{v}_i$, tem-se então que:

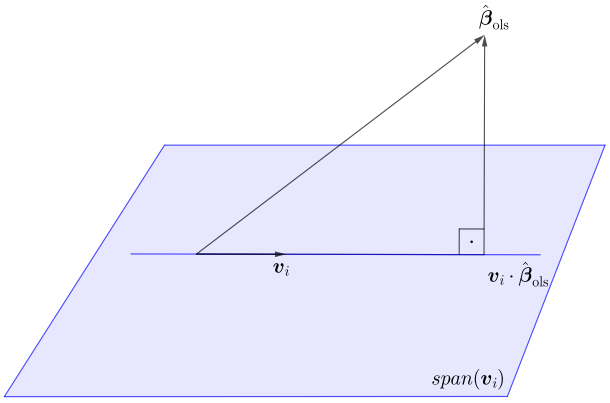
$$\lambda_i\boldsymbol{v}_i=\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{v}_i\Rightarrow\lambda_i\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v}_i=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{v}_i$$
$$\Rightarrow\lambda_i\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v}_i=\boldsymbol{v}_i$$
$$\Rightarrow\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v}_i=\frac{1}{\lambda_i}\boldsymbol{v}_i. \tag{2.160}$$

Logo, os autovetores são preservados, enquanto os autovalores correspondentes de $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ são inversos em relação aos autovalores de $\boldsymbol{X}'\boldsymbol{X}$. Com esses resultados pode-se determinar a variância da combinação linear $\boldsymbol{v}_i'\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, obtida pela projeção do vetor $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ na direção do subespaço gerado pelo vetor unitário $\boldsymbol{v}_i'$ (FIGURA 2.33):

$$\mathrm{var}\left[\boldsymbol{v}_i'\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right]=\boldsymbol{v}_i'\mathrm{cov}\left[\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right]\boldsymbol{v}_i=\sigma^2\boldsymbol{v}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v}_i=\sigma^2\frac{1}{\lambda_i}\boldsymbol{v}_i'\boldsymbol{v}_i=\sigma^2\frac{1}{\lambda_i}. \tag{2.161}$$

Observe que a variância de $\boldsymbol{v}_i'\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ existe, mas depende do parâmetro σ^2 que é desconhecido. Nesse caso, torna-se necessária a estimação do parâmetro σ^2 para garantir que a variância dessa combinação linear possa ser determinada. A filosofia para a obtenção dos componentes

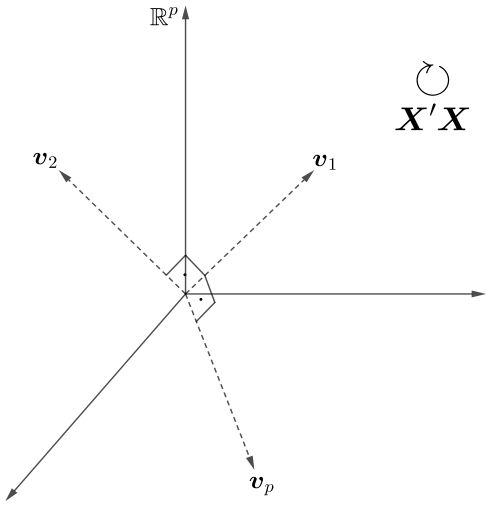
Figura 2.33 – Projeção do vetor $\hat{\boldsymbol{\beta}}_{\text{ols}}$ no subespaço gerado pelo vetor unitário $\boldsymbol{v}_i$.



Fonte: Do autor (2020).

principais pode mais uma vez ser utilizada, pois os vetores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ presentes em $\mathbb{R}^p$ e associados aos autovalores $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ indicam as direções de máxima variabilidade.

Figura 2.34 – Representação geométrica dos autovetores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ da matriz $\boldsymbol{X}'\boldsymbol{X}$.



Fonte: Adaptado de Silveira (2014).

Nesse sentido, o problema também pode ser tratado no espaço de observações $\mathbb{R}^n$, pela combinação linear de $\boldsymbol{X}\boldsymbol{v}_i$ com $\boldsymbol{y}$. A variância de $\boldsymbol{X}\boldsymbol{v}_i \cdot \boldsymbol{y}$ é então dada por:

$$\begin{aligned}
\text{var}\left[\boldsymbol{X}\boldsymbol{v}_i \cdot \boldsymbol{y}\right] &= \left(\boldsymbol{X}\boldsymbol{v}_i\right)' \text{cov}\left[\boldsymbol{y}\right] \left(\boldsymbol{X}\boldsymbol{v}_i\right) \\
&= \left(\boldsymbol{X}\boldsymbol{v}_i\right)' \sigma^2 \boldsymbol{I} \left(\boldsymbol{X}\boldsymbol{v}_i\right) \\
&= \sigma^2 \left(\boldsymbol{X}\boldsymbol{v}_i\right)' \left(\boldsymbol{X}\boldsymbol{v}_i\right) \\
&= \sigma^2 \boldsymbol{v}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right) \boldsymbol{v}_i \\
&= \sigma^2 \lambda_i \boldsymbol{v}_i' \boldsymbol{v}_i \\
&= \sigma^2 \lambda_i. \qquad (2.162)
\end{aligned}$$

A variância de $\boldsymbol{X}\boldsymbol{v}_i \cdot \boldsymbol{y}$ é maximizada em relação a $\|\boldsymbol{v}_i\| = 1$ quando $\boldsymbol{v}_i$ é um autovetor relativo ao maior autovalor de $\boldsymbol{X}'\boldsymbol{X}$. Para o vetor aleatório $\hat{\boldsymbol{\beta}}_{\text{ols}}$, o componente principal é $\boldsymbol{v}_i \cdot \hat{\boldsymbol{\beta}}_{\text{ols}}$, isto é, a projeção de $\hat{\boldsymbol{\beta}}_{\text{ols}}$ no subespaço gerado por $\boldsymbol{v}_i$.

Os vetores $(\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p)$ são ortogonais, pois são os autovetores de $\boldsymbol{X}\boldsymbol{X}'$. De fato,

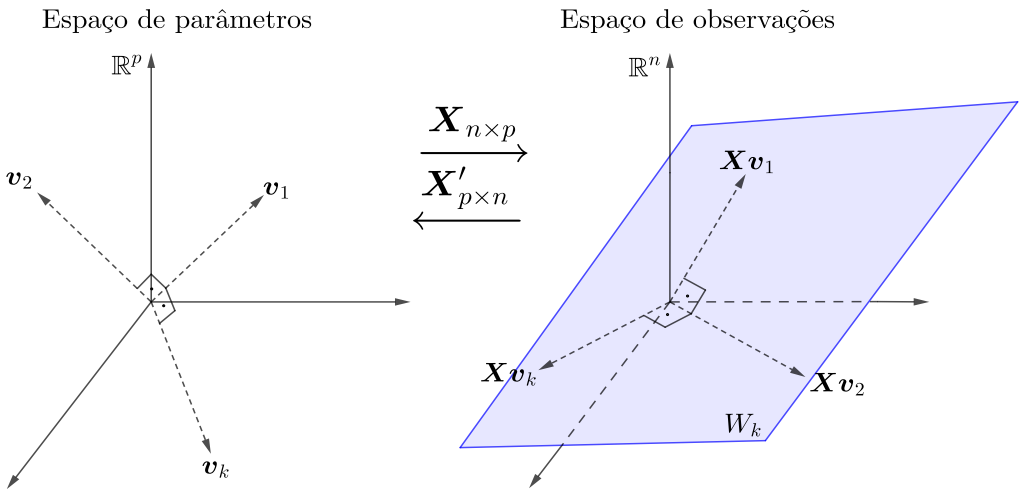
$$\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{X}\boldsymbol{v}_i = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{v}_i = \boldsymbol{X}\lambda_i\boldsymbol{v}_i = \lambda_i\boldsymbol{X}\boldsymbol{v}_i. \tag{2.163}$$

A regressão em componentes principais pode então ser dividida em duas etapas. A primeira consiste na obtenção dos componentes principais usuais $(\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p)$, que originam-se das direções $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ que fornecem a maior variabilidade em termos de $(X_1, X_2, \ldots, X_p)$. A segunda etapa consiste na projeção ortogonal do vetor aleatório $\boldsymbol{y}$ nos subespaços gerados por $(\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p)$ no espaço de observações. Isso equivale a projetar ortogonalmente o vetor $\hat{\boldsymbol{\beta}}_{\text{ols}}$ nos subespaços gerados pelos autovetores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ de $\boldsymbol{X}'\boldsymbol{X}$, no espaço de parâmetros. Dado que os vetores $(\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p)$ pertencem a $\mathbb{R}^n$, o conjunto de todas as combinações lineares desses vetores é referido como *span* de $(\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p)$. A ideia no método PCR consiste em se obter os subespaços $W_1, W_2, \ldots, W_k$ utilizando-se as direções $\boldsymbol{X}\boldsymbol{v}_i$, da seguinte maneira:

$$\begin{aligned}
W_1 &= span\{\boldsymbol{X}\boldsymbol{v}_1\} \\
W_2 &= span\{\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2\} \\
W_3 &= span\{\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \boldsymbol{X}\boldsymbol{v}_3\} \\
&\vdots \\
W_k &= span\{\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \boldsymbol{X}\boldsymbol{v}_3, \ldots, \boldsymbol{X}\boldsymbol{v}_k\}
\end{aligned} \tag{2.164}$$

Percebe-se com isso que a própria imagem de $\boldsymbol{X}$ pode ser reconstruída pela utilização de todos os componentes principais, ou seja, $\text{Im}(\boldsymbol{X}) = span\left\{\boldsymbol{X}\boldsymbol{v}_1, \boldsymbol{X}\boldsymbol{v}_2, \ldots, \boldsymbol{X}\boldsymbol{v}_p\right\}$. Na Figura 2.35 é apresentada uma representação geométrica do subespaço W_k. Nessa figura é possível observar a ação de duas transformações lineares, sendo que a matriz $\boldsymbol{X}_{n\times p}$ atua como uma transformação do espaço de parâmetros para o espaço de observações, enquanto a matriz $\boldsymbol{X}'_{p\times n}$ atua como uma transformação do espaço de observações para o espaço de parâmetros.

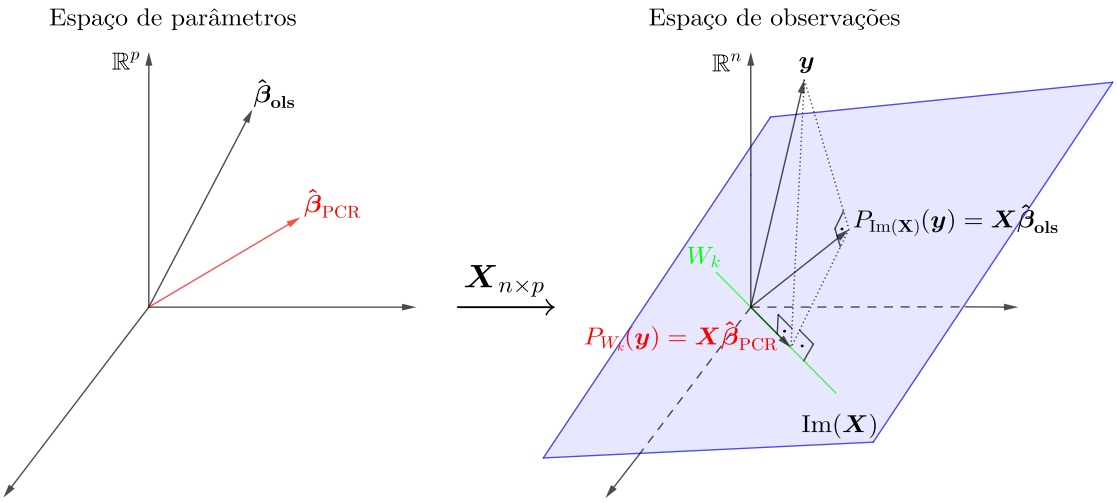
Figura 2.35 – Subespaço W_k gerado pelos k primeiros autovetores da matriz $\boldsymbol{X}'\boldsymbol{X}$.



Fonte: Adaptado de Silveira (2014).

Como as direções $\boldsymbol{X}\boldsymbol{v}_i$ explicam a variabilidade das covariáveis $(X_1, X_2, \ldots, X_p)$ e os k primeiros componentes principais retêm a maior parte da variabilidade das variáveis originais, a ideia é regredir $\boldsymbol{y}$ aos k-primeiros componentes principais, isto é, projetar $\boldsymbol{y}$ ortogonalmente no subespaço W_k. O estimador da regressão em componentes principais é denotado por $\hat{\boldsymbol{\beta}}_{\text{PCR}}$ e a projeção de $\boldsymbol{y}$ no subespaço W_k no espaço de observações possui equações normais dadas por $P_{W_k}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{PCR}}$. Geometricamente, essa projeção está representada na Figura 2.36.

Figura 2.36 – Projeção ortogonal de $\boldsymbol{y}$ no subespaço W_k para a obtenção do estimador $\hat{\boldsymbol{\beta}}_{\text{PCR}}$.



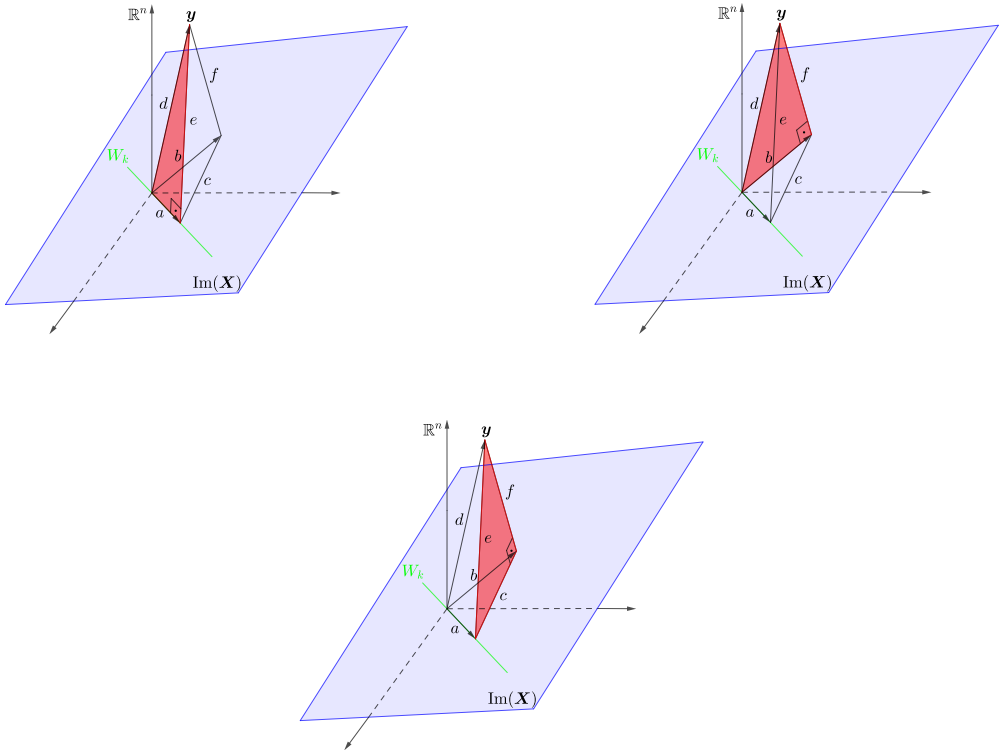
Fonte: Adaptado de Silveira (2014).

A projeção ortogonal de $\boldsymbol{y}$ no subespaço W_k e na $\text{Im}(\boldsymbol{X})$, equivale a projetar ortogonalmente $P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})$ em W_k (FIGURA 2.36). A demonstração desse fato, a projeção ortogonal de

$P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$ em W_k, segue da aplicação do teorema de Pitágoras nos triângulos retângulos ade, bdf e cef apresentados na Figura 2.37, em que para uma melhor visualização considerou-se $\boldsymbol{a} = P_{W_k}(\boldsymbol{y})$ e $\boldsymbol{b} = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$. O objetivo passa a ser então demonstrar que os vetores $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$, $P_{W_k}(\boldsymbol{y})$ e o vetor dado pela diferença entre eles formam um triângulo retângulo.

De fato, como ade, bdf e cef são triângulos retângulos e com base na Figura 2.37, segue do teorema de Pitágoras que $d^2 = a^2 + e^2$, $d^2 = b^2 + f^2$ e $e^2 = c^2 + f^2$, respectivamente. Uma vez que $d^2 = b^2 + f^2$ e $d^2 = a^2 + e^2$ então $b^2 + f^2 = a^2 + e^2$. Desse modo, $b^2 + f^2 = a^2 + c^2 + f^2$ se e somente se $b^2 = a^2 + c^2$. Portanto, abc é um triângulo retângulo, em que $\boldsymbol{a} = P_{W_k}(\boldsymbol{y})$, $\boldsymbol{b} = P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$ e $\boldsymbol{c}$ é o vetor obtido pela diferença entre $P_{W_k}(\boldsymbol{y})$ e $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$. Consequentemente, a projeção de $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$ em W_k é ortogonal.

Figura 2.37 – Triângulos retângulos ade, bdf e cef obtidos pela projeção ortogonal de $\boldsymbol{y}$ no subespaço W_k e na imagem de $\boldsymbol{X}$.



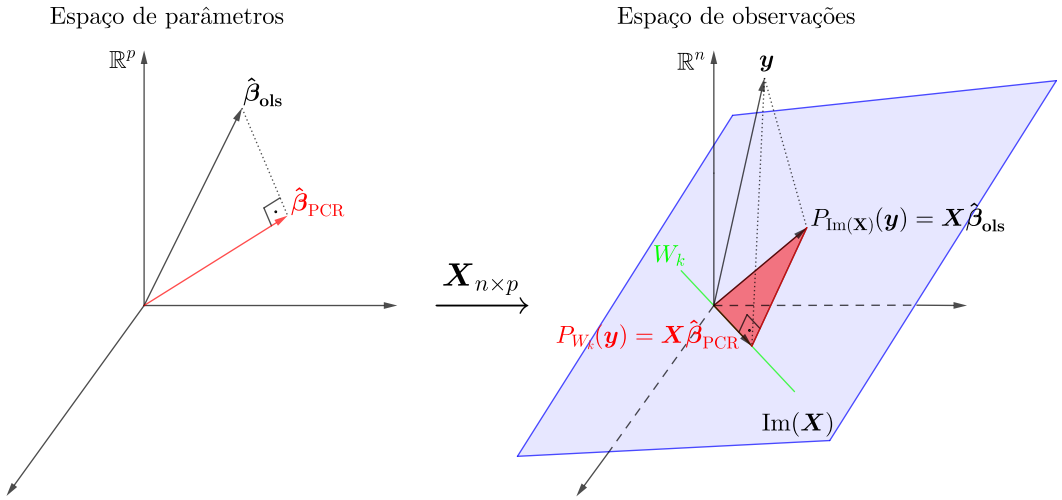
Fonte: Do autor (2020).

Desse resultado, segue para o triângulo retângulo formado pelos vetores $P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})$, $P_{W_k}(\boldsymbol{y})$ e pelo vetor dado pela diferença entre eles que:

$$\left\|P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y})\right\|^2 = \left\|P_{W_k}(\boldsymbol{y})\right\|^2 + \left\|P_{\mathrm{Im}(\boldsymbol{X})}(\boldsymbol{y}) - P_{W_k}(\boldsymbol{y})\right\|^2. \tag{2.165}$$

Desse modo, por meio dessa relação vem que $\|P_{W_k}(\boldsymbol{y})\| \leq \left\|P_{\text{Im}(\boldsymbol{X})}(\boldsymbol{y})\right\|$. Apesar desse fato não implicar em $\left\|\hat{\boldsymbol{\beta}}_{\text{PCR}}\right\| < \left\|\hat{\boldsymbol{\beta}}_{\text{ols}}\right\|$, o procedimento da regressão em componentes principais é claramente conservador, no sentido que fornece estimativas menores para o vetor de parâmetros $\boldsymbol{\beta}$. Contudo, pode ser demonstrado que $\left\|\hat{\boldsymbol{\beta}}_{\text{PCR}}\right\| < \left\|\hat{\boldsymbol{\beta}}_{\text{ols}}\right\|$. Essa demonstração não será feita aqui, mas pode ser vista em Silveira (2014). As expressões matriciais para as equações normais do estimador $\hat{\boldsymbol{\beta}}_{\text{PCR}}$ são obtidas com base em uma nova transformação, ou seja, o projetor $P_{W_k}(\boldsymbol{y})$ é expresso matricialmente baseando-se em uma nova transformação entre os espaços paramétrico e de observações. Ao final mostra-se que $\hat{\boldsymbol{\beta}}_{\text{PCR}}$ é obtido como projeção ortogonal do vetor $\hat{\boldsymbol{\beta}}_{\text{ols}}$, o que garante que $\left\|\hat{\boldsymbol{\beta}}_{\text{PCR}}\right\| < \left\|\hat{\boldsymbol{\beta}}_{\text{ols}}\right\|$. Nesse sentido, $\hat{\boldsymbol{\beta}}_{\text{PCR}}$ é um estimador de encolhimento em relação ao estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

Figura 2.38 – Projeção ortogonal do vetor $\hat{\boldsymbol{\beta}}_{\text{ols}}$ no vetor $\hat{\boldsymbol{\beta}}_{\text{PCR}}$ no espaço de parâmetros e de $\boldsymbol{y}$ no subespaço W_k e na e na imagem de $\boldsymbol{X}$ no espaço de observações.



Fonte: Do autor (2020).

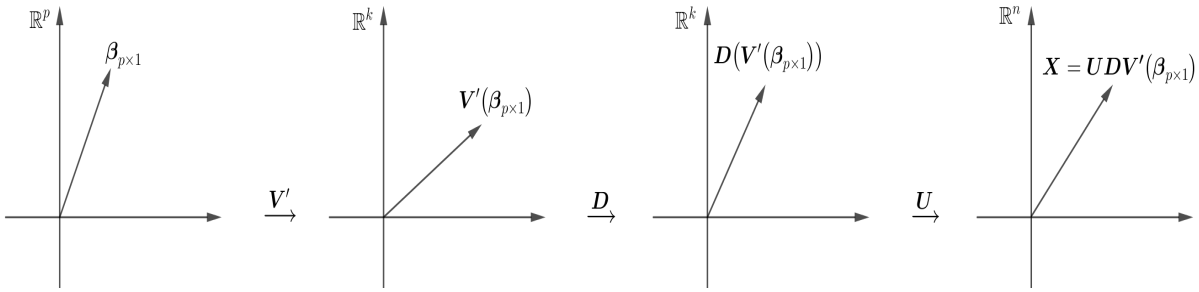
Diante dos objetivos mais importantes ao utilizar-se a PCA, apresentados na seção 2.7.2, é nítido que o método PCR consegue atender os dois primeiros objetivos. Todavia, para um grande número de variáveis ainda persiste o problema de interpretação dos componentes principais. Como também enfatizado na seção 2.7.2, um dos avanços para o problema de interpretação dos componentes principais foi o trabalho de Zou, Hastie e Tibshirani (2006). No artigo, os autores relacionam os métodos de regressão *Ridge*, LASSO e Elastic Net a PCA. Zou, Hastie e Tibshirani (2006) propuseram a análise de componentes principais esparsos (SPCA, do inglês "Sparse Principal Component Analysis") com o objetivo de aproximar as propriedades da PCA padrão, mantendo um pequeno número de coeficientes ou *loadings* diferentes de zero. Nesse sentido, a SPCA torna-se o tema central das próximas seções desse trabalho.

2.8 Relação entre os componentes principais e o estimador *Ridge*

A SPCA é uma abordagem para se obterem componentes modificados com *loadings* esparsos e baseia-se na capacidade de se escrever a PCA como um problema de otimização do tipo regressão. Inicialmente, Zou, Hastie e Tibshirani (2006) mostraram que os componentes podem ser obtidos utilizando-se o método *Ridge*. Como extensão, os autores incluíram a restrição do método LASSO (*Elastic Net*) ao problema, integrando dessa forma as boas propriedades desse método, de modo que a PCA resultante produz *loadings* esparsos (*loadings* nulos). Primeiramente, será destacada a relação entre os componentes principais e o estimador *Ridge*.

Como já visto, os componentes principais podem ser determinados pela decomposição da matriz de delineamento em valores singulares $\boldsymbol{X}_{n\times p}=\boldsymbol{U}_{n\times k}\boldsymbol{D}_{k\times k}\boldsymbol{V}'_{k\times p}$, em que as colunas $\boldsymbol{z}_{(i)}$ da matriz $\boldsymbol{Z}_{n\times k}=\boldsymbol{U}_{n\times k}\boldsymbol{D}_{k\times k}$ são os componentes principais, as linhas $\boldsymbol{v}'_i$ da matriz $\boldsymbol{V}'_{k\times p}$ são os vetores de *loadings* correspondentes e a matriz $\boldsymbol{D}$ é uma matriz diagonal com entradas λ_i, $i=1,2,\ldots,k$. Os vetores coluna da matriz $\boldsymbol{U}$ são ortonormais ($\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}_{k\times k}$), o mesmo ocorrendo para os vetores linha de $\boldsymbol{V}'$ ($\boldsymbol{V}'\boldsymbol{V}=\boldsymbol{I}_{k\times k}$). Na Figura 2.39 apresenta-se uma representação gráfica da decomposição em valores singulares da matriz $\boldsymbol{X}$.

Figura 2.39 – Decomposição em valores singulares da matriz $\boldsymbol{X}$.



Fonte: Do autor (2020).

A partir das definições das matrizes $\boldsymbol{U}$ e $\boldsymbol{V}'$, note que essas matrizes podem ser expressas da seguinte forma:

$$\boldsymbol{U}=\left[\boldsymbol{u}_{(1)},\boldsymbol{u}_{(2)},\ldots,\boldsymbol{u}_{(k)}\right] \text{ e } \boldsymbol{V}'=\begin{bmatrix}\boldsymbol{v}'_1\\ \boldsymbol{v}'_2\\ \vdots\\ \boldsymbol{v}'_k\end{bmatrix}. \tag{2.166}$$

Segue do fato de $\boldsymbol{U}$ possuir colunas ortonormais que $\boldsymbol{u}'_j\boldsymbol{u}_j=1$ e $\boldsymbol{u}'_j\boldsymbol{u}_i=0$, para $i,j=1,2,\ldots,k$ $(i\neq j)$. Por sua vez, segue do fato de $\boldsymbol{V}'$ possuir linhas ortonormais que $\boldsymbol{v}'_i\boldsymbol{v}_i=1$ e $\boldsymbol{v}'_i\boldsymbol{v}_j=0$, para $i,j=1,2,\ldots,k$ $(i\neq j)$. Portanto,

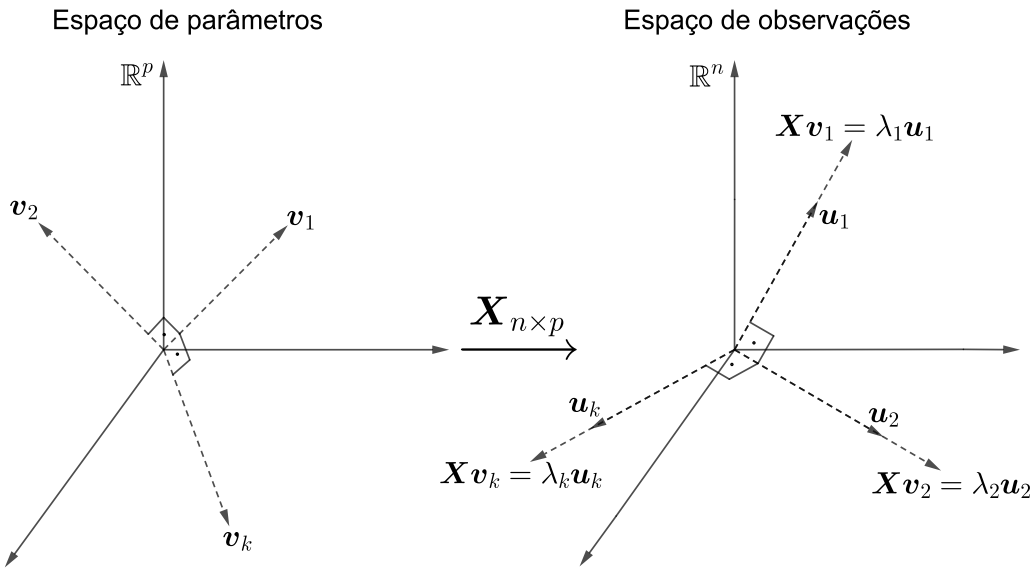
$$
\begin{aligned}
\boldsymbol{X}\boldsymbol{v}_i &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'\boldsymbol{v}_i \\
&= \left[\boldsymbol{u}_{(1)},\ldots,\boldsymbol{u}_{(i)},\ldots,\boldsymbol{u}_{(k)}\right]
\begin{bmatrix}
\lambda_1 & \ldots & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
0 & \ldots & \lambda_i & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & \ldots & \lambda_k
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{v}_1' \\ \vdots \\ \boldsymbol{v}_i' \\ \vdots \\ \boldsymbol{v}_k'
\end{bmatrix}
\boldsymbol{v}_i \\
&= \left[\boldsymbol{u}_{(1)},\ldots,\boldsymbol{u}_{(i)},\ldots,\boldsymbol{u}_{(k)}\right]
\begin{bmatrix}
\lambda_1 & \ldots & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
0 & \ldots & \lambda_i & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & \ldots & \lambda_k
\end{bmatrix}
\begin{bmatrix}
0 \\ \vdots \\ 1 \\ \vdots \\ 0
\end{bmatrix} \\
&= \left[\boldsymbol{u}_{(1)},\ldots,\boldsymbol{u}_{(i)},\ldots,\boldsymbol{u}_{(k)}\right](\lambda_i\boldsymbol{e}_i) \\
&= \lambda_i\boldsymbol{u}_i, \qquad (2.167)
\end{aligned}
$$

em que $\boldsymbol{e}_i' = (0,\ldots,1,\ldots,0)$ é o vetor canônico que possui argumento unitário na posição i e argumentos nulos nas demais posições.

Novamente, os vetores $(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k)$ correspondem às direções que capturam sequencialmente e de forma decrescente a variância dos dados, ou equivalentemente, são os autovetores da matriz de covariâncias da amostra $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$, com $\boldsymbol{X}$ em sua forma centrada. Portanto, $(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k)$ são estimadores dos *loadings* populacionais, os autovetores da matriz de covariâncias populacionais $\boldsymbol{\Sigma}$. Os vetores $(\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_k)$ são projeções dos dados nas direções dos *loadings* e os vetores $(\lambda_1\boldsymbol{u}_1,\lambda_2\boldsymbol{u}_2,\ldots,\lambda_k\boldsymbol{u}_k)$ são os componentes principais (FIGURA 2.40).

Por ser largamente utilizada tanto na Estatística como em outras áreas, ao longo das décadas ocorreu uma proliferação da terminologia associada a PCA. Wilks (2011, p. 471-472) fornece em seu livro um breve resumo da terminologia variada da PCA. Segundo o autor, o nome mais comum para os elementos individuais dos autovetores na literatura estatística é *loading*. Outro termo muito usual para esse fim na literatura estatística é "coeficiente". Jolliffe (2002, p. 6) ressalta que algumas vezes os vetores $\boldsymbol{v}_i$ $(i = 1,2,\ldots,p)$ são referidos como "componentes principais". Segundo o autor, é preferível reservar o termo "componentes principais" para as variáveis derivadas $\boldsymbol{X}\boldsymbol{v}_i = \lambda_i\boldsymbol{u}_i$ e se referir a $\boldsymbol{v}_i$ como o vetor de coeficientes ou de "*loa-*

Figura 2.40 – Vetores de *loadings* no espaço paramétrico e de componentes principais no espaço de observações.



Fonte: Do autor (2020).

dings" para o *i*-ésimo componente. Neste trabalho, para cada um dos vetores $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ será utilizado a terminologia "vetor de *loadings*", em concordância com Cadima e Jolliffe (1995), Jolliffe (2002), Zou, Hastie e Tibshirani (2006), entre outros trabalhos na literatura.

Como observado por Zou, Hastie e Tibshirani (2006), pode-se considerar a regressão *Ridge* para cada componente principal. Utilizando o vetor de pseudodados $\boldsymbol{z}_i = \lambda_i \boldsymbol{u}_i$, o estimador *Ridge* para essa situação é definido como:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \underset{\boldsymbol{\beta}}{\arg\min} \left(\|\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma \|\boldsymbol{\beta}\|_2^2 \right). \tag{2.168}$$

Como $\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{z}_i$ e $\boldsymbol{X}_{n\times p} = \boldsymbol{U}_{n\times p}\boldsymbol{D}_{p\times p}\boldsymbol{V}'_{p\times p}$ (posto coluna completo), a regressão *Ridge* para o *i*-ésimo componente principal é dada por:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{Ridge}} &= (\boldsymbol{X}'\boldsymbol{X} + \tau\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{z}_i \\
&= (\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' + \tau\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i \\
&= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}' + \tau\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i \\
&= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}' + \tau\boldsymbol{V}\boldsymbol{V}')^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \tau\boldsymbol{I})^{-1}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \tau\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i,
\end{aligned} \tag{2.169}$$

em que $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$, pois $\boldsymbol{V}$ é uma matriz quadrada $(p \times p)$ com colunas ortonormais e a matriz inversa de $(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}' + \tau\boldsymbol{V}\boldsymbol{V}')$ é $\boldsymbol{V}(\boldsymbol{D}^2 + \tau\boldsymbol{I})^{-1}\boldsymbol{V}'$ (APÊNDICE).

Para $\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1}$ vem que:

$$
\begin{aligned}
\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1} &= \left(\begin{bmatrix} \lambda_1^2 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \lambda_i^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \lambda_p^2 \end{bmatrix} + \tau \begin{bmatrix} 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} \lambda_1^2 + \tau & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \lambda_i^2 + \tau & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \lambda_p^2 + \tau \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \frac{1}{\lambda_1^2+\tau} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \frac{1}{\lambda_i^2+\tau} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \frac{1}{\lambda_p^2+\tau} \end{bmatrix}. \qquad (2.170)
\end{aligned}
$$

De posse desse último resultado segue para $\boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1}$ que:

$$
\begin{aligned}
\boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1} &= \begin{bmatrix} \boldsymbol{v}_{(1)} & \dots & \boldsymbol{v}_{(i)} & \dots & \boldsymbol{v}_{(p)} \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_1^2+\tau} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \frac{1}{\lambda_i^2+\tau} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \frac{1}{\lambda_p^2+\tau} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\lambda_1^2+\tau}\boldsymbol{v}_{(1)} & \dots & \frac{1}{\lambda_i^2+\tau}\boldsymbol{v}_{(i)} & \dots & \frac{1}{\lambda_p^2+\tau}\boldsymbol{v}_{(p)} \end{bmatrix}. \qquad (2.171)
\end{aligned}
$$

Uma vez que $\boldsymbol{u}_i'\boldsymbol{u}_i = 1$ e $\boldsymbol{u}_i'\boldsymbol{u}_j = 0$, decorre para $\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i$ que:

$$
\begin{aligned}
\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i &= \begin{bmatrix} \lambda_1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \lambda_i & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \lambda_p \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_i' \\ \vdots \\ \boldsymbol{u}_p' \end{bmatrix} \boldsymbol{u}_i\lambda_i \\
&= \begin{bmatrix} \lambda_1\boldsymbol{u}_1' \\ \vdots \\ \lambda_i\boldsymbol{u}_i' \\ \vdots \\ \lambda_p\boldsymbol{u}_p' \end{bmatrix} \boldsymbol{u}_i\lambda_i \\
&= \begin{bmatrix} 0 \\ \vdots \\ \lambda_i^2 \\ \vdots \\ 0 \end{bmatrix}. \qquad (2.172)
\end{aligned}
$$

Logo, substituindo (2.171) e (2.172) em (2.169) resulta que:

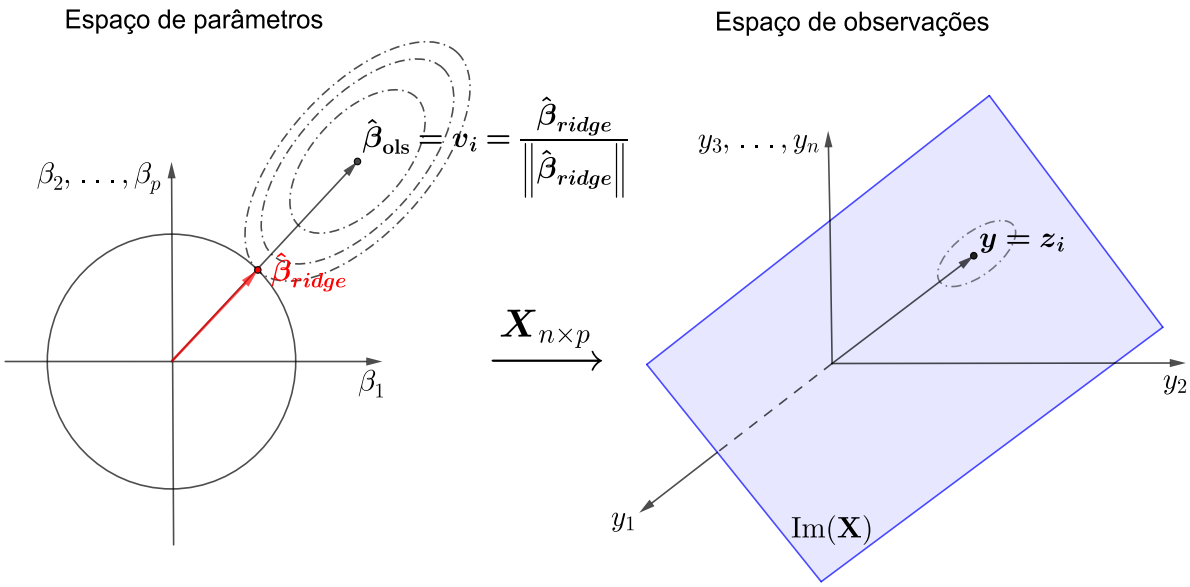
$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{Ridge}} &= \boldsymbol{V}\left(\boldsymbol{D}^2+\tau\boldsymbol{I}\right)^{-1}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{u}_i\lambda_i \\
&= \begin{bmatrix} \frac{1}{\lambda_1^2+\tau}\boldsymbol{v}_{(1)} & \cdots & \frac{1}{\lambda_i^2+\tau}\boldsymbol{v}_{(i)} & \cdots & \frac{1}{\lambda_p^2+\tau}\boldsymbol{v}_{(p)} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ \lambda_i^2 \\ \vdots \\ 0 \end{bmatrix} \\
&= \frac{\lambda_i^2}{\lambda_i^2+\tau}\boldsymbol{v}_i, \qquad (2.173)
\end{aligned}
$$

em que $i = 1,2,\dots,p$.

Como $\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \frac{\lambda_i^2}{\lambda_i^2+\tau}\boldsymbol{v}_i$, tem-se que o estimador *Ridge* é proporcional ao i-ésimo vetor de *loadings* $\boldsymbol{v}_i$. A demonstração algébrica acima não é muito esclarecedora em relação ao que de fato está ocorrendo e, neste caso, uma demonstração geométrica é interessante. Como

$\boldsymbol{X}\boldsymbol{v}_i = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'\boldsymbol{v}_i = \lambda_i \boldsymbol{u}_i = \boldsymbol{z}_i$, o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é igual a $\boldsymbol{v}_i$. De fato, na seção 2.2 foi visto que devido a matriz $\boldsymbol{X}$ ser injetiva, existe um $\hat{\boldsymbol{\beta}}$ no espaço de parâmetros $\mathbb{R}^p$ que é levado por essa transformação linear de forma única na projeção ortogonal de $\boldsymbol{y}$ na $\text{Im}(\boldsymbol{X})$, que é um subespaço do $\mathbb{R}^n$. Esse $\hat{\boldsymbol{\beta}}$ é a solução de mínimos quadrados $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Uma vez que $\boldsymbol{X}\boldsymbol{v}_i = \boldsymbol{z}_i$, segue agora que é o vetor $\boldsymbol{v}_i$ que é levado de forma única pela matriz $\boldsymbol{X}$ em $\boldsymbol{y} = \boldsymbol{z}_i$ (FIGURA 2.41). Isso justifica a afirmação de que o estimador $\hat{\boldsymbol{\beta}}_{\text{ols}}$ é igual a $\boldsymbol{v}_i$. Considerando uma hiperesfera centrada em $\boldsymbol{z}_i$ na $\text{Im}(\boldsymbol{X})$, a pré-imagem dessa hiperesfera é um elipsoide centrado em $\boldsymbol{v}_i$ no espaço de parâmetros. Além disso, $\boldsymbol{v}_i$ é também um dos eixos principais desse elipsoide. Neste caso, é possível observar que o ponto de tangência entre a hiperesfera centrada na origem e o elipsoide é o vetor $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$, que é proporcional ao vetor $\boldsymbol{v}_i$, isto é, $\boldsymbol{v}_i = \dfrac{\hat{\boldsymbol{\beta}}_{\text{Ridge}}}{\left\|\hat{\boldsymbol{\beta}}_{\text{Ridge}}\right\|}$ (FIGURA 2.41).

Figura 2.41 – Descrição geométrica da obtenção dos componentes principais utilizando o estimador *Ridge*.



Fonte: Do autor (2020).

A mesma construção vale para os outros vetores singulares $\boldsymbol{v}_i$ e $\boldsymbol{u}_i$. Dessa forma, todos os componentes principais podem ser obtidos a partir da regressão *Ridge*.

2.8.1 Componentes principais com esparsidade

Para se obter a esparsidade nos componentes principais, a construção anterior pode ser estendida, modificando-se a regressão *Ridge* para uma regressão *Elastic Net* pelo acréscimo da restrição $\|\boldsymbol{\beta}\|_1$ (restrição LASSO) na penalização. Novamente, os pseudodados $\boldsymbol{z}_i = \lambda_i \boldsymbol{u}_i$ são

utilizados, considerando-se agora a regressão *Elastic Net*. Logo, para o estimador *Elastic Net* vem que:

$$\hat{\boldsymbol{\beta}}_{\text{en}} = (1+\gamma_2)\left\{\arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma_1\|\boldsymbol{\beta}\|_1 + \gamma_2\|\boldsymbol{\beta}\|_2^2\right)\right\}. \tag{2.174}$$

De modo análogo à construção anterior, o elipsoide está centrado em $\boldsymbol{v}_i$ com um de seus eixos na direção de $\boldsymbol{v}_i$, definindo uma reta que passa pela origem e por esse ponto. A estimativa $\hat{\boldsymbol{\beta}}_{\text{en}}$ é obtida pela interseção do elipsoide com o convexo $\gamma_1\|\boldsymbol{\beta}\|_1 + \gamma_2\|\boldsymbol{\beta}\|_2^2 \leq t$, para um determinado valor de t. Note então que essa interseção não ocorre mais na reta que liga o ponto $\boldsymbol{v}_i$ a origem, como ocorre na regressão *Ridge*, mas em um ponto próximo (FIGURA 2.42). Logo, tem-se a seguinte aproximação:

$$\frac{\hat{\boldsymbol{\beta}}_{\text{en}}}{\left\|\hat{\boldsymbol{\beta}}_{\text{en}}\right\|} \approx \boldsymbol{v}_i. \tag{2.175}$$

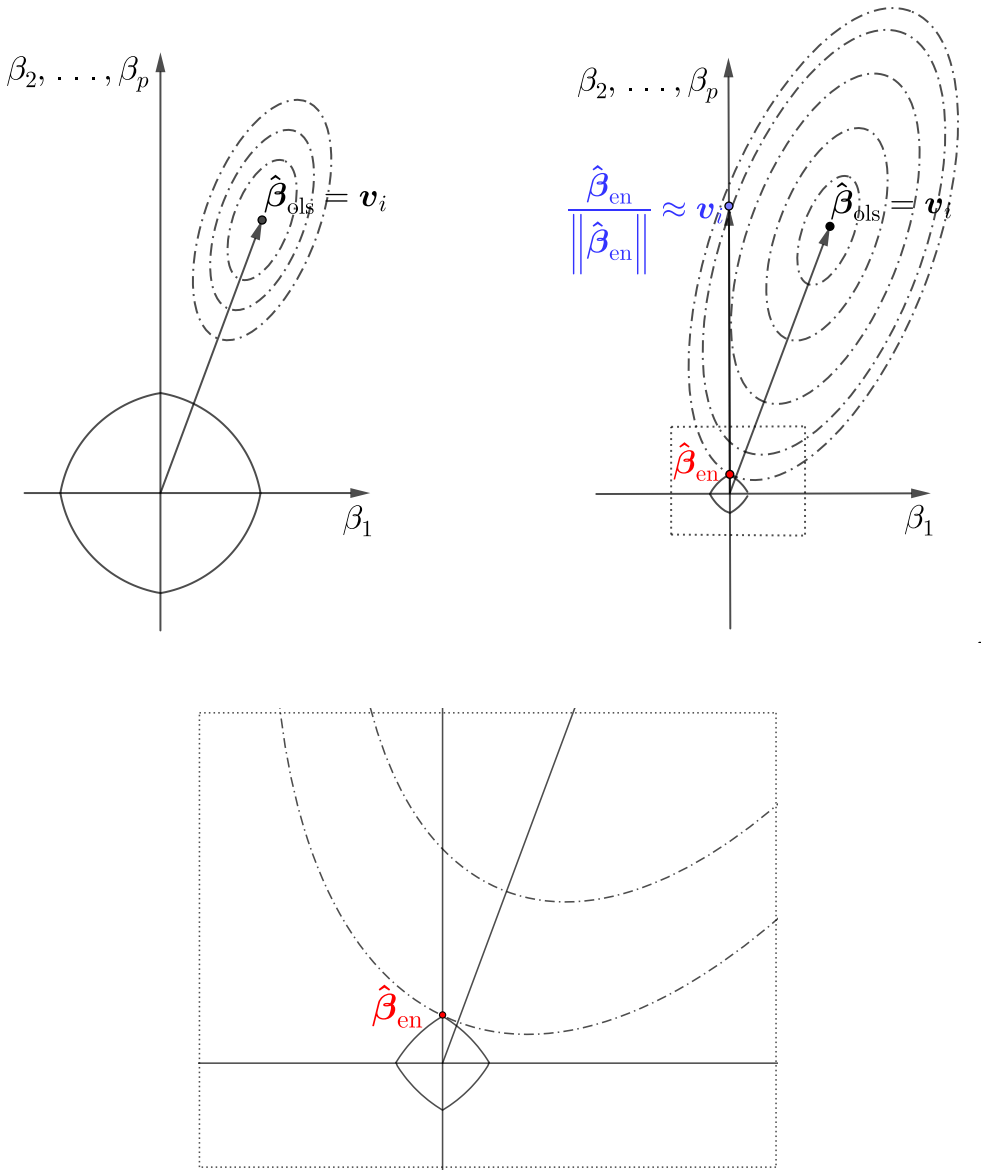
Variando-se o valor do parâmetro de ajuste t é possível obter-se estimativas $\hat{\boldsymbol{\beta}}_{\text{en}}$ com esparsidade, quando o elipsoide intercepta o convexo em algum eixo coordenado. Portanto, tem-se uma aproximação dos componentes principais com esparsidade, que serão denominados componentes principais esparsos. Esse método foi introduzido por Zou, Hastie e Tibshirani (2006) e foi denominado de "*Sparse Principal Component Analysis*". Na Figura 2.42 é apresentada a descrição geométrica da obtenção dos vetores de *loadings* utilizando o estimador *Elastic Net*.

Duas propriedades inerentes aos vetores de *loadings*, o fato de serem unitários e ortogonais, acarretam por consequência as duas propriedades mais importantes das variáveis latentes, os componentes principais. Como já observado na seção 2.7.5, a restrição da norma unitária aos vetores de *loadings* é essencial para tornar o problema bem definido, uma vez que a variância dos componentes principais pode ser arbitrariamente grande sem essa restrição. Por sua vez, a ortogonalidade dos vetores de *loadings* torna os componentes principais ortogonais (não correlacionados).

Se os componentes são modificados de alguma forma, como para componentes rotacionados para uma melhor interpretação (JOLLIFFE, 1995) ou pela formação de componentes esparsos (ZOU; HASTIE; TIBSHIRANI, 2006), uma dessas duas propriedades dos vetores de *loadings* precisará ser sacrificada, uma vez que os vetores de *loadings* modificados não são mais exatamente os autovetores da matriz de covariâncias. O que habitualmente ocorre é sacrificar

a ortogonalidade dos vetores de *loadings* e, nesse caso, os componentes principais passam a ser correlacionados. Dessa forma, as duas propriedades básicas dos componentes principais, a ortogonalidade e a independência, só valem de forma aproximada para os componentes modificados. Em razão disto, a variância explicada para os componentes principais esparsos deve ser estimada e torna-se necessário uma noção modificada de variância explicada, que é descrita na subseção seguinte.

Figura 2.42 – Descrição geométrica da obtenção dos vetores de *loadings* utilizando o estimador *Elastic Net*.



Fonte: Do autor (2020).

2.8.2 A variância total generalizada

Se $\hat{\boldsymbol{Z}} = \left[\hat{\boldsymbol{Z}}_{(1)}, \hat{\boldsymbol{Z}}_{(2)}, \ldots, \hat{\boldsymbol{Z}}_{(k)}\right]$ são os componentes principais aproximados, obtidos por algum método, certamente não se observará a não correlação entre os componentes. Dessa forma, a variância total não deve ser mais estimada pelo traço de $\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}$. A presença de correlação certamente acarretaria superestimação (ZOU; HASTIE; TIBSHIRANI, 2006). Como os componentes principais estão naturalmente ordenados a partir dos autovalores, a ideia para se calcular a variância explicada pelos k primeiros componentes modificados $\left\{\hat{\boldsymbol{Z}}_i,\ i = 1, \ldots, k\right\}$ é a seguinte construção: obtidos $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$, projeta-se ortogonalmente $\hat{\boldsymbol{Z}}_2$ em $\hat{\boldsymbol{Z}}_1$, tomando-se a diferença $\hat{\boldsymbol{Z}}_{2|1} = \hat{\boldsymbol{Z}}_2 - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_2\right)$. Essa é uma forma de se eliminar a dependência linear entre $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$. Obtido $\hat{\boldsymbol{Z}}_3$, projeta-se $\hat{\boldsymbol{Z}}_3$ ortogonalmente no subespaço gerado por $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$ e toma-se a diferença $\hat{\boldsymbol{Z}}_{3|1,2} = \hat{\boldsymbol{Z}}_3 - P_{\hat{\boldsymbol{Z}}_1|\hat{\boldsymbol{Z}}_2}\left(\hat{\boldsymbol{Z}}_3\right)$. Assim sucessivamente obtém-se $\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1}$:

$$\begin{aligned}
&\hat{\boldsymbol{Z}}_1 \\
&\hat{\boldsymbol{Z}}_{2|1} = \hat{\boldsymbol{Z}}_2 - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_2\right) \\
&\hat{\boldsymbol{Z}}_{3|1,2} = \hat{\boldsymbol{Z}}_3 - P_{\hat{\boldsymbol{Z}}_2}\left(\hat{\boldsymbol{Z}}_3\right) - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_3\right) \\
&\vdots \\
&\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1} = \hat{\boldsymbol{Z}}_k - P_{\hat{\boldsymbol{Z}}_{k-1}}\left(\hat{\boldsymbol{Z}}_k\right) - \ldots - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_k\right)
\end{aligned} \tag{2.176}$$

Segue então que a variância explicada por $\hat{\boldsymbol{Z}}_k$ é $\left\|\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1}\right\|^2$ e a variância total explicada até o k-ésimo componente principal é dada por $\sum_{j=1}^{k}\left\|\hat{\boldsymbol{Z}}_{j|1,\ldots,j-1}\right\|^2$.

2.9 O método *Octagonal Shrinkage and Clustering Algorithm for Regression* (OSCAR)

Após a ampla utilização do método de regressão de penalização LASSO, tanto como um método de encolhimento quanto para seleção de covariáveis em modelos de regressão linear, foram propostas outras construções semelhantes visando acrescentar propriedades e sanar deficiências desse método. Um dos procedimentos mais bem sucedidos talvez seja o método *Elastic Net*.

Essencialmente, todas as variantes que foram propostas partem da construção básica relacionada a um conjunto convexo e o estimador de mínimos quadrados $\hat{\boldsymbol{\beta}}_{\text{ols}}$, que é projetado nesse convexo via métrica de Mahalanobis (ANEXO A). Uma dessas construções, particularmente elegante, foi proposta por Bondell e Reich (2008), sendo denominada de algoritmo de encolhimento e de agrupamento octogonal para regressão (OSCAR, do inglês "*Octagonal Sh-*

rinkage and Clustering Algorithm for Regression"). De forma semelhante ao método *Elastic Net*, o método OSCAR também possui a capacidade de selecionar simultaneamente covariáveis. Além de possuir essa propriedade, esse método também impõe o agrupamento supervisionado de covariáveis altamente correlacionadas, que possuem um efeito semelhante sobre a resposta, formando assim *clusters* que são representados por um único valor para os coeficientes dos parâmetros. A palavra "*octogonal*" em OSCAR é motivada pela geometria da região convexa que define a penalidade.

Para a formulação do OSCAR, considere o modelo de regressão linear usual com dados observados em n observações e p covariáveis preditoras. Seja $\boldsymbol{y}' = (y_1, y_2, \ldots, y_n)$ o vetor de respostas e $\boldsymbol{x}'_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ a j-ésima covariável preditora, $j = 1, 2, \ldots, p$. Assuma que a resposta foi centralizada e que cada covariável está em sua forma padronizada. Desse modo:

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \ \text{ e } \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, 2, \ldots, p. \tag{2.177}$$

Como nas abordagens anteriores, o OSCAR é construído por meio de um problema de mínimos quadrados com restrição. A escolha da restrição utilizada nesse método consiste da combinação ponderada das normas L_1 e L_∞ par a par para os coeficientes. Especificamente, o problema de otimização de mínimos quadrados restrito para o OSCAR é dado por:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{OSCAR}} = \arg\min_{\boldsymbol{\beta}} & \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 \\ & \text{restrito a} \\ & \sum_{j=1}^{p} \left|\beta_j\right| + c \sum_{j<k} \max\left\{\left|\beta_j\right|, \left|\beta_k\right|\right\} \leq t, \end{aligned} \tag{2.178}$$

em que $c \geq 0$ e $t > 0$ são parâmetros de *tunning*, sendo que c controla a ponderação relativa das normas e t controla a magnitude. Note a semelhança com a definição do convexo no método *Elastic Net*, em que se tem a métrica L_1 em combinação com a métrica L_2^2.

Pode-se notar pela formulação que a métrica L_∞ par a par foi utilizada em detrimento da métrica L_∞ global. Embora em duas dimensões elas sejam iguais, seus comportamentos em $p > 2$ dimensões são bem diferentes. A utilização da métrica L_∞ global só permitiria a possibilidade de se obter um único *cluster*, que deve conter o maior coeficiente. Definindo o OSCAR através da métrica L_∞ par a par, permite-se que múltiplos grupos de diferentes tamanhos possam ser obtidos (BONDELL; REICH, 2008).

A construção do convexo K_p é então obtida da forma,

$$K_p = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p; \sum_{j=1}^{p} \left|\beta_j\right| + c \sum_{j<k} \max\left\{\left|\beta_j\right|, \left|\beta_k\right|\right\} \leq t \right\}, \tag{2.179}$$

em que $c \geq 0$ e $t > 0$ são parâmetros de *tunning*.

A descrição geométrica de K_p para o caso $p = 2$ é útil para se entender o objetivo básico do método. Vamos nos restringir ao primeiro quadrante ($\beta_1 \geq 0, \beta_2 \geq 0$), pois a descrição para os demais quadrantes pode ser feita de forma semelhante devida à simetria do convexo K_p. Inicialmente, vamos supor que $\beta_2 = 0$. Logo,

$$\begin{aligned} |\beta_1| + |\beta_2| + c\max\{|\beta_1|, |\beta_2|\} \leq t &\Rightarrow |\beta_1| + 0 + c\max\{|\beta_1|, 0\} \leq t \\ &\Rightarrow \beta_1 + c\beta_1 \leq t \\ &\Rightarrow \beta_1 \leq \frac{t}{1+c}. \end{aligned} \tag{2.180}$$

Tomando-se a igualdade, resulta que $\beta_1 = \frac{t}{1+c}$. De modo análogo, para $\beta_1 = 0$ tem-se $\beta_2 = \frac{t}{1+c}$. Com esses resultados, no plano (β_1, β_2) tem-se que a região de restrição OSCAR possui vértices $\left(0, \frac{t}{1+c}\right)$ e $\left(\frac{t}{1+c}, 0\right)$ nos eixos coordenados, para o primeiro quadrante. Agora, podem ser considerados mais três casos:

i) Caso $\beta_2 < \beta_1$

$$\begin{aligned} |\beta_1| + |\beta_2| + c\max\{|\beta_1|, |\beta_2|\} \leq t &\Rightarrow \beta_1 + \beta_2 + c\beta_1 \leq t \\ &\Rightarrow (1+c)\beta_1 + \beta_2 \leq t \\ &\Rightarrow \beta_2 \leq -(1+c)\beta_1 + t. \end{aligned} \tag{2.181}$$

Tomando-se a igualdade tem-se o segmento de reta $\beta_2 = -(1+c)\beta_1 + t$, para $\beta_2 < \beta_1$.

ii) Caso $\beta_2 > \beta_1$

$$\begin{aligned} |\beta_1| + |\beta_2| + c\max\{|\beta_1|, |\beta_2|\} \leq t &\Rightarrow \beta_1 + \beta_2 + c\beta_2 \leq t \\ &\Rightarrow \beta_1 + (1+c)\beta_2 \leq t \\ &\Rightarrow \beta_2 \leq -\frac{1}{1+c}\beta_1 + \frac{t}{1+c}. \end{aligned} \tag{2.182}$$

Tomando-se a igualdade tem-se o segmento de reta $\beta_2 = -\frac{1}{1+c}\beta_1 + \frac{t}{1+c}$, para $\beta_2 > \beta_1$.

iii) Caso $\beta_1 = \beta_2$

Estamos interessados também no caso $\beta_1 = \beta_2$, pois é nessa situação que ocorre a igualdade dos coeficientes e, nesse caso, o agrupamento das covariáveis. Em relação aos segmentos de retas obtidos para os casos $\beta_2 < \beta_1$ e $\beta_2 > \beta_1$, é possível verificar que esses segmentos de reta são concorrentes para $c > 0$, uma vez que para $c = 0$ (LASSO) isso não faz sentido. Logo, determinar o vértice da região OSCAR no plano (β_1, β_2) para o qual $\beta_1 = \beta_2$ equivale a determinar o ponto em comum entre esses segmentos de reta. Sem maiores dificuldades, é possível mostrar que esse vértice é dado por $\left(\frac{t}{2+c}, \frac{t}{2+c}\right)$. De fato, igualando os segmentos de retas obtidos em i) e ii) resulta que:

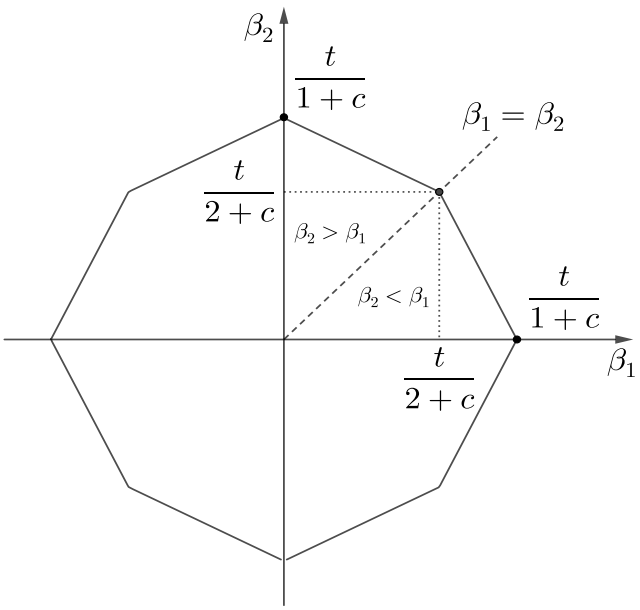
$$\begin{aligned}
-(1+c)\beta_1 + t = -\frac{1}{1+c}\beta_1 + \frac{t}{1+c} &\Rightarrow -(1+c)\beta_1 + \frac{1}{1+c}\beta_1 = \frac{t}{1+c} - t \\
&\Rightarrow \frac{-(1+c)^2\beta_1 + \beta_1}{1+c} = \frac{t - t(1+c)}{1+c} \\
&\Rightarrow \left[-(1+c)^2 + 1\right]\beta_1 = t - t(1+c) \\
&\Rightarrow \left(-1 - 2c - c^2 + 1\right)\beta_1 = t - t - tc \\
&\Rightarrow \left(-2c - c^2\right)\beta_1 = -tc \\
&\Rightarrow \beta_1 = \frac{-tc}{-2c - c^2} \\
&\Rightarrow \beta_1 = \frac{-tc}{-c(2+c)} \\
&\Rightarrow \beta_1 = \frac{t}{2+c}. \qquad (2.183)
\end{aligned}$$

Como $\beta_1 = \beta_2$, tem-se as coordenadas $\left(\frac{t}{2+c}, \frac{t}{2+c}\right)$ do ponto em comum entre os segmentos de reta obtidos nos casos i) e ii). Com comportamento semelhante nos outros quadrantes, tem-se que a forma da região de restrição OSCAR em duas dimensões é exatamente um octógono, de onde provém o nome do método (FIGURA 2.43). A partir dessa figura, a razão para o termo *octogonal* no nome do método torna-se agora clara. O parâmetro c determina o ângulo entre as semi-retas. Com vértices nos eixos coordenados e nas diagonais, o OSCAR estimula a esparsidade e a igualdade de coeficientes em graus variados, dependendo da força da correlação entre as covariáveis, do valor do parâmetro c e da localização da solução de mínimos quadrados.

Para $p = 3$, a construção é bem mais complexa pois,

$$K_p = \left\{ \boldsymbol{\beta} \in \mathbb{R}^3; \sum_{j=1}^{3} |\beta_j| + c \sum_{j<k}^{3} \max\left\{ |\beta_j|, |\beta_k| \right\} \leq t \right\}. \qquad (2.184)$$

Figura 2.43 – Representação gráfica da região de restrição OSCAR no plano (β_1, β_2).



Fonte: Do autor (2020).

Ao se tomar os máximos dois a dois no conjunto $\sum_{j=1}^{3} |\beta_j| + c \sum_{j<k}^{3} \max\{|\beta_j|, |\beta_k|\} \leq t$, segue que:

$$|\beta_1| + |\beta_2| + |\beta_3| + c\max\{|\beta_1|, |\beta_2|\} + c\max\{|\beta_1|, |\beta_3|\} + c\max\{|\beta_2|, |\beta_3|\} \leq t.$$

Ordenando $|\beta_1|$, $|\beta_2|$ e $|\beta_3|$ por $|\beta_{(1)}| \leq |\beta_{(2)}| \leq |\beta_{(3)}|$, segue que:

$$|\beta_{(1)}| + |\beta_{(2)}| + |\beta_{(3)}| + c\max\{|\beta_{(1)}|, |\beta_{(2)}|\} + c\max\{|\beta_{(1)}|, |\beta_{(3)}|\} + c\max\{|\beta_{(2)}|, |\beta_{(3)}|\} = t.$$

Desse modo,

$$|\beta_{(1)}| + |\beta_{(2)}| + |\beta_{(3)}| + c|\beta_{(2)}| + c|\beta_{(3)}| + c|\beta_{(3)}| = t \Rightarrow |\beta_{(1)}| + (1+c)|\beta_{(2)}| + (1+2c)|\beta_{(3)}| = t.$$

Essa equação define vários planos. Os vértices são obtidos pela interseção desses planos e, portanto, são retas e pontos. Note que a configuração $|\beta_{(1)}| \leq |\beta_{(2)}| \leq |\beta_{(3)}|$ é apenas uma dentre outras configurações possíveis. De forma geral, a região de penalidade OSCAR é um convexo que possui $3^p - 1$ vértices (PETRY; TUTZ, 2012). Para $p = 3$, existem 26 vértices. Esses 26 vértices são apresentados a seguir, mas a forma como eles podem ser obtidos é reservada ao Apêndice. Ao todo, existem:

1. 6 vértices considerando uma entrada não nula, obtendo-se os vértices da forma $\left(\pm\frac{t}{1+2c},0,0\right)$, $\left(0,\pm\frac{t}{1+2c},0\right)$ e $\left(0,0,\pm\frac{t}{1+2c}\right)$.

2. 12 vértices com duas entradas não nulas e iguais (em valor absoluto), considerando os vértices da forma $\left(\pm\frac{t}{2+3c},\pm\frac{t}{2+3c},0\right)$, $\left(\pm\frac{t}{2+3c},0,\pm\frac{t}{2+3c}\right)$ e $\left(0,\pm\frac{t}{2+3c},\pm\frac{t}{2+3c}\right)$.

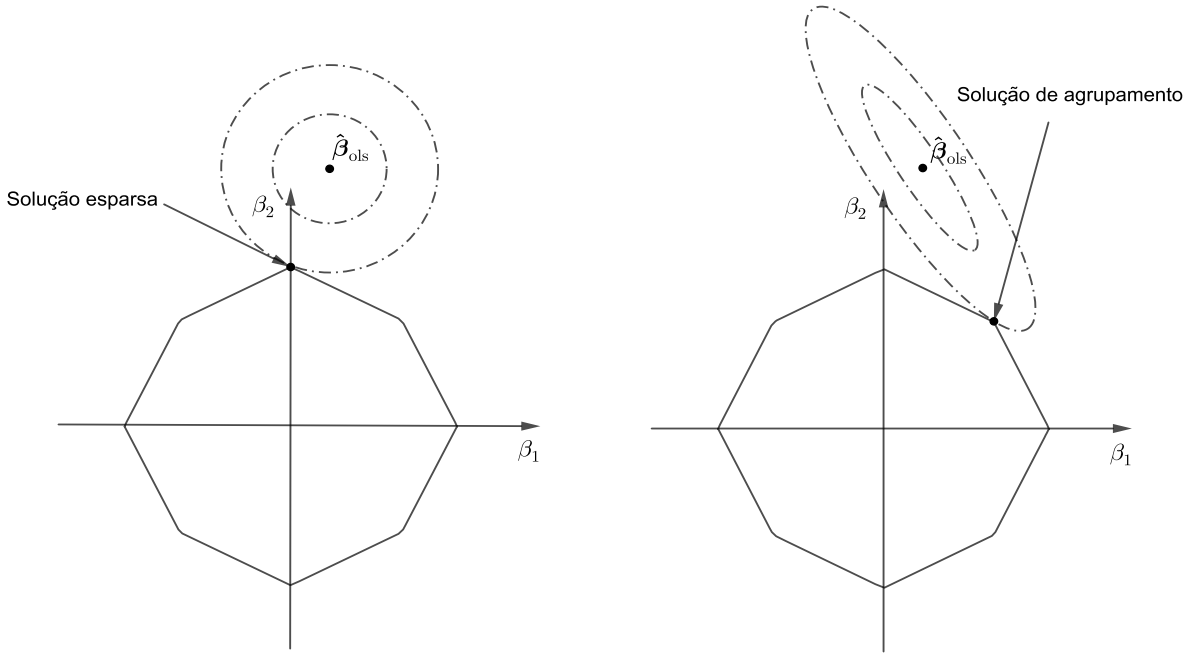
3. 8 vértices para o caso $\beta_1=\beta_2=\beta_3$, considerando nessa situação todas as permutações simples de $\left(\pm\frac{t}{3+3c},\pm\frac{t}{3+3c},\pm\frac{t}{3+3c}\right)$.

Com base na interpretação geométrica das soluções de mínimos quadrados, restritas a região de penalidade OSCAR, pode-se observar que esse método favorece as soluções a apresentarem estimativas esparsas, devido à presença de vértices nos eixos coordenados ou a apresentarem a igualdade de alguns coeficientes nos demais vértices, gerando consequentemente o agrupamento de covariáveis. Considerando o mesmo valor de c e a mesma localização das estimativas de mínimos quadrados, nas Figuras 2.44a) e 2.44b) pode ser observado como a correlação entre as covariáveis pode afetar as soluções do método OSCAR, fornecendo uma solução esparsa ou uma solução de agrupamento. Na Figura 2.44 ilustra-se o fato de que o agrupamento torna-se mais provável de ocorrer à medida que os preditores tornam-se mais correlacionados.

Figura 2.44 – Representação gráfica da solução OSCAR no plano (β_1,β_2).

(a) Contornos centrados na estimativa $\hat{\boldsymbol{\beta}}_{\text{ols}}$ com baixa correlação (Solução $\hat{\beta}_1=0$).

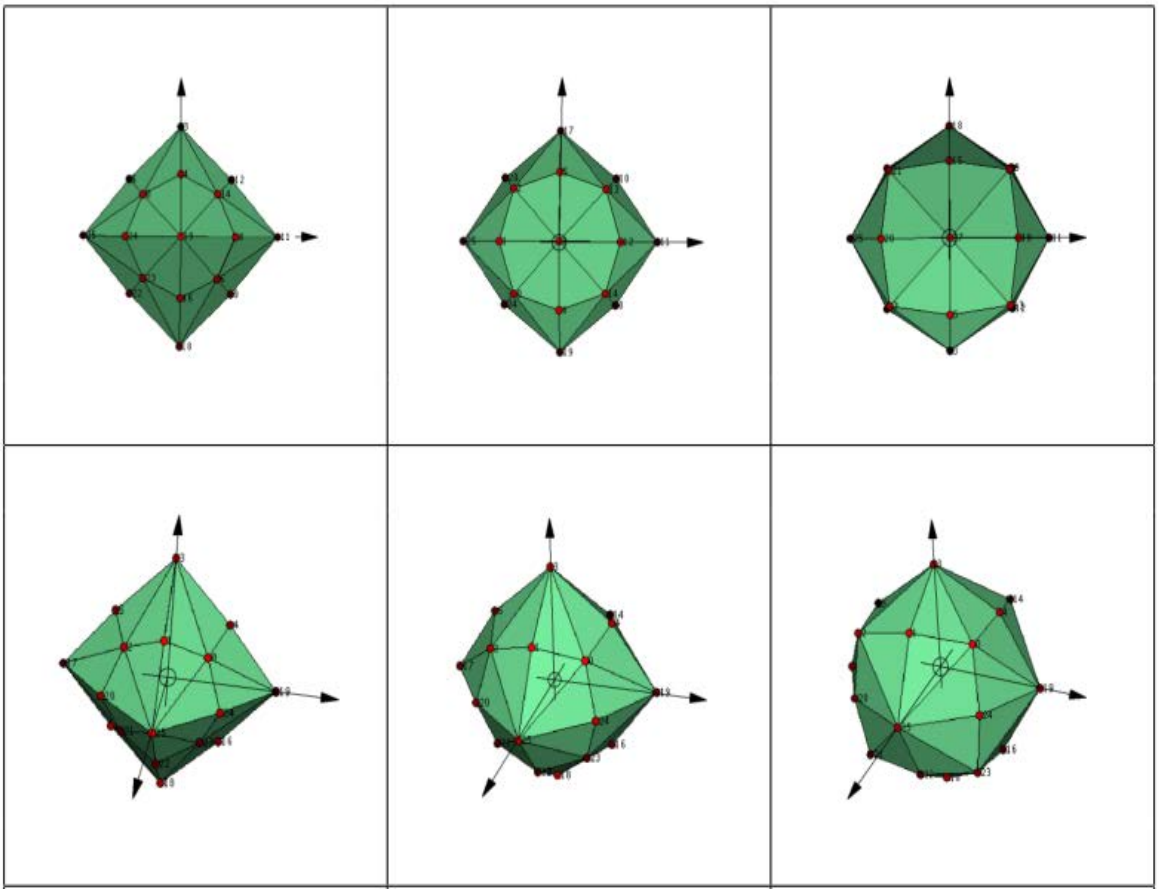
(b) Contornos centrados na estimativa $\hat{\boldsymbol{\beta}}_{\text{ols}}$ com alta correlação (Solução $\hat{\beta}_1=\hat{\beta}_2$).



Fonte: Adaptado de Bondell e Reich (2008).

Na Figura 2.45 é mostrada a região de penalidade OSCAR no $\mathbb{R}^3$ para diferentes valores do parâmetro de *tunning* c. A primeira linha da Figura 2.45 reflete mais uma vez a nomenclatura do método OSCAR. Nessa figura ilustra-se que as projeções ortogonais da região de penalidade OSCAR formam um octógono, o que pode ser mostrado usando as projeções ortogonais dos vértices em qualquer plano $\left(\beta_i, \beta_j\right)$. Por causa da simetria, na Figura 2.45 apenas uma projeção é mostrada.

Figura 2.45 – A região de penalidade OSCAR com três diferentes valores do parâmetro de *tunning* c. Na primeira linha, as projeções para um plano (β_i, β_j) são mostradas. Na segunda linha, uma visão oblíqua das penalidades são exibidas.



Fonte: Petry e Tutz (2012).

A função lagrangeana do problema de otimização do método OSCAR é dada por:

$$\begin{aligned} L(\boldsymbol{\beta}, \gamma_1, \gamma_2) &= \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 + \gamma_1 \left[\sum_{j=1}^{p} \left|\beta_j\right| + c \sum_{1 \le j < k \le p} \max\left\{ \left|\beta_j\right|, \left|\beta_k\right| \right\} - t \right] \\ &= \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 + \gamma_1 \sum_{j=1}^{p} \left|\beta_j\right| + \gamma_1 c \sum_{1 \le j < k \le p} \max\left\{ \left|\beta_j\right|, \left|\beta_k\right| \right\} - \gamma_1 t \\ &= \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 + \gamma_1 \sum_{j=1}^{p} \left|\beta_j\right| + \gamma_2 \sum_{1 \le j < k \le p} \max\left\{ \left|\beta_j\right|, \left|\beta_k\right| \right\} - \gamma_1 t, \end{aligned} \tag{2.185}$$

em que $\gamma_1, \gamma_2 \geq 0$.

De modo semelhante ao que foi realizado para o método *Elastic Net*, observe que:

$$\begin{aligned}\gamma_1 \sum_{j=1}^{p} |\beta_j| + \gamma_2 \sum_{1\leq j<k\leq p} \max\left\{|\beta_j|, |\beta_k|\right\} \leq \gamma_1 t &\Rightarrow \frac{\gamma_1}{\gamma_1+\gamma_2} \sum_{j=1}^{p} |\beta_j| + \frac{\gamma_2}{\gamma_1+\gamma_2} \sum_{1\leq j<k\leq p} \max\left\{|\beta_j|, |\beta_k|\right\} \leq \frac{\gamma_1 t}{\gamma_1+\gamma_2} \\ &\Rightarrow \frac{\gamma_1}{\gamma_1+\gamma_2} \sum_{j=1}^{p} |\beta_j| + \left(1 - \frac{\gamma_1}{\gamma_1+\gamma_2}\right) \sum_{1\leq j<k\leq p} \max\left\{|\beta_j|, |\beta_k|\right\} \leq t' \\ &\Rightarrow \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{1\leq j<k\leq p} \max\left\{|\beta_j|, |\beta_k|\right\} \leq t', \qquad (2.186)\end{aligned}$$

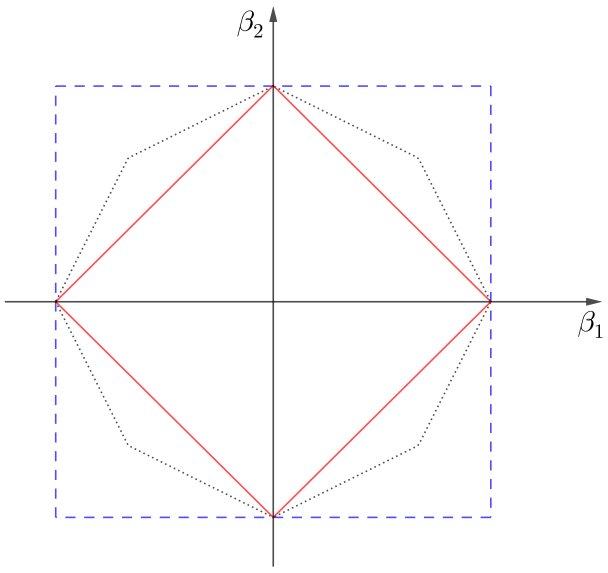
em que $\alpha = \frac{\gamma_1}{\gamma_1+\gamma_2}$ e por construção $0 \leq \alpha \leq 1$.

Portanto, a função lagrangeana do problema de otimização do método OSCAR pode ser reexpressa da seguinte forma:

$$\left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 + \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{1\leq j<k\leq p} \max\left\{|\beta_j|, |\beta_k|\right\}. \qquad (2.187)$$

Considerando essa nova reparametrização, na Figura 2.46 apresenta-se uma representação da região de restrição OSCAR no plano (β_1, β_2) para $\alpha = 0,5$ (octógono, pontilhado) e $\alpha = 0$ (quadrado, tracejado). Note para esse último caso que o método OSCAR fornece uma região de penalidade quadrada que reforça o agrupamento entre as covariáveis, mas não a seleção das mesmas. Pode-se observar ainda nessa figura que a restrição LASSO é um caso particular da restrição OSCAR, para $\alpha = 1$ (contínuo).

Figura 2.46 – Representação gráfica da região de restrição OSCAR, para os casos em que $\alpha = 0$ (quadrado, tracejado), $\alpha = 0,5$ (octógono, pontilhado) e $\alpha = 1$ (LASSO, contínuo).



Fonte: Do autor (2020).

2.10 O método *Pairwise Absolute Clustering and Sparsity* (PACS)

Sharma, Bondell e Zhang (2013) propuseram uma modificação na penalização do método OSCAR, definindo um novo método que foi denominado agrupamento e esparsidade absoluta de pares (PACS, do inglês "*Pairwise Absolute Clustering and Sparsity*"). O método visa identificar grupos de preditores que formam *clusters*, como grupos de preditores altamente correlacionados.

A igualdade de coeficientes no método PACS é obtida pela penalização das diferenças e as somas par a par entre os coeficientes, com a utilização de pesos não negativos. O problema de otimização de mínimos quadrados restritos para a regressão PACS é dado por:

$$\begin{gathered}\hat{\boldsymbol{\beta}}_{\text{PACS}} = \arg\min_{\boldsymbol{\beta}} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 \\ \text{restrito a} \\ \sum_{j=1}^{p} w_j \left|\beta_j\right| + \sum_{1\leq j<k\leq p} w_{jk(-)} \left|\beta_k - \beta_j\right| + \sum_{1\leq j<k\leq p} w_{jk(+)} \left|\beta_k + \beta_j\right| \leq t\end{gathered}, \tag{2.188}$$

em que t é o parâmetro de ajuste e $\boldsymbol{w}$ são os pesos.

As estimativas são obtidas pela minimização da seguinte expressão:

$$\left\| \boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \right\|^2 + \gamma \left\{ \sum_{j=1}^{p} w_j \left|\beta_j\right| + \sum_{1\leq j<k\leq p} w_{jk(-)} \left|\beta_k - \beta_j\right| + \sum_{1\leq j<k\leq p} w_{jk(+)} \left|\beta_k + \beta_j\right| \right\}. \tag{2.189}$$

A penalidade em (2.189) consiste da norma ponderada L_1 dos coeficientes, o que favorece a esparsidade, e uma penalidade nas diferenças e somas de pares dos coeficientes, o que permite a igualdade entre eles. A penalidade ponderada nas diferenças de pares de coeficientes favorece que coeficientes com o mesmo sinal sejam definidos como iguais. Por sua vez, a penalidade ponderada nas somas de pares de coeficientes favorece que os coeficientes com sinais opostos sejam definidos como iguais. Os pesos são números não negativos pré-especificados e a forma como eles podem ser escolhidos será estudada posteriormente. A função objetivo do método PACS é uma função convexa, pois é uma soma de funções convexas. Em particular, se $\boldsymbol{X}'\boldsymbol{X}$ é de posto coluna completo, essa função é estritamente convexa (SHARMA; BONDELL; ZHANG, 2013).

Pode-se observar que o método OSCAR é um caso particular do método PACS, uma vez que $\max\left\{\left|\beta_j\right|,\left|\beta_k\right|\right\} = 0,5\left(\left|\beta_k-\beta_j\right|+\left|\beta_k+\beta_j\right|\right)$. A demonstração desse resultado pode ser vista no Apêndice. Dessa maneira,

$$\begin{aligned}(1-\alpha)\sum_{1\leq j<k\leq p}\max\left\{\left|\beta_j\right|,\left|\beta_k\right|\right\} &= 0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left(\left|\beta_k-\beta_j\right|+\left|\beta_k+\beta_j\right|\right)\\ &= 0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_k-\beta_j\right|+0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_k+\beta_j\right|. \quad (2.190)\end{aligned}$$

Com base nesse último resultado e utilizando a expressão (2.187), segue então que a Lagrangeana do método OSCAR pode ser equivalentemente expressa como o mínimo de:

$$\left\|\boldsymbol{y}-\sum_{j=1}^{p}\beta_j\boldsymbol{x}_j\right\|^2+\gamma\left\{\alpha\sum_{j=1}^{p}\left|\beta_j\right|+0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_j-\beta_k\right|+0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_j+\beta_k\right|\right\}, \quad (2.191)$$

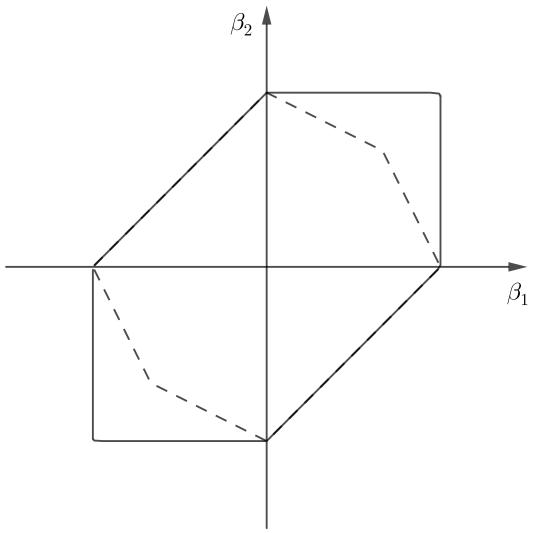
e, portanto, o OSCAR é um caso particular do PACS.

Além disso, podemos observar como cada parte da penalidade em (2.191) contribui para o agrupamento das covariáveis. Seja o plano (β_1,β_2). Considerando na restrição em (2.191) somente $\alpha\sum_{j=1}^{p}\left|\beta_j\right|+0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_j-\beta_k\right|$, é possível verificar que conforme o valor de α diminui, a restrição tende a favorecer agrupamentos pela maior evidência dos vértices no primeiro e terceiro quadrantes (FIGURA 2.47a)). Nessa situação, o termo $\left|\beta_j-\beta_k\right|$ favorece o agrupamento de variáveis preditoras que possuam o mesmo sinal em seus coeficientes (primeiro e terceiro quadrantes no $\mathbb{R}^2$). A restrição $\alpha\sum_{j=1}^{p}\left|\beta_j\right|+0,5\,(1-\alpha)\sum_{1\leq j<k\leq p}\left|\beta_j+\beta_k\right|$ possui uma interpretação análoga. O que difere nesse caso é que o termo $\left|\beta_j+\beta_k\right|$ favorece o agrupamento de variáveis preditoras que possuam sinais diferentes em seus coeficientes no $\mathbb{R}^2$, o que ocorre no segundo e quarto quadrantes (FIGURA 2.47b)).

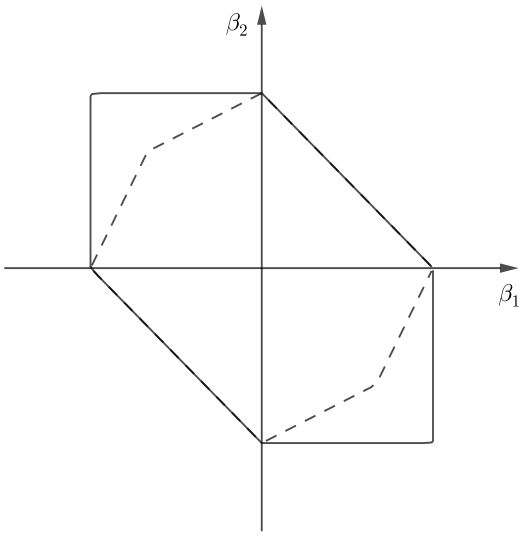
Sharma, Bondell e Zhang (2013) destacam que a formulação PACS apresenta certas vantagens em relação a formulação original OSCAR, apresentada em (2.185). Como no método OSCAR, as soluções PACS podem ser calculadas via programação quadrática. No entanto, as soluções também podem ser calculadas usando uma aproximação quadrática local da penalidade, que não é diretamente aplicável à formulação original do OSCAR. A última estratégia é superior à programação quadrática, uma vez que a programação quadrática torna-se dispendiosa em computação e não é viável para um número grande e até moderado de parâmetros,

Figura 2.47 – Diferentes restrições na estimação OSCAR, considerando $\alpha = 0,50$ (contínuo) e $\alpha = 0,75$ (tracejado).

(a) $\alpha \sum_{j=1}^{p} |\beta_j| + 0,5\,(1-\alpha) \sum_{1 \le j < k \le p} |\beta_j - \beta_k|$.



(b) $\alpha \sum_{j=1}^{p} |\beta_j| + 0,5\,(1-\alpha) \sum_{1 \le j < k \le p} |\beta_j + \beta_k|$.



Fonte: Do autor (2020).

enquanto a estratégia de aproximação quadrática local continua a ser viável nesses casos e pode ser convenientemente implementada em um *software* padrão.

Com a formulação proposta, Sharma, Bondell e Zhang (2013) mostram em seu trabalho que o método PACS necessita apenas de um parâmetro de *tunning*, enquanto o OSCAR possui dois parâmetros de *tunning*. Essa redução no número de parâmetros de *tunning* melhora consideravelmente o custo computacional. Na Figura 2.48 ilustra-se a flexibilidade da abordagem PACS em relação a abordagem OSCAR, em termos da região de restrição no plano (β_1, β_2). Na Figura 2.48a), pode-se observar que, embora α varie entre 0 e 1, a região de restrição permanece simétrica nos quatro eixos de simetria, ou seja, a região de restrição OSCAR tem a mesma forma em todos os quatro quadrantes. Nas Figuras 2.48b) e 2.48c) são exibidas duas regiões particulares da restrição PACS, obtidas a partir de duas escolhas diferentes de pesos. A escolha dos pesos interfere na forma da região de restrição, o que sugere a flexibilidade da abordagem PACS em detrimento da abordagem OSCAR.

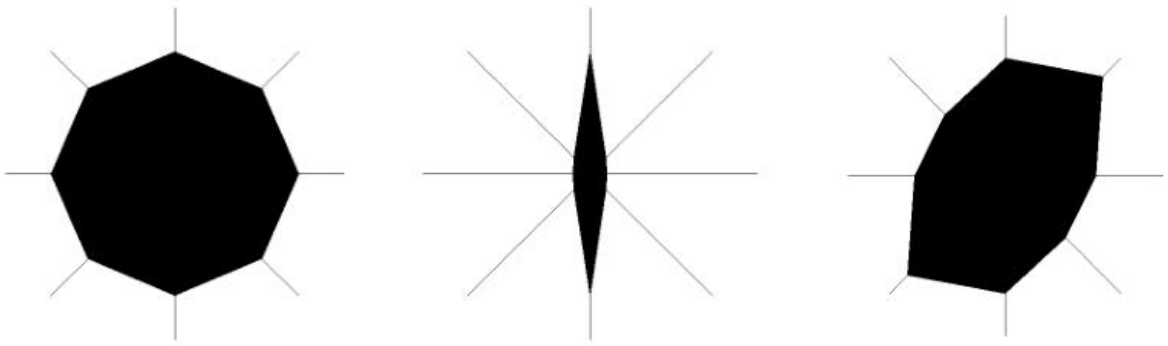
Os autores apresentam três estratégias diferentes para a escolha dos pesos $\boldsymbol{w}$. As três opções são: pesos determinados por um esquema de escalonamento dos preditores, pesos adaptativos de dados para propriedades oraculares e uma abordagem para incorporar as correlações nos pesos entre as covariáveis. Em relação a segunda estratégia, um procedimento oracular

Figura 2.48 – Ilustração para representar a flexibilidade da abordagem PACS sobre a abordagem OSCAR em termos das regiões de restrição sombreadas no plano (β_1, β_2).

(a) Região de restrição OSCAR para um valor fixo de c.

(b) Região de restrição PACS com um esquema de ponderação.

(c) Região de restrição PACS com um esquema de ponderação diferente.



Fonte: Sharma, Bondell e Zhang (2013).

é aquele que deve consistentemente identificar o modelo correto e obter estimativas acuradas ótimas.

Devido aos objetivos do presente trabalho será utilizada a terceira estratégia para a escolha dos pesos, ou seja, o agrupamento de preditores será realizado com base na correlação entre eles. Essa abordagem é explorada em Tutz e Ulbricht (2009), no contexto da regressão *Ridge*. Para esse fim, considere a incorporação de correlações no esquema de ponderação, tal como é dado por: $w_j = 1$, $w_{jk(-)} = \left(1 - r_{jk}\right)^{-1}$ e $w_{jk(+)} = \left(1 + r_{jk}\right)^{-1}$, em que r_{jk} representa a correlação entre as covariáveis j e k $(1 \leq j < k \leq p)$. Intuitivamente, os pesos $w_{jk(-)}$ penalizam mais fortemente as diferenças nos coeficientes quando as covariáveis são altamente correlacionadas positivamente e os pesos $w_{jk(+)}$ penalizam fortemente as somas quando as correlações são altas, porém negativas. Embora essa ponderação não favoreça preditores não correlacionados a apresentarem coeficientes iguais, ela favorece mais fortemente os pares de preditores altamente correlacionados a apresentarem igualdade em seus coeficientes. Os pesos adaptativos que incorporam esses termos de correlações são então dados por: $w_j = \left|\tilde{\beta}_j\right|^{-1}$, $w_{jk(-)} = \left(1 - r_{jk}\right)^{-1}\left|\tilde{\beta}_k - \tilde{\beta}_j\right|^{-1}$ e $w_{jk(+)} = \left(1 + r_{jk}\right)^{-1}\left|\tilde{\beta}_k + \tilde{\beta}_j\right|^{-1}$, para $1 \leq j < k \leq p$.

3 MATERIAL E MÉTODOS

3.1 Métodos propostos

A ideia base do método SPCA consiste em utilizar o fato de que os componentes principais podem ser obtidos a partir de uma regressão *Ridge*, utilizando como pseudodados os componentes principais $\boldsymbol{z}_i = \lambda_i \boldsymbol{u}_i$, em que o autovalor λ_i e o autovetor $\boldsymbol{u}_i$ podem ser obtidos pela decomposição em valores singulares de $\boldsymbol{X}$ ($\boldsymbol{X} = \boldsymbol{UDV}'$). A partir desse fato acrescenta-se a penalização L_1, definindo-se uma regressão *Elastic Net* e, consequentemente, soluções esparsas para os componentes principais podem ser obtidas (ZOU; HASTIE; TIBSHIRANI, 2006).

O fundamento teórico dos métodos propostos neste trabalho foi norteado por uma ideia semelhante, consistindo em se utilizar o convexo K_p definido pelas restrições dos métodos OSCAR e PACS para se obterem aproximações para os componentes principais que possuirão as propriedades de esparsidade e de agrupamento (classificação em "*clusters*"). Em relação ao método OSCAR, o primeiro método será denominado *Sparse Group for Principal Component Analysis* (SGPCA). Por sua vez, em relação ao método PACS, o segundo método será denominado *Pairwise Absolute Clustering and Sparsity for Principal Component Analysis* (PACSPCA). A vantagem de se obterem componentes principais com fator de agrupamento por coeficientes iguais reside na possível melhora e facilidade de interpretação dos componentes, o que é um dos grandes problemas da teoria.

Teorema 3.1.1: Para cada $i = 1, 2, \ldots, p$, seja $\boldsymbol{z}_i = \lambda_i \boldsymbol{u}_i$ o i-ésimo componente principal. Se:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma \left\{ \sum_{j=1}^{p} |\beta_j| + c \sum_{j<k}^{p} \max \left\{ |\beta_j|, |\beta_k| \right\} \right\} \tag{3.1}$$

ou

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \gamma \left\{ \sum_{j=1}^{p} w_j |\beta_j| + \sum_{1 \le j < k \le p} w_{jk(-)} |\beta_k - \beta_j| + \sum_{1 \le j < k \le p} w_{jk(+)} |\beta_k + \beta_j| \right\}, \tag{3.2}$$

então $\hat{\boldsymbol{V}}_i = \dfrac{\hat{\boldsymbol{\beta}}}{\left\|\hat{\boldsymbol{\beta}}\right\|}$ é a uma aproximação para o vetor de *loadings* $\boldsymbol{V}_i$, com propriedades de esparsidade e de agrupamento, considerando os métodos SGPCA e PACSPCA, respectivamente.

3.2 Formação de grupos via simulação e cenários de simulação

Para a compreensão de como os grupos de variáveis podem ser construídos em exemplos simulados, uma devida atenção deve ser dada à estrutura de correlação das variáveis de interesse. Uma vez que o método dos componentes principais é uma técnica da estatística multivariada que modela a estrutura de covariâncias, no presente trabalho adotou-se como critério para a formação de grupos, a estrutura de correlação entre as variáveis. Nesse sentido, em cada exemplo buscou-se definir como grupo, as variáveis que possuíssem correlações (positiva ou negativa) fortes $[0,7-0,9)$ ou muito fortes $[0,9-1)$ entre si e que apresentassem correlações (positiva ou negativa) desprezíveis $[0-0,3)$, fracas $[0,3-0,5)$ ou moderadas $[0,5-0,7)$ com as variáveis de outros grupos.

Com o objetivo de avaliar os métodos propostos foram apresentados dois cenários sintéticos. Neste estudo de simulação foram comparados os resultados dos métodos SGPCA e PACSPCA com os resultados de métodos conhecidos na literatura e amplamente utilizados, os métodos PCA padrão e SPCA. O objetivo específico consistiu em avaliar o desempenho desses métodos na identificação de grupos e na capacidade de cada um deles de fornecer uma representação esparsa em termos da complexidade resultante de cada um dos componentes principais nos exemplos propostos. Nesse sentido, a seguir são apresentados os dois cenários utilizados para a avaliação dos métodos propostos.

3.2.1 Cenário 1

Neste primeiro cenário utilizou-se o exemplo concebido por Zou, Hastie e Tibshirani (2006). Os autores consideraram três variáveis, as quais eles denominaram de "fatores ocultos". São elas:

$$V_1 \sim N(0,290),\ V_2 \sim N(0,300) \text{ e } V_3 = -0,3V_1 + 0,925V_2 + \varepsilon, \tag{3.3}$$

em que $\varepsilon \sim N(0,1)$, sendo V_1, V_2 e ε independentes.

Dessa maneira, foram construídas 10 variáveis a partir de V_1, V_2 e V_3, como se segue:

$$\begin{aligned} X_i &= V_1 + \varepsilon_i,\ \varepsilon_i \sim N(0,1),\ i = 1,2,3,4. \\ X_i &= V_2 + \varepsilon_i,\ \varepsilon_i \sim N(0,1),\ i = 5,6,7,8. \\ X_i &= V_3 + \varepsilon_i,\ \varepsilon_i \sim N(0,1),\ i = 9,10, \end{aligned} \tag{3.4}$$

em que $\{\varepsilon_i\}$ são independentes, $i = 1,\ldots,10$.

Podemos observar inicialmente que os fatores V_1, V_2 e V_3 são normalmente distribuídos, sendo a variável V_3 uma combinação linear das duas primeiras. Além disso, as variâncias dos fatores subjacentes V_1, V_2 e V_3 são 290, 300 e 283,79, respectivamente. O número de variáveis associadas aos fatores V_1, V_2 e V_3 são 4 (X_1, X_2, X_3, X_4), 4 (X_5, X_6, X_7, X_8) e 2 (X_9, X_{10}), respectivamente. Em razão de V_1 e V_2 gerarem um número maior de variáveis, Zou, Hastie e Tibshirani (2006) destacam que esses fatores são igualmente importantes, apresentando uma relevância superior em relação ao fator V_3.

Mediante a forma como as variáveis $(X_1, X_2, \ldots, X_8)$ foram obtidas, preliminarmente podemos afirmar que existem dois grupos bem definidos neste cenário, a saber o grupo I formado por (X_1, X_2, X_3, X_4) e o grupo II formado por (X_5, X_6, X_7, X_8).

3.2.2 Cenário 2

Esse novo cenário é uma adaptação do exemplo idealizado por Zou, Hastie e Tibshirani (2006) e que foi utilizado no cenário 1. No presente cenário serão consideradas $p = 12$ variáveis e cinco "fatores ocultos", sendo esses últimos dados por:

$$\begin{array}{c} V_1 \sim N(0, 290),\ V_2 \sim N(0, 300),\ V_3 \sim N(0, 310) \\ V_4 = 0,7V_1 - 0,8V_2 + \varepsilon, V_5 = -0,6V_1 + 0,55V_2 + \varepsilon \end{array}, \tag{3.5}$$

em que $\varepsilon \sim N(0, 1)$, sendo V_1, V_2, V_3 e ε independentes.

Dessa maneira, foram construídas 12 variáveis a partir de V_1, V_2, V_3, V_4 e V_5, como se segue:

$$\begin{aligned} X_i &= V_1 + \varepsilon_i,\ \varepsilon_i \sim N(0, 1),\ i = 1, 2, 3. \\ X_i &= V_2 + \varepsilon_i,\ \varepsilon_i \sim N(0, 1),\ i = 4, 5, 6. \\ X_i &= V_3 + \varepsilon_i,\ \varepsilon_i \sim N(0, 1),\ i = 7, 8. \\ X_i &= V_4 + \varepsilon_i,\ \varepsilon_i \sim N(0, 1),\ i = 9, 10. \\ X_i &= V_5 + \varepsilon_i,\ \varepsilon_i \sim N(0, 1),\ i = 11, 12, \end{aligned} \tag{3.6}$$

em que $\{\varepsilon_i\}$ são independentes, $i = 1, \ldots, 12$.

Podemos observar que os fatores V_1, V_2, V_3, V_4 e V_5 são normalmente distribuídos, sendo V_4 e V_5 formados a partir de combinações lineares de V_1 e V_2. Além disso, as variâncias dos fatores subjacentes V_1, V_2, V_3, V_4 e V_5 são 290, 300, 310, 335,1 e 196,15, respectivamente. O número de variáveis associadas aos fatores V_1, V_2, V_3, V_4 e V_5 são 3 (X_1, X_2, X_3), 3 (X_4, X_5, X_6), 2

(X_7, X_8), 2 (X_9, X_{10}) e 2 (X_{11}, X_{12}), respectivamente. Novamente, em razão de gerarem um número maior de variáveis, os fatores V_1 e V_2 são igualmente importantes, apresentando relevância superior em relação aos fatores V_3, V_4 e V_5.

Com base nas considerações sobre como os grupos podem ser formados em exemplos sintéticos, pode-se verificar que no cenário 2 existem três grupos bem definidos. A saber, as variáveis (X_1, X_2, X_3) definem o grupo I, enquanto as variáveis (X_4, X_5, X_6) definem o grupo II e as variáveis (X_7, X_8) definem o grupo III. Em conformidade com a maneira como as variáveis $(X_9, X_{10}, X_{11}, X_{12})$ foram construídas, espera-se que essas apresentem correlações moderadas ou fortes com as variáveis dos grupos I e II. Isso é esperado em razão das variáveis $(X_9, \ldots, X_{12})$, juntamente com as variáveis dos grupos I e II, possuírem em comum os fatores subjacentes V_1 e V_2. Diante da forma como as variáveis foram construídas, no presente cenário espera-se que os métodos propostos consigam identificar corretamente os grupos subjacentes. Além disso, buscou-se analisar, neste cenário, como os novos métodos se comportam em uma situação em que as variáveis de um grupo possam se correlacionar com variáveis de outro grupo, apresentando correlações moderadas ou fortes.

No artigo original, Zou, Hastie e Tibshirani (2006) utilizaram a matriz de covariâncias exata ou populacional de $(X_1, X_2, \ldots, X_{10})$. Para obter os componentes a partir dos métodos PCA, SPCA, SGPCA e PACSPCA foi simulada a matriz $\boldsymbol{X}$ nos dois cenários, adotando como tamanho amostral $n = 100$ observações. Nesse caso, optou-se por utilizar a matriz de covariâncias amostrais para a formação dos componentes principais.

3.2.3 Validação dos resultados dos cenários 1 e 2

Na presente subseção objetivou-se avaliar a porcentagem de acerto dos métodos PCA padrão, SPCA, SGPCA e PACSPCA em agrupar corretamente as variáveis nos seus respectivos grupos formulados nos cenários 1 e 2. Para isso, foram realizadas $N = 1000$ simulações, considerando os tamanhos de amostra iguais a $n = 30$, $n = 50$ e $n = 100$ observações.

Em cada simulação, foram consideradas as seguintes etapas gerais:

1. Gerar a matriz de observações $\boldsymbol{X}_{n \times p}$ conforme a descrição apresentada nos cenários 1 e 2;
2. Padronizar a matriz de observações $\boldsymbol{X}_{n \times p}$;
3. Obter para cada método os vetores de *loadings* do primeiro e segundo componentes principais.

No cenário 1 definiu-se como grupos bem definidos ou bem caracterizados os grupos I e II, constituídos pelas variáveis (X_1, X_2, X_3, X_4) e (X_5, X_6, X_7, X_8), respectivamente. Por sua vez, no cenário 2 definiu-se como grupos bem definidos ou bem caracterizados os grupos I (X_1, X_2, X_3), II (X_4, X_5, X_6) e III (X_7, X_8).

A cada simulação, os métodos PCA padrão, SPCA, SGPCA e PACSPCA foram avaliados conforme a função indicadora S_i $(i = 1, 2, \ldots, N)$, definida por:

$$S_i = \begin{cases} 1, & \text{se as variáveis são agrupadas corretamente} \\ 0, & \text{se as variáveis não são agrupadas corretamente} \end{cases} \tag{3.7}$$

Um conjunto de variáveis é dito como pertencente a um grupo caso apresente *loadings* iguais. Nessa avaliação serão considerados somente os dois primeiros componentes principais de cada método avaliado. Para o entendimento de como foi feita a avaliação dos métodos quanto aos agrupamentos, vamos nos restringir ao método PCA padrão e ao seu primeiro componente principal (PC1) no cenário 1. Em cada uma das $N = 1000$ simulações, a função S_i receberá o valor 1 quando o método PCA padrão for capaz de fornecer no componente PC1 o mesmo valor de *loading* para as variáveis (X_1, X_2, X_3, X_4) e o mesmo valor de *loading* para as variáveis (X_5, X_6, X_7, X_8), de tal forma que cada um desses valores seja diferente um do outro. O mesmo raciocínio pode ser aplicado para os outros métodos no cenário 1. Esse raciocínio também pode ser estendido para a avaliação dos métodos quanto aos agrupamentos no cenário 2.

Uma vez que cada *loading* é uma quantidade pertencente ao intervalo $[0, 1)$ e com o intuito de que a avaliação de cada método fosse realizada da forma mais rigorosa possível, considerou-se que um conjunto de variáveis são pertencentes ao mesmo grupo caso apresentem *loadings* iguais até a quinta casa decimal. A porcentagem (P_o) de vezes que cada método agrupou corretamente as variáveis nas simulações foi calculada por:

$$P_\text{o} = \frac{1}{N} \sum_{i=1}^{N} S_i \times 100, \tag{3.8}$$

em que S_i representa o valor da função indicadora na i-ésima simulação e N é o número de simulações $(N = 1000)$.

3.3 Exemplos com dados reais

Os métodos propostos SGPCA e PACSPCA também foram aplicados a dois conjuntos de dados reais para a avaliação dos mesmos.

3.3.1 Exemplo 1

Os dados utilizados no primeiro exemplo foram extraídos da Revista *Motor Trend US Magazine* de 1974 e posteriormente foram incorporados à biblioteca padrão do *R* no pacote *dataset*. Este conjunto de dados consiste em observações sobre o consumo de combustível de 32 modelos de carros entre os anos de 1973 e 1974, sendo que para cada carro foram avaliadas 11 variáveis ou recursos, expressos em unidades de medida dos Estados Unidos. As variáveis em questão são:

1. *mpg* (milhas por galão (EUA)) - carros mais potentes e mais pesados tendem a consumir mais combustível.
2. *cyl* (número de cilindros) - carros mais potentes geralmente têm mais cilindros.
3. *disp* (deslocamento) (cu.in.) - o volume combinado dos cilindros do motor.
4. *hp* (potência bruta) - é uma medida da potência gerada pelo carro.
5. *drat* (relação do eixo traseiro) - descreve como uma rotação do eixo de transmissão corresponde à rotação das rodas. Valores mais altos diminuirão a eficiência do combustível.
6. *wt* (peso (1000 lbs)) - peso dos carros.
7. *qsec* (tempo de 1/4 de milha) - a velocidade e a aceleração dos carros.
8. *vs* (bloco do motor) - indica se o motor do veículo tem o formato de um "V" ($vs = 0$) ou se é um formato "em linha" ($vs = 1$), que é mais comum.
9. *am* (transmissão) - indica se a transmissão do carro é automática (0) ou manual (1).
10. *gear* (número de marchas à frente) - os carros esportivos tendem a ter mais marchas.
11. *carb* (número de carburadores) - associado a motores mais potentes.

Primeiramente, pode-se observar que as variáveis apresentadas podem ser classificadas em duas categorias, sendo uma relacionada à potência dos carros (*cyl*, *disp*, *hp*, *wt*, *carb*) e as varáveis relacionadas à economia (*mpg*, *drat*, *qsec*, *vs*, *am*, *gear*). Como já destacado, as variáveis possuem escalas de medida diferentes. Esse fato reforça a necessidade da padronização da matriz de observações.

3.3.2 Exemplo 2

O conjunto de dados que será analisado no segundo exemplo é resultado de uma análise química de vinhos cultivados em uma região específica da Itália, derivados de três cultivares diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. As variáveis (ou atributos) são:

X_1: álcool - teor de álcool, relatado em unidades de álcool por volume.

X_2: ácido málico - um dos principais ácidos orgânicos encontrados em vinhos. Embora seja encontrado em quase todas as frutas e bagas, seu sabor é mais proeminente nas maçãs verdes. Nesse caso, ele projeta um sabor azedo no vinho.

X_3: cinza - A cinza é simplesmente a matéria inorgânica deixada após a evaporação e a incineração.

X_4: alcalinidade da cinza - a alcalinidade da cinza determina o quão básica a cinza em um vinho pode ser, em oposição à sua acidez.

X_5: magnésio - o magnésio é um metal que afeta o sabor do vinho.

X_6: fenóis totais - os fenóis são substâncias químicas que afetam o sabor, a cor e a sensação do vinho na boca, ou seja, a textura do vinho.

X_7: flavonóides - os flavonóides são um tipo de fenol. Possuem ação antioxidante até maior que a vitamina E, e por isso, protegem o coração dos efeitos das gorduras.

X_8: fenóis não flavonóides - os não flavonóides são outro tipo de fenol.

X_9: proantocianinas - as proantocianidinas são outro tipo de fenol. São responsáveis pelas sensações gustativas dos vinhos, nomeadamente ao nível da adstringência, assumindo ainda um importante papel no envelhecimento do vinho.

X_{10}: intensidade de cor - representa o quão escuro é o vinho.

X_{11}: matiz - a matiz de um vinho, normalmente é determinada pela cor da cultivar utilizada.

X_{12}: OD280/OD315 de vinhos diluídos - medições do teor de proteínas.

X_{13}: prolina - um aminoácido presente nos vinhos.

O conjunto de dados possui $n = 178$ observações e nenhum valor ausente. Esse conjunto de dados pode ser obtido no pacote *bootcluster* (YU, 2017) da biblioteca do *R*.

3.4 Padronização e decomposição em valores singulares da matriz de observações

Com base nos cenários de simulação e nos exemplos com dados reais, o procedimento inicial foi padronizar a matriz de observações $\boldsymbol{X}_{n\times p}$, em que n e p são o número de observações e o número de variáveis, respectivamente. Supondo que se tem um conjunto com p variáveis aleatórias $(X_1, X_2, \ldots, X_p)$, segue que a matriz de dados $\boldsymbol{X}$ pode ser representada da seguinte forma:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \tag{3.9}$$

em que $\boldsymbol{x}'_{(j)} = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ corresponde ao vetor amostral da variável X_j $(j = 1, 2, \ldots, p)$.

Para padronizar a matriz $\boldsymbol{X}$ seguiram-se duas etapas. A primeira consistiu em expressar $\boldsymbol{X}$ em sua forma centrada $\boldsymbol{X}_{\mathrm{c}}$, de dimensões $(n \times p)$, que é obtida da seguinte forma:

$$\boldsymbol{X}_{\mathrm{c}} = \left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\boldsymbol{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \ldots & x_{np} - \bar{x}_p \end{bmatrix}, \tag{3.10}$$

em que $\boldsymbol{I}_{n\times n}$ é a matriz identidade e $\boldsymbol{J}_{n\times n}$ é uma matriz quadrada de 1's. A matriz $\boldsymbol{I} - \left(1/n\right)\boldsymbol{J}$, que pré-multiplica $\boldsymbol{X}$, é chamada de matriz de *centering*.

Para completar a padronização de $\boldsymbol{X}$, na segunda etapa foi necessário reduzir cada vetor $\boldsymbol{x}'_{(j)} = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ a sua escala unitária. Para isso, utilizou-se a matriz diagonal $\boldsymbol{D}$, que é dada por:

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{s_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{s_p} \end{bmatrix}, \tag{3.11}$$

em que s_j corresponde ao desvio padrão amostral observado para cada vetor de observações $\boldsymbol{x}'_{(j)} = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)$ de $\boldsymbol{X}$.

A matriz de observações em sua forma padronizada $\boldsymbol{X}_{\mathrm{p}}$ é obtida pós multiplicando $\boldsymbol{X}_{\mathrm{c}}$ por $\boldsymbol{D}$, pois as colunas de $\boldsymbol{X}_{\mathrm{c}}$ são multiplicadas pelo inverso de cada desvio padrão correspondente da diagonal de $\boldsymbol{D}$. Logo,

$$\begin{aligned}\boldsymbol{X}_{\mathrm{p}} = \boldsymbol{X}_{\mathrm{c}}\boldsymbol{D} &= \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \dots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \dots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \dots & x_{np}-\bar{x}_p \end{bmatrix} \begin{bmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{s_p} \end{bmatrix} \\ &= \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \dots & \frac{x_{1p}-\bar{x}_p}{s_p} \\ \frac{x_{21}-\bar{x}_1}{s_1} & \frac{x_{22}-\bar{x}_2}{s_2} & \dots & \frac{x_{2p}-\bar{x}_p}{s_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \frac{x_{n2}-\bar{x}_2}{s_2} & \dots & \frac{x_{np}-\bar{x}_p}{s_p} \end{bmatrix}. \end{aligned} \tag{3.12}$$

Para a utilização dos métodos SGPCA e PACSPCA, os componentes principais foram obtidos por meio da decomposição em valores singulares da matriz de dados $\boldsymbol{X}_{\mathrm{p}}$. Sem perda de generalidade, a decomposição em valores singulares de $\boldsymbol{X}_{\mathrm{p}}$ é dada por:

$$\boldsymbol{X}_{\mathrm{p}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}', \tag{3.13}$$

em que $\boldsymbol{U}_{n\times p}$ e $\boldsymbol{V}_{p\times p}$ são matrizes com colunas ortonormais e $\boldsymbol{D}_{p\times p}$ é uma matriz diagonal com elementos nulos ou positivos, chamados de valores singulares.

Segue então que $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$ é a matriz cujas colunas são os componentes principais e as colunas de $\boldsymbol{V}$ são os vetores de *loadings* correspondentes de cada um dos componentes. Logo, para cada $i = 1,2,\dots,p$, $\boldsymbol{Z}_i = \boldsymbol{U}_i D_{ii} = \lambda_i \boldsymbol{u}_i$ denota o i-ésimo componente principal. Cabe ressaltar que, de forma semelhante ao que ocorre na teoria da regressão linear, cada pseudodado $\boldsymbol{z}_i$ utilizado foi também centralizado. Uma vez que os componentes principais são obtidos à partir da matriz $\boldsymbol{X}_{\mathrm{p}}$, a média de cada componente é zero.

3.5 Porcentagem de variância explicada dos componentes obtidos pelos métodos SPCA, SGPCA e PACSPCA

Se $\hat{\boldsymbol{Z}} = \left\{\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2, \dots, \hat{\boldsymbol{Z}}_k\right\}$ são os componentes principais aproximados, obtidos pelos métodos SPCA, SGPCA e PACSPCA, certamente eles serão correlacionados. Dessa forma, a variância total não deve mais ser estimada pelo traço de $\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}$. A ideia para se calcular a vari-

ância explicada pelos k primeiros componentes modificados $\left\{\hat{\boldsymbol{Z}}_i,\, i=1,\ldots,k\right\}$ é norteada pela seguinte construção: obtidos $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$, projeta-se ortogonalmente $\hat{\boldsymbol{Z}}_2$ em $\hat{\boldsymbol{Z}}_1$, tomando-se a diferença $\hat{\boldsymbol{Z}}_{2|1} = \hat{\boldsymbol{Z}}_2 - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_2\right)$. Essa é uma forma de se eliminar a dependência linear entre $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$. Obtido $\hat{\boldsymbol{Z}}_3$, projeta-se $\hat{\boldsymbol{Z}}_3$ ortogonalmente no subespaço gerado por $\hat{\boldsymbol{Z}}_1$ e $\hat{\boldsymbol{Z}}_2$ e toma-se a diferença $\hat{\boldsymbol{Z}}_{3|1,2} = \hat{\boldsymbol{Z}}_3 - P_{\hat{\boldsymbol{Z}}_1|\hat{\boldsymbol{Z}}_2}\left(\hat{\boldsymbol{Z}}_3\right)$. Assim sucessivamente obtém-se $\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1}$:

$$\begin{aligned}
&\hat{\boldsymbol{Z}}_1 \\
&\hat{\boldsymbol{Z}}_{2|1} = \hat{\boldsymbol{Z}}_2 - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_2\right) \\
&\hat{\boldsymbol{Z}}_{3|1,2} = \hat{\boldsymbol{Z}}_3 - P_{\hat{\boldsymbol{Z}}_2}\left(\hat{\boldsymbol{Z}}_3\right) - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_3\right) \\
&\vdots \\
&\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1} = \hat{\boldsymbol{Z}}_k - P_{\hat{\boldsymbol{Z}}_{k-1}}\left(\hat{\boldsymbol{Z}}_k\right) - \ldots - P_{\hat{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{Z}}_k\right)
\end{aligned} \quad . \tag{3.14}$$

A variância explicada pelo componente principal $\hat{\boldsymbol{Z}}_k$ é dada pelo quadrado da norma de sua projeção ortogonal nos demais $k-1$ componentes, $\left\|\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1}\right\|^2$ e a variância total explicada até o k-ésimo componente principal é dada por $\sum_{j=1}^{k}\left\|\hat{\boldsymbol{Z}}_{j|1,\ldots,j-1}\right\|^2$.

Com o propósito de identificar diferenças quanto ao desempenho dos métodos foi utilizada a noção de porcentagem de variância explicada, que é obtida pela divisão da variância explicada pelo componente principal pela variância total explicada pelo PCA padrão. Nesse sentido, a proporção da variância (em porcentagem) explicada pelo k-ésimo componente principal obtido pelos métodos SPCA, SGPCA ou PACSPCA é dada por:

$$\frac{\left\|\hat{\boldsymbol{Z}}_{k|1,2,\ldots,k-1}\right\|^2}{\operatorname{tr}\left[(\boldsymbol{UD})'\boldsymbol{UD}\right]} \times 100, \tag{3.15}$$

em que $\operatorname{tr}\left[(\boldsymbol{UD})'\boldsymbol{UD}\right]$ é a variância total explicada pelos componentes do método PCA padrão.

3.6 Recursos computacionais

3.6.1 *Softwares* e pacotes

Todas as análises estatísticas foram realizadas utilizando-se o *software MatLab* e o *software R* 4.0.2, conforme *R Development Core Team* (2020). Os pacotes *MASS* (VENABLES; RIPLEY, 2002) e *elasticnet* (ZOU; HASTIE, 2018) da biblioteca do *R* foram utilizados para a simulação dos dados e para a obtenção dos componentes principais via método SPCA, respectivamente. Para obtenção dos componentes principais esparsos foi utilizada a função *spca* do pacote *elasticnet*. Um primeiro argumento interessante dessa função permite que se determine a quantidade de componentes a serem fornecidos. Um segundo argumento importante é

o argumento `sparse="varnum"`, que especifica que desejamos impor a propriedade de esparsidade em cada um dos vetores de *loadings* dos componentes a serem gerados, sendo nesse caso necessário a indicação do número de *loadings* (coeficientes) que serão diferentes de zero. Isso é feito com o argumento "`para=c(·)`".

Para a geração dos componentes via método OSCAR (BONDELL; REICH, 2008), denominado no presente trabalho de SGPCA, foi utilizado um *script* desenvolvido pelos autores para o *software MatLab*. Para a obtenção dos vetores de *loadings* dos componentes via método PACS (SHARMA; BONDELL; ZHANG, 2013), denominado no presente trabalho de PACSPCA, utilizou-se um *script R* também desenvolvido pelos autores. Ambos os *scripts* dos métodos OSCAR e PACS encontram-se disponíveis na página do professor Howard Bondell.

Sharma, Bondell e Zhang (2013) sugerem que para o argumento `betawt` da função PACS, que sejam atribuídos os coeficientes obtidos a partir de um modelo de regressão *Ridge*. Em estudos com preditores colineares, os autores observaram que a utilização de estimativas *Ridge* para os pesos adaptativos apresentou uma performance melhor em comparação a aquelas com estimativas de mínimos quadrados. Nesse sentindo, utilizou-se o pacote *glmnet* (FRIEDAN et. al, 2020) para a obtenção das estimativas *Ridge*, que foram empregadas posteriormente na função PACS para a obtenção dos vetores de *loadings* dos componentes principais para o método PACSPCA. Por sua vez, na função OSCAR exige-se a especificação dos parâmetros α e γ. Nesse sentido, utilizou-se uma grade de valores para esses parâmetros, que é apresentada a seguir:

$$\alpha = (0; .01; .05; .1; .25; .5; .75; .9; 1), \tag{3.16}$$

$$\gamma = (.003; .004; .005; .0075; .01; .025; .05; .1; .15; .2; .3; .4; .5; .6). \tag{3.17}$$

Para garantir que os resultados obtidos por todos os métodos possam ser reproduzíveis, utilizou-se a função `set.seed` do *R*. Essa função permite, por meio de uma "semente", que a geração de uma mesma sequência de números pseudo-aleatórios possa ser obtida por diferentes usuários. Dessa maneira, para garantir a reprodutibilidade dos resultados obtidos no presente trabalho, adotou-se a semente `set.seed`(1). Em razão dos resultados do método SGPCA serem calculados no *software MatLab* e com o objetivo de manter todos os métodos em condições semelhantes de análise, utilizou-se as mesmas matrizes padronizadas $\boldsymbol{X}_{\mathrm{p}}$ obtidas no *software R* para o método SGPCA. Nesse sentido, as matrizes $\boldsymbol{X}_{\mathrm{p}}$ para os cenários 1 e 2 foram primeiramente obtidas no *R* e depois empregadas no *MatLab* para a aplicação do método SGPCA.

4 RESULTADOS E DISCUSSÃO

4.1 Cenário 1

Na Tabela 4.1 são apresentadas as correlações amostrais de Pearson observadas entre as variáveis $(X_1, X_2, \ldots, X_{10})$. Uma vez que as variáveis (X_1, X_2, X_3, X_4) apresentam correlações muito fortes entre si e correlações desprezíveis com as demais variáveis, segue que essas variáveis constituem o primeiro grupo (Grupo I). Como as variáveis (X_5, X_6, X_7, X_8) foram formadas a partir de V_2, seria natural pensar que essas variáveis formassem um segundo grupo. Contudo, isso não ocorre na prática.

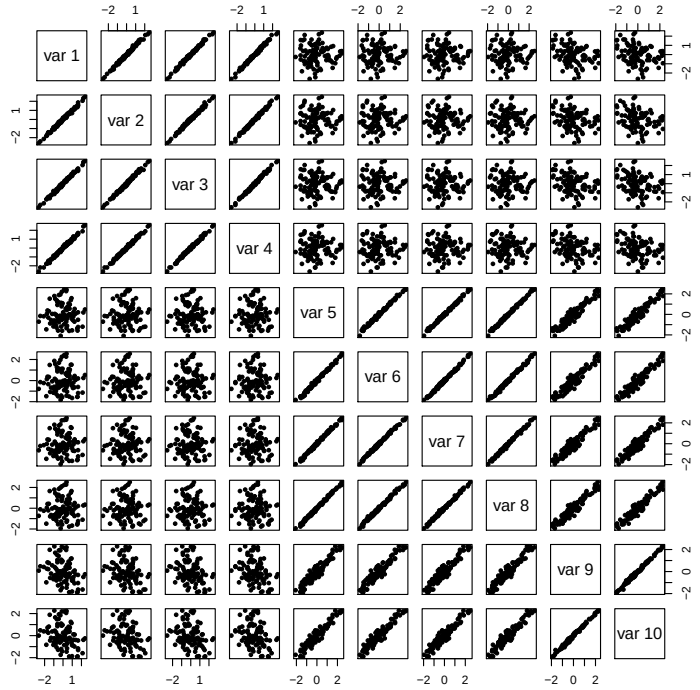
Tabela 4.1 – Correlações amostrais de Pearson observadas entre as variáveis $(X_1, X_2, \ldots, X_{10})$, obtidas a partir dos fatores V_1, V_2 e V_3.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,000	0,995	0,995	0,995	-0,002	0,001	0,001	-0,013	-0,293	-0,287
X_2	0,995	1,000	0,995	0,995	0,012	0,014	0,014	0,001	-0,280	-0,273
X_3	0,995	0,995	1,000	0,996	-0,001	0,003	0,003	-0,011	-0,292	-0,285
X_4	0,995	0,995	0,996	1,000	-0,004	-0,003	-0,002	-0,016	-0,296	-0,289
X_5	-0,002	0,012	-0,001	-0,004	1,000	0,996	0,996	0,996	0,950	0,951
X_6	0,001	0,014	0,003	-0,003	0,995	1,000	0,997	0,995	0,949	0,951
X_7	0,001	0,014	0,003	-0,002	0,996	0,997	1,000	0,996	0,949	0,951
X_8	-0,013	0,001	-0,011	-0,016	0,996	0,995	0,996	1,000	0,954	0,955
X_9	-0,293	-0,280	-0,292	-0,296	0,950	0,949	0,949	0,954	1,000	0,996
X_{10}	-0,287	-0,273	-0,285	-0,289	0,951	0,951	0,951	0,955	0,996	1,000

Os fatores V_1 e V_2 possuem maior relevância em relação a V_3, em razão do fato de gerarem um número maior de variáveis. Além disso, o fator V_3 recebe um maior peso do fator V_2 em sua formação, quando comparado ao fator V_1 $(V_3 = -0,3V_1 + 0,925V_2 + \varepsilon)$. Uma vez que (X_9, X_{10}) possuem como fator latente a variável V_3, isso naturalmente implicou que as variáveis (X_5, X_6, X_7, X_8) apresentassem altas correlações com (X_9, X_{10}). Dessa maneira, espera-se que as variáveis $(X_5, X_6, X_7, X_8, X_9, X_{10})$ formem o segundo grupo (Grupo II). A forma como as variáveis dos dois grupos se correlacionam dentro e entre os grupos formados pode ser melhor compreendida com base nos gráficos de dispersão entre as variáveis (FIGURA 4.1).

Na Tabela 4.2 são apresentados os vetores *loadings* dos componentes principais, obtidos a partir dos métodos PCA padrão, SPCA, SGPCA e PACSPCA. Uma primeira decisão que deve ser levada em consideração diz respeito ao número de componentes principais a serem retidos. Adotando como critério a porcentagem acumulada da explicação da variância total, podem ser retidos os k $(k < p)$ primeiros componentes que expliquem entre 70% e 90% da

Figura 4.1 – Gráficos de dispersão das variáveis $(X_1, X_2, \ldots, X_{10})$.



Fonte: Do autor (2020).

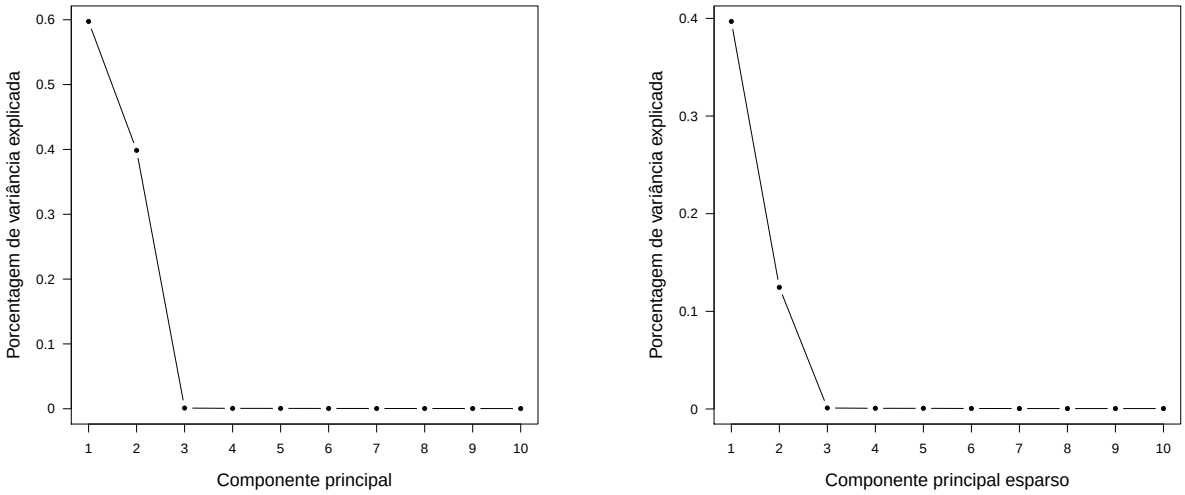
variância total (JOLLIFFE, 2002, p. 113). Porém, essa porcentagem pode ser maior ou menor dependendo dos detalhes práticos de um determinado conjunto de dados. Segundo Jolliffe (2002), um valor maior que 90% será apropriado quando um ou dois componentes principais representarem fontes de variação muito dominantes e óbvios. Esse fato ocorre nesse primeiro cenário, em razão dos fatores V_1 e V_2.

Considerando o método PCA padrão para a obtenção dos componentes principais, pode-se observar que os dois primeiros componentes explicaram juntos 99,60% da variância total (TABELA 4.2). Além de utilizar-se como critério para a retenção dos k primeiros componentes a porcentagem cumulativa da explicação da variância total, pode-se utilizar também um gráfico denominado *scree plot*, em que é plotado no eixo das abcissas a ordem j de cada um dos componentes principais e no eixo das ordenadas os seus respectivos autovalores ou a porcentagem de variância explicada de cada componente. Pode-se observar na Figura 4.2*a* que o acréscimo do terceiro componente não provoca alterações expressivas na porcentagem de variância explicada, pois a partir desse componente o coeficiente angular no gráfico é aproximadamente nulo.

Ainda em relação aos componentes obtidos pelo método PCA padrão, pode-se verificar que esse método não conseguiu identificar os dois grupos que foram construídos. Esse fato

Figura 4.2 – Gráficos *scree plot* dos 10 componentes principais obtidos a partir dos métodos PCA padrão e *Sparse Principal Component Analysis* (SPCA), considerando o tamanho amostral $n = 100$.

(a) Método PCA padrão. (b) Método SPCA.



Fonte: Do autor (2020).

ocorreria se as variáveis com correlações muito fortes entre si apresentassem um único valor de *loading*. Todavia, no PC1 esse método forneceu uma boa aproximação nos *loadings* das variáveis X_1 (0,119), X_2 (0,114), X_3 (0,118) e X_4 (0,120) (TABELA 4.2). Em relação a identificação do segundo grupo, o desempenho não foi tão eficiente. Nos dois componentes, o método PCA padrão conseguiu uma boa aproximação nos *loadings* das variáveis (X_5, X_6, X_7, X_8), diferindo basicamente na terceira casa decimal. Porém, tanto para o PC1 quanto para o PC2, o método forneceu valores diferentes para as variáveis (X_9, X_{10}) em comparação aos valores dos *loadings* de (X_5, X_6, X_7, X_8), não levando em consideração a alta correlação presente entre essas variáveis.

Na Figura 4.2b é apresentado o gráfico *scree plot* dos 10 componentes principais obtidos utilizando-se o método SPCA. Uma vez que a porcentagem de variância explicada do terceiro ao décimo componente principal esparso não é relevante, considerou-se somente os dois primeiros componentes. Zou, Hastie e Tibshirani (2006) sugeriram que é necessário considerar-se apenas os dois primeiros componentes com representações esparsas "corretas", devido à importância dos fatores V_1 e V_2 e, pelo fato dos dois primeiros componentes exatos explicarem 99,60% da variabilidade total no PCA padrão. Conforme os autores, o primeiro componente derivado deveria recuperar o fator V_2 usando apenas (X_5, X_6, X_7, X_8), enquanto o segundo componente deveria recuperar o fator V_1 usando apenas (X_1, X_2, X_3, X_4). Nesse sentido, no presente trabalho foram então gerados 10 componentes esparsos e para isso foi utilizado o argumento

`sparse="varnum"` da função *spca* do pacote *elasticnet*, impondo que a esparsidade em cada um dos componentes deve apresentar como número de *loadings* não nulos um número igual a 4 (`para=c(4, ..., 4)`).

Zou, Hastie e Tibshirani (2006) verificaram para o PC1 esparso que para cada uma das variáveis (X_5, X_6, X_7, X_8) foi obtido o mesmo valor de *loading* (0,5), enquanto as demais variáveis apresentaram *loadings* nulos. Por sua vez, para o PC2 esparso foi obtido o valor de *loading* de 0,5 para cada uma das variáveis (X_1, X_2, X_3, X_4) e *loadings* nulos para as variáveis restantes nesse componente. Ainda nesse artigo foi possível observar que os dois primeiros componentes esparsos apresentaram variância acumulada de $80,40\%$, sendo que os componentes esparsos PC1 e PC2 retiveram, respectivamente, $40,90\%$ e $39,50\%$ da variância total.

Na Tabela 4.2 também são apresentados os resultados dos componentes principais utilizando o método SPCA. Apesar de se tratar da análise do mesmo exemplo sintético, pode-se verificar que os resultados apresentados nesse trabalho diferem dos resultados fornecidos por Zou, Hastie e Tibshirani (2006). A primeira diferença diz respeito a ordem dos vetores de *loadings* dos componentes. Ao se calcular os vetores de *loadings* esparsos utilizando a função *spca* do *R* verificou-se que os mesmos não preservaram a ordem em relação a porcentagem de variância explicada para os componentes formados. Para exemplificar, vamos nos restringir aos dois primeiros vetores de *loadings*. O primeiro vetor de *loadings*, obtido de $\boldsymbol{y} = \boldsymbol{z}_1$, forneceu um componente esparso com menor variância explicada (12,50%) quando comparado ao componente esparso (39,70%) gerado a partir do segundo vetor de *loadings*, obtido de $\boldsymbol{y} = \boldsymbol{z}_2$. Os vetores $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$ são os dois primeiros componentes principais do método PCA padrão, que foram utilizados como pseudodados.

Diante disso, os vetores de *loadings* foram apresentados na Tabela 4.2 levando em consideração os valores de variância explicada dos componentes principais esparsos gerados a partir deles, do maior para o menor. O vetor de *loadings* do componente PC1 esparso foi obtido a partir de $\boldsymbol{y} = \boldsymbol{z}_2$ e esse componente apresentou variância explicada de $39,70\%$. Por sua vez, o vetor de *loadings* do componente PC2 esparso foi obtido a partir de $\boldsymbol{y} = \boldsymbol{z}_1$ e esse componente apresentou variância explicada de $12,50\%$. Ainda que a ordem dos vetores de *loadings* não tenha sido preservada em relação aos resultados do artigo de Zou, Hastie e Tibshirani (2006), algumas observações podem ser feitas. Claramente, o método SPCA conseguiu identificar corretamente os conjuntos de variáveis importantes nos dois componentes principais gerados. Pode-se observar ainda que os 4 *loadings* diferentes de zero no PC1 correspondem as variáveis relacionadas

ao fator V_1, enquanto que os 4 *loadings* diferentes de zero para o PC2 correspondem as variáveis cujo fator subjacente é V_2. Com base nesses resultados, nota-se que o fato de V_2 possuir uma variância maior quando comparado ao fator V_1 não é refletida no *ranking* entre os dois componentes esparsos. Essa afirmação é fornecida porque o PC1, que apresenta maior variância (%), é dominado por variáveis geradas a partir de V_1.

A segunda diferença está relacionada aos valores de *loadings* para os dois componentes esparsos e em relação as variâncias explicadas por cada um deles. Os resultados apresentados na Tabela 4.2 diferiram dos resultados apresentados por Zou, Hastie e Tibshirani (2006). Por exemplo, os dois componentes esparsos no presente trabalho retiveram juntos apenas 52,20% da variância total, enquanto que em Zou, Hastie e Tibshirani (2006) esse valor foi de 80,40%, conforme mencionado anteriormente. Naturalmente, uma pergunta que surge é: o que faz com que os resultados nos dois trabalhos sejam tão distintos? A resposta a ser dada para essa questão não é definitiva, mas apenas sugere algumas possibilidades.

A primeira causa que justificaria as diferenças é que para gerar os componentes esparsos os autores utilizam a matriz de covariâncias exata ou populacional das variáveis $(X_1, X_2, \ldots, X_{10})$, enquanto que no presente trabalho foi utilizada a matriz de correlações amostrais das mesmas. Outra justificativa mais plausível é em relação a especificação dos argumentos da função *spca* no *R*. Pode ser visto que especificações diferentes nos argumentos da função *spca* podem gerar diferentes vetores de *loadings* para um mesmo conjunto de dados. Esse fato é exemplificado no Apêndice e o leitor é convidado a observar o que pode ocorrer aos resultados para pequenas mudanças nos valores dos argumentos dessa função. Por último, não é claro que a ordem dos vetores de *loadings* esparsos e, por consequência, a dos componentes modificados com esparsidade pelo método SPCA deve preservar a mesma ordem dos componente principais $(\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_p)$ do método PCA padrão, que foram utilizados como pseudodados. Pelo fato dos vetores de *loadings* esparsos serem aproximações aos vetores de *loadings* do método PCA padrão, é possível que a ordem não seja mantida. Nesses casos, os componentes esparsos devem ser ordenados conforme os valores de variância explicada por cada um deles, do maior para o menor.

Em síntese, os resultados obtidos utilizando-se o método SPCA mostram que esse método foi eficiente no tocante à identificação das variáveis importantes. Porém, é importante destacar que no tocante a identificação dos grupos definidos com base na alta correlação entre as variáveis, o método SPCA não se mostrou tão eficaz. Pode-se verificar que no segundo componente esparso são fornecidos *loadings* nulos para (X_9, X_{10}), apesar de ambas variáveis

Tabela 4.2 – Vetores de *loadings* para a formação dos componentes principais utilizando os métodos PCA, SPCA, SGPCA e PACSPCA.

Variável	PCA		SPCA		SGPCA		PACSPCA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
X_1	0,119	-0,478	-0,496	0	0,057	-0,495	0,141	-0,500
X_2	0,114	-0,480	-0,473	0	0,057	-0,495	0,141	-0,500
X_3	0,118	-0,479	-0,553	0	0,057	-0,495	0,141	-0,500
X_4	0,120	-0,478	-0,474	0	0,057	-0,495	0,141	-0,500
X_5	-0,391	-0,145	0	-0,007	-0,406	-0,067	-0,392	0
X_6	-0,390	-0,146	0	-0,078	-0,406	-0,067	-0,392	0
X_7	-0,390	-0,146	0	-0,035	-0,406	-0,067	-0,392	0
X_8	-0,392	-0,140	0	-0,996	-0,406	-0,067	-0,392	0
X_9	-0,408	0,002	0	0	-0,406	0	-0,392	0
X_{10}	-0,408	-0,002	0	0	-0,406	0	-0,392	0
Variância (%)	59,74	39,86	39,70	12,50	59,42	39,74	59,66	35,71
Variância acumulada (%)	59,74	99,60	39,70	52,20	59,42	99,16	59,66	95,37

apresentarem correlações superiores a 94% em relação as variáveis (X_5, X_6, X_7, X_8) (TABELA 4.1).

Uma questão pertinente relacionada a esses resultados pode ser abordada. No exemplo simulado foi possível determinar, com certa tranquilidade, o nível de esparsidade para cada componente. Todavia, determinar o nível de esparsidade em dados reais torna-se um desafio muito grande, pois geralmente desconhecemos os "fatores ocultos" que podem afetar as variáveis que estejam em análise. Em um cenário de uma análise com dados reais, como desprezar variáveis (X_9, X_{10}) que sejam porventura altamente correlacionadas com outras variáveis? A fim de não desprezar a alta correlação que possa existir entre as variáveis sob investigação, os métodos propostos podem ser utilizados para auxiliar em situações como essa, pois são procedimentos de penalização que permitem a seleção de variáveis importantes (esparsidade) enquanto agrupam variáveis correlacionadas.

Na Tabela 4.2 são apresentadas as soluções do método SGPCA para cada componente. Como foi empregada uma grade de valores para os parâmetros α e γ, optou-se por utilizar $\alpha = 0,25$ e $\gamma = 0,60$. A escolha foi subjetiva e baseou-se unicamente nos resultados fornecidos, por apresentarem resultados mais interessantes quando comparado aos demais. Inicialmente, pode-se observar em relação ao componente de maior relevância, PC1, que o método proposto SGPCA mostrou-se eficiente ao identificar corretamente os dois grupos de variáveis, atribuindo a cada uma das variáveis $(X_1, \ldots, X_4)$ o valor de *loading* igual a $0,057$ e para as va-

riáveis $(X_5,\ldots,X_{10})$ o valor de *loading* igual a $-0,406$. Por sua vez, o método SGPCA também conseguiu capturar no componente PC2 a mesma informação fornecida nos dois componentes esparsos do método SPCA. O resultado obtido no segundo componente pode ser interpretado como a identificação das variáveis associadas aos fatores de maior importância, V_1 e V_2. Nesse segundo componente foram obtidos *loadings* iguais e não nulos para as variáveis $(X_1,\ldots,X_4)$ $(-0,495)$ e $(X_5,\ldots,X_8)$ $(-0,067)$, enquanto foram fornecidos *loadings* nulos para (X_9,X_{10}) (esparsidade).

Na Tabela 4.2 também são apresentados os resultados do método PACSPCA. Os pesos dos coeficientes no método PACS podem ser atribuídos de algumas maneiras diferentes. Uma vez que a correlação desempenha um papel fundamental na forma como os grupos são formados, utilizou-se como pesos a correlação entre as variáveis. Essa abordagem é denominada de PACS adaptativo, por incorporar correlações nos pesos (*AdCorr* PACS). Pode ser observado na Tabela 4.2 que o método PACSPCA proposto fornece resultados análogos ao método SGPCA no tocante ao primeiro componente, por identificar corretamente os dois grupos de variáveis. Em relação ao PC1, o método PACSPCA forneceu para $(X_1,\ldots,X_4)$ o valor de *loading* igual a $0,141$ e para $(X_5,\ldots,X_{10})$ o valor foi de $-0,392$. O resultado obtido para o PC2 é semelhante ao obtido pelo método SPCA no que se refere ao segundo componente, no sentido de fornecer *loadings* não nulos para as variáveis $(X_1,\ldots,X_4)$ e por fornecer solução esparsa para as demais variáveis.

Em termos da variância acumulada (em porcentagem), nesse cenário pode-se verificar que o melhor desempenho foi obtido pelo SGPCA (99,16%), seguido do PACSPCA (95,36%). A pior performance foi verificada para o SPCA (52,20%). Esse desempenho do método SPCA, em comparação ao bom desempenho dos métodos SGPCA e PACSPCA, pode ser justificado em razão do grau de esparsidade imposta aos componentes. Neste caso, quanto maior o número de *loadings* nulos menor a variância explicada por cada um dos componentes. Esse fato reforça a potencialidade dos métodos propostos, pois ambos preservam a propriedade de esparsidade enquanto agrupam variáveis correlacionadas. Isso se reflete numa maior porcentagem de variância explicada para os componentes principais obtidos por esses métodos. É importante destacar também que, uma vez que os métodos SPCA, SGPCA e PACSPCA fornecem vetores de *loadings* aproximados, nem sempre o primeiro componente principal possuirá a maior porcentagem de variância explicada.

Uma consideração mais geral pode ser feita acerca dos resultados apresentados nesse cenário. Os métodos SGPCA e PACSPCA podem ser empregados como métodos de seleção de variáveis nos componentes e as soluções esparsas (*loadings* nulos) são os resultados mais naturais nesse sentido. Contudo, as soluções de agrupamento também podem ser empregadas para essa finalidade. Pôde-se verificar que os dois métodos formaram dois grupos de variáveis, sendo que as correlações das variáveis dentro dos grupos foram em ambos os casos muito fortes. Em situações como essa, em cada grupo constituído pode-se tomar um representante de cada conjunto de variáveis. No presente exemplo, os PC1's obtidos pelos métodos SGPCA e PACSPCA poderiam ser redefinidos como a combinação linear de apenas duas variáveis, uma representante das variáveis $(X_1,\ldots,X_4)$ e outra representante das variáveis $(X_5,\ldots,X_{10})$. Do ponto de vista meramente estatístico qualquer variável dentro de um grupo poderia ser escolhida como representante, uma vez que as variáveis não se distinguem em razão da forte estrutura das correlações observadas. Todavia, pode existir o interesse prático para que uma ou até mais variáveis sejam preservadas em cada grupo no componente principal.

4.2 Cenário 2

Na Tabela 4.3 são apresentadas as correlações amostrais de Pearson observadas entre as variáveis $(X_1,X_2,\ldots,X_{12})$. De forma semelhante ao que ocorreu no cenário 1, algumas variáveis apresentaram correlações muito fortes entre si e correlações desprezíveis, moderadas ou fortes com as demais variáveis. Isso pode ser verificado principalmente entre as variáveis (X_1,X_2,X_3), (X_4,X_5,X_6) e (X_7,X_8). Naturalmente, esses conjuntos de variáveis constituem três grupos que devem ser considerados, pela forma como foram obtidos. Seguindo o raciocínio apresentado para o exemplo do cenário 1, no presente cenário também pode-se considerar que os fatores V_1 e V_2 são igualmente importantes e mais relevantes em detrimento aos fatores V_3, V_4 e V_5.

As variáveis (X_9,X_{10}) foram formadas a partir do fator V_4, enquanto as variáveis (X_{11},X_{12}) foram formadas a partir do fator V_5. Esses dois fatores possuem uma particularidade em relação a forma como foram constituídos, sendo $V_4 = 0,7V_1 - 0,8V_2 + \varepsilon$ e $V_5 = -0,6V_1 + 0,55V_2 + \varepsilon$. Essa particularidade decorre do fato de que, tanto para o fator V_4 quanto para o fator V_5, não se verifica a predominância de um dos fatores V_1 ou V_2, em razão dos pesos atribuídos a cada um desses fatores. Nesse caso, não é possível definir de antemão se cada um dos conjuntos de variáveis (X_9,X_{10}) e (X_{11},X_{12}) formarão novos *clusters* ou se serão agrupados ao grupo I (X_1,X_2,X_3) ou ao grupo II (X_4,X_5,X_6). Essa dúvida surge primeiramente do fato de V_1 ser o

fator subjacente das variáveis do grupo I e de V_2 ser o fator subjacente das variáveis do grupo II. Além disso, esses dois fatores também são utilizados de forma indireta na formação das variáveis (X_9, X_{10}) e (X_{11}, X_{12}) por meio de V_4 e V_5, respectivamente.

Tabela 4.3 – Correlações amostrais de Pearson observadas entre as variáveis $(X_1, X_2, \ldots, X_{12})$, obtidas a partir dos fatores V_1, V_2, V_3, V_4 e V_5.

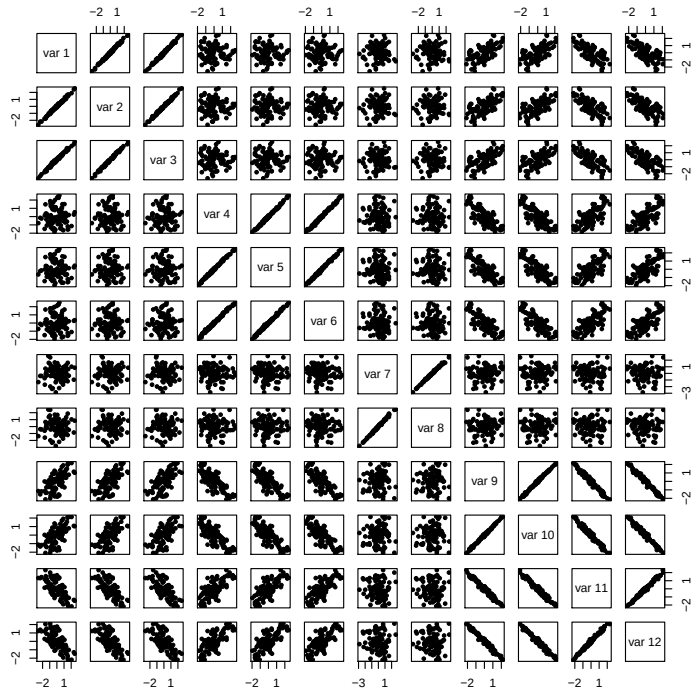
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1,000	0,996	0,995	0,003	0,003	-0,011	0,012	0,015	0,619	0,618	-0,693	-0,682
X_2	0,996	1,000	0,995	-0,003	-0,002	-0,016	0,012	0,016	0,623	0,622	-0,696	-0,684
X_3	0,995	0,995	1,000	0,008	0,010	-0,004	0,004	0,007	0,615	0,615	-0,686	-0,677
X_4	0,003	-0,003	0,008	1,000	0,997	0,995	-0,051	-0,041	-0,777	-0,776	0,708	0,719
X_5	0,003	-0,002	0,010	0,997	1,000	0,996	-0,062	-0,052	-0,777	-0,776	0,707	0,718
X_6	-0,011	-0,016	-0,004	0,995	0,996	1,000	-0,053	-0,043	-0,786	-0,785	0,717	0,727
X_7	0,012	0,012	0,004	-0,051	-0,062	-0,053	1,000	0,997	0,053	0,044	-0,055	-0,037
X_8	0,015	0,016	0,007	-0,041	-0,052	-0,043	0,997	1,000	0,046	0,038	-0,051	-0,032
X_9	0,619	0,623	0,615	-0,777	-0,777	-0,786	0,052	0,046	1,000	0,997	-0,984	-0,986
X_{10}	0,618	0,622	0,615	-0,776	-0,776	-0,785	0,043	0,038	0,997	1,000	-0,982	-0,985
X_{11}	-0,693	-0,696	-0,686	0,708	0,707	0,717	-0,055	-0,051	-0,984	-0,982	1,000	0,993
X_{12}	-0,682	-0,684	-0,677	0,719	0,718	0,727	-0,037	-0,032	-0,986	-0,985	0,993	1,000

Com base nas correlações apresentadas na Tabela 4.3, pode-se verificar que as variáveis $(X_9, \ldots, X_{12})$ se correlacionam de forma moderada ou forte com as variáveis $(X_1, \ldots, X_6)$. As variáveis (X_9, X_{10}) apresentam correlações moderadas e positivas $(0,615\,;\,0,623)$ com as variáveis (X_1, X_2, X_3) e correlações fortes e negativas $(-0,776\,;\,-0,786)$ com as variáveis (X_4, X_5, X_6). Por sua vez, as variáveis (X_{11}, X_{12}) apresentam correlações moderadas e negativas $(-0,677\,;\,-0,696)$ com as variáveis (X_1, X_2, X_3) e correlações fortes e positivas $(0,707\,;\,0,727)$ com as variáveis (X_4, X_5, X_6).

Novamente, a questão relacionada às variáveis (X_9, X_{10}) e (X_{11}, X_{12}) pode ser suscitada, levando em consideração não somente a forma como essas variáveis originaram-se, mas também mediante a intensidade das correlações que essas apresentaram com as variáveis dos grupos I e II. Este tipo de situação também é importante para se analisar o comportamento dos métodos propostos SGPCA e PACSPCA, num cenário em que algumas variáveis correlacionam-se de forma moderada ou forte com outras variáveis. Na Figura 4.3 são apresentados os gráficos de dispersão das variáveis $(X_1, \ldots, X_{12})$. Podemos observar mais claramente na Figura 4.3 como as variáveis dos grupos I, II e III se correlacionam entre si, dentro e entre os grupos formados e com as variáveis $(X_9, \ldots, X_{12})$. Além disso, podemos observar também que as variáveis (X_9, X_{10}) apresentam correlação muito forte e positiva entre si e ambas apresentam correlações

muito fortes e negativas com as variáveis (X_{11}, X_{12}). Essa análise gráfica reforça também a influência dos fatores V_1 e V_2 sobre essas variáveis.

Figura 4.3 – Gráficos de dispersão das variáveis $(X_1, X_2, \ldots, X_{12})$.

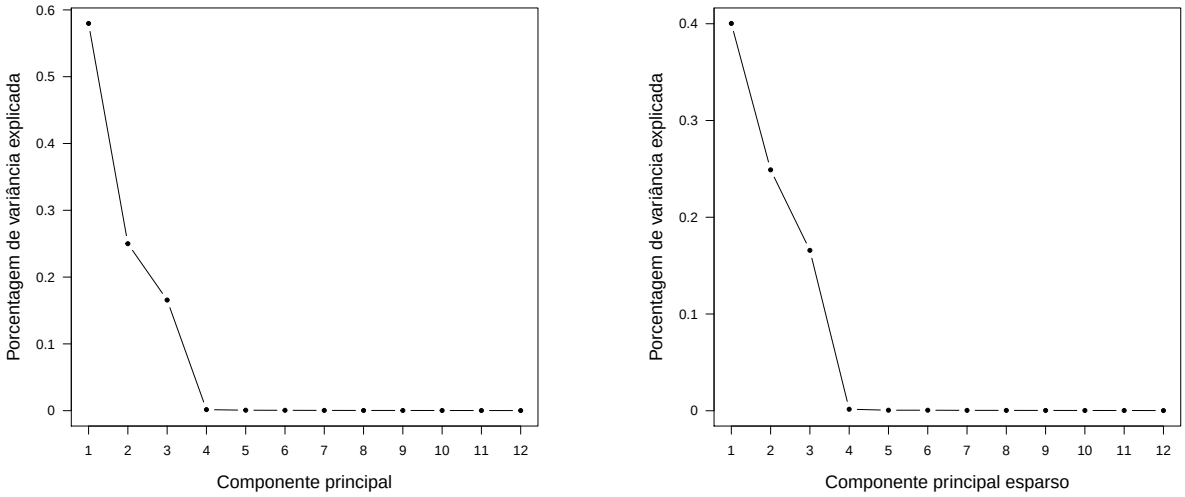


Fonte: Do autor (2020).

Na Figura 4.4a apresenta-se o gráfico *scree plot* para os doze componentes principais obtidos a partir do método PCA padrão. Pode-se verificar que não ocorrem alterações significativas na porcentagem de variância explicada a partir do quarto componente principal, o que pode ser observado pelo fato do coeficiente angular do gráfico ser aproximadamente nulo a partir desse componente. A análise dos gráficos *scree plot* sugere que os três primeiros componentes principais devam ser retidos para explicar a estrutura de variâncias e covariâncias das variáveis $(X_1, X_2, \ldots, X_{12})$. Porém, ao se adotar como critério a porcentagem acumulada da variância total, optou-se por reter somente os dois primeiros componentes, pois juntos esses componentes explicam 82,97% da variabilidade total dos dados (TABELA 4.4). Por sua vez, na Figura 4.4b apresenta-se o gráfico *scree plot* para os doze componentes principais esparsos obtidos a partir do método SPCA. Nesse cenário foi possível se observar que os componentes esparsos preservaram a ordem de magnitude (ou importância) em relação a porcentagem de variância explicada.

Figura 4.4 – Gráficos *scree plot* dos 12 componentes principais obtidos a partir dos métodos PCA padrão e *Sparse Principal Component Analysis* (SPCA), considerando o tamanho amostral $n = 100$.

(a) Método PCA padrão. (b) Método SPCA.



Fonte: Do autor (2020).

Na Tabela 4.4 são apresentados os resultados dos métodos PCA, SPCA, SGPCA e PACSPCA, relacionados à formação dos componentes principais. Primeiramente, em relação ao método PCA padrão pode-se verificar que esse método já consegue obter resultados próximos ao ideal em relação à identificação dos grupos formados. Tomando-se como base os *loadings* fornecidos para o PC1, por exemplo, nota-se que mesmo esse método já consegue fornecer *loadings* aproximadamente iguais para os grupos I (X_1, X_2, X_3), II (X_4, X_5, X_6) e III (X_7, X_8), diferindo na segunda casa decimal para o grupo I ($0,249; 0,250; 0,247$) e na terceira casa decimal para os grupos II ($-0,284; -0,284; -0,288$) e III ($0,025; 0,023$). Pode-se verificar também que o método PCA padrão forneceu para as variáveis (X_9, X_{10}) (fator subjacente V_4) *loadings* aproximadamente iguais ($0,378; 0,377$), enquanto para as variáveis (X_{11}, X_{12}) (fator subjacente V_5) os *loadings* obtidos foram iguais ($-0,377$). Resultados semelhantes podem ser observados para o PC2 obtido por esse método.

Uma vez que foram formados três grupos bem definidos nesse cenário, abrangendo as 8 primeiras variáveis $(X_1, \ldots, X_8)$, em ambos os componentes gerados pelo método SPCA indicou-se no argumento `sparse="varnum"` que o número de *loadings* diferentes de zero em cada um desses componentes seria 8 (`para=c` (8, ..., 8)). Dessa forma, foram gerados 12 componentes a partir do método SPCA e os resultados dos dois primeiros encontram-se na Tabela 4.4. Pode-se verificar em relação o PC1, que o método SPCA não conseguiu identificar corre-

Tabela 4.4 – Vetores de *loadings* para a formação dos componentes principais utilizando os métodos PCA, SPCA, SGPCA e PACSPCA.

Variável	PCA		SPCA		SGPCA		PACSPCA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
X_1	0,249	-0,434	0	-0,452	0,316	-0,422	0,316	-0,408
X_2	0,250	-0,432	0,137	-0,447	0,316	-0,422	0,316	-0,408
X_3	0,247	-0,437	0	-0,457	0,316	-0,422	0,316	-0,408
X_4	-0,284	-0,379	-0,070	-0,359	-0,316	-0,394	-0,316	-0,408
X_5	-0,284	-0,380	-0,089	-0,363	-0,316	-0,394	-0,316	-0,408
X_6	-0,288	-0,373	-0,049	-0,354	-0,316	-0,394	-0,316	-0,408
X_7	0,025	0,045	0	0,029	0	0	0	0
X_8	0,023	0,040	0	0,021	0	0	0	0
X_9	0,378	0,026	0,582	0	0,316	0	0,316	0
X_{10}	0,377	0,026	0,476	0	0,316	0	0,316	0
X_{11}	-0,377	0,032	-0,445	0	-0,316	0	-0,316	0
X_{12}	-0,377	0,024	-0,451	0	-0,316	0	-0,316	0
Variância (%)	57,97	25,00	40,03	24,90	56,19	24,89	56,19	24,87
Variância acumulada (%)	57,97	82,97	40,03	64,93	56,19	81,08	56,20	81,06

tamente os três grupos que foram formados nesse cenário, fornecendo *loadings* nulos para as variáveis (X_1, X_3, X_7, X_8). Essa performance em relação ao agrupamento de variáveis já era esperado, pois a principal característica do método SPCA consiste em fornecer soluções esparsas. Todavia, podemos perceber que o método SPCA, no PC2, consegue identificar as variáveis de maior relevância, fornecendo *loadings* nulos para as variáveis $(X_9, \ldots, X_{12})$.

Na Tabela 4.4 também são apresentados os resultados obtidos pelos métodos SGPCA e PACSPCA. Para o presente cenário, adotou-se como valores para os parâmetros α e λ, os valores $\alpha = 0,25$ e $\gamma = 0,60$. Em relação ao agrupamento das variáveis foi possível observar que ambos os métodos obtiveram resultados iguais em relação ao componente PC1. Pode-se verificar que os métodos SGPCA e PACSPCA forneceram *loadings* iguais para as variáveis $(X_1, X_2, X_3, X_9, X_{10})$, agrupando essas variáveis. Um resultado semelhante foi obtido para as variáveis $(X_4, X_5, X_6, X_{11}, X_{12})$, pela igualdade de seus *loadings* e, portanto, pelo respectivo agrupamento dessas variáveis. Em um cenário de análise com um conjunto de dados reais, um aspecto importante a ser considerado seria investigar esses agrupamentos resultantes, para se compreender o que contribui para que as variáveis agrupadas tenham um comportamento geralmente semelhante.

Devido à forma como cada variável foi formada, sabia-se claramente que (X_1, X_2, X_3) formam o grupo I e que as variáveis (X_4, X_5, X_6) formam o grupo II. Uma questão pertinente

nesse momento seria: qual a justificativa mais plausível para que as variáveis (X_9, X_{10}) serem agrupadas no grupo I e as variáveis (X_{11}, X_{12}) serem agrupadas no grupo II? Como já salientado, no presente cenário não há a predominância de um dos fatores V_1 ou V_2 na formação dos fatores V_4 e V_5. Isso implicou, em primeiro lugar, que não foram observadas correlações muito fortes (superior a 90%) entre as variáveis $(X_9, X_{10}, X_{11}, X_{12})$ com as variáveis dos grupos I ou II (TABELA 4.3). As correlações observadas (positivas ou negativas) entre essas variáveis e cada variável dos grupos I ou II foram moderadas ou fortes. Talvez a justificativa mais plausível para a questão levantada seja em relação ao sentido das correlações entre as variáveis.

Tomando como referência as variáveis (X_1, X_2, X_3) pode-se observar, com base nos resultados da Tabela 4.3 ou em relação aos gráficos de dispersão apresentados na Figura 4.3, que essas variáveis apresentam correlações moderadas e positivas $(0,615\ ;\ 0,623)$ com (X_9, X_{10}) e correlações moderadas e negativas $(-0,677\ ;\ -0,696)$ com (X_{11}, X_{12}). Dessa forma, uma explicação admissível para $(X_1, X_2, X_3, X_9, X_{10})$ formarem um grupo é que os métodos propostos não agrupam somente levando em consideração a força das correlações envolvidas, mas também consideram o sentido das correlações entre as variáveis. Na Tabela 4.3 pode ser observado que (X_1, X_2, X_3) também apresentam correlações moderadas com (X_{11}, X_{12}). Todavia, essas correlações são negativas, ou seja, o sentido das correlações das variáveis do grupo I é o oposto das correlações que essas mesmas variáveis estabelecem com cada uma das variáveis (X_{11}, X_{12}) (FIGURA 4.3).

Ainda em relação ao componente PC1, pode-se observar também que os resultados dos métodos SGPCA e PACSPCA concordam entre si em relação as variáveis (X_7, X_8) (TABELA 4.4). Ambos os métodos agrupam corretamente essas variáveis, porém a solução fornecida para os coeficientes dessas variáveis é esparsa (*loadings* nulos). Em relação aos resultados para o PC2, novamente o SGPCA conseguiu identificar nesse componente as variáveis associadas aos fatores de maior importância, V_1 e V_2. Para cada uma das variáveis (X_1, X_2, X_3) foi atribuído o valor de *loading* $-0,422$ e para cada uma das variáveis (X_4, X_5, X_6) o valor de *loading* foi igual a $-0,394$. Por sua vez, o PACSPCA forneceu o mesmo valor *loading* de $-0,408$ para cada umas das variáveis $(X_1, X_2, X_3, X_4, X_5, X_6)$. Ambos os métodos forneceram *loadings* nulos para as variáveis (X_7, X_8) do componente PC2.

Em termos da variância acumulada (em porcentagem), os métodos SGPCA e PACSPCA apresentaram novamente os melhores desempenhos, com leve superioridade do método SGPCA (81,08%) em relação ao método PACSPCA (81,06%). Novamente, a pior performance foi

verificada para o SPCA (64,9%). Além disso, os resultados obtidos nesse cenário permitiram que se verificasse que os métodos propostos, SGPCA e PACSPCA, não identificam as variáveis a serem agrupadas somente em situações em que existem correlações muito fortes entre as mesmas, mas também agrupam em situações em que são observadas correlações moderadas ou fortes. Além disso, os resultados desse cenário indicam que o sentido das correlações, negativo ou positivo, também desempenha um papel fundamental na forma como as variáveis poderão ou não ser agrupadas.

4.3 Validação dos resultados dos cenários 1 e 2

Na Tabela 4.5 são apresentadas as porcentagens de acerto dos métodos PCA, SPCA, SGPCA e PACSPCA quanto ao agrupamento correto das variáveis nos cenários 1 e 2, considerando $N = 1000$ simulações e $n = 30$, $n = 50$ e $n = 100$ observações. Como era esperado, os métodos PCA e SPCA apresentaram os piores desempenhos. Em nenhuma das situações de avaliação propostas os dois métodos conseguiram agrupar corretamente as variáveis. Por sua vez, os métodos SGPCA e PACSPCA apresentaram porcentagens de acerto superiores a 85% a partir do menor tamanho amostral ($n = 30$). Considerando os dois componentes, o método SGPCA apresentou uma melhor performance quando comparado ao PACSPCA, principalmente quando se considera o PC2 no cenário 2. Quando a comparação se baseia unicamente no PC1, a medida que o tamanho das amostras aumenta de $n = 30$ para $n = 100$ observações os dois métodos passam a ter comportamentos muito semelhantes quanto as porcentagens de acerto.

Tabela 4.5 – Porcentagens de acerto dos métodos PCA, SPCA, SGPCA e PACSPCA quanto ao agrupamento correto das variáveis nos cenários 1 e 2, considerando $N = 1000$ simulações e $n = 30$, $n = 50$ e $n = 100$ observações.

Cenário	n	PCA		SPCA		SGPCA		PACSPCA	
		PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
1	30	0	0	0	0	98,30	99,80	93,00	85,50
	50	0	0	0	0	99,90	99,90	98,30	99,60
	100	0	0	0	0	99,40	99,80	99,90	100,00
2	30	0	0	0	0	92,90	60,80	80,00	0,30
	50	0	0	0	0	97,10	55,30	91,90	1,80
	100	0	0	0	0	99,60	42,20	99,70	8,50

Em cada simulação também foram calculadas as variâncias explicadas por cada método e ao final foram calculadas as variâncias médias explicadas para as $N = 1000$ simulações (TABELA 4.6). Novamente, entre os métodos com componentes principais modificados o pior

desempenho em termos de variância média explicada foi observado para o método SPCA. Todavia, é importante salientar que o grau de esparsidade imposta em cada componente para esse método implica em uma menor ou maior variância explicada em cada situação. Por exemplo, se consideramos novamente $k = 10$ componentes no cenário 1, mas alteramos a esparsidade de 4 para 8 componentes com *loadings* não nulos utilizando o argumento `para` da função *spca* do pacote *elasticnet* (`para=c(8, ..., 8)`), as variâncias médias em $N = 1000$ simulações passam de $0,21$ em todos os valores de n para 0,52 ($n = 30$), 0,54 ($n = 50$) e 0,56 ($n = 100$) no componente PC1 obtido pelo método SPCA. Consequentemente, como as variâncias desses componentes são calculadas com base na norma vetorial dos mesmos (ver 2.8.2), exigir um maior grau de esparsidade (*loadings* nulos) em um componente principal pode melhorar a sua interpretação. Todavia, essa decisão implica necessariamente em uma menor variância explicada.

Essa situação envolvendo o método SPCA é mais um dos casos de *trade-off* que ocorrem com determinada frequência na Estatística. De forma geral, o *trade-off* é um termo da língua inglesa que pode ser traduzido como "perda-e-ganho". O termo refere-se, geralmente, a perder uma qualidade ou aspecto de algo em detrimento de outra qualidade ou aspecto. Nesse sentido, a tomada de decisão exige a compreensão tanto de um aspecto bom quanto do aspecto ruim para uma particular escolha. A escolha entre o grau de esparsidade nos vetores de *loadings* e a porcentagem de variância explicada em cada componente deve assinalar ao analista ou pesquisador a compreensão de que a escolha de uma propriedade para um eventual ganho implica na possível perda de outra propriedade ou característica que também possa ser desejável.

Tabela 4.6 – Variâncias médias explicadas pelos componentes principais 1 (PC1) e 2 (PC2) obtidos pelo métodos PCA, SPCA, SGPCA e PACSPCA, considerando $N = 1000$ simulações e $n = 30$, $n = 50$ e $n = 100$ observações.

Cenário	n	PCA		SPCA		SGPCA		PACSPCA	
		PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
1	30	0,62	0,38	0,37	0,21	0,62	0,37	0,61	0,33
	50	0,61	0,38	0,39	0,21	0,61	0,37	0,60	0,35
	100	0,61	0,39	0,39	0,21	0,61	0,38	0,60	0,37
2	30	0,59	0,26	0,40	0,26	0,57	0,26	0,55	0,25
	50	0,58	0,26	0,39	0,26	0,57	0,26	0,55	0,25
	100	0,58	0,25	0,38	0,25	0,57	0,25	0,56	0,25

Zou, Hastie e Tibshirani (2006) apresentaram uma análise com um conjunto de dados reais em que o método SPCA apresentou melhor performance, em termos de proporção de variância acumulada, quando comparado a outro método modificado de PCA. O conjunto ana-

lisado, também chamado de dados *pitprops*, possui $n = 180$ observações e $p = 13$ variáveis. Esse conjunto de dados foi introduzido primeiramente por Jeffers (1967) e ilustra a dificuldade de interpretação na PCA. Ao considerar os seis primeiros componentes principais esparsos, os autores verificaram que o método SPCA reteve 75,80% da variância total até o sexto componente, enquanto que o método SCoTLASS, proposto por Jolliffe, Trendafilov e Uddin em 2003, reteve 69,30%. Além desse melhor desempenho em relação a proporção de variância acumulada até o sexto componente, os vetores de *loadings* fornecidos pelo método SPCA também apresentaram maior esparsidade.

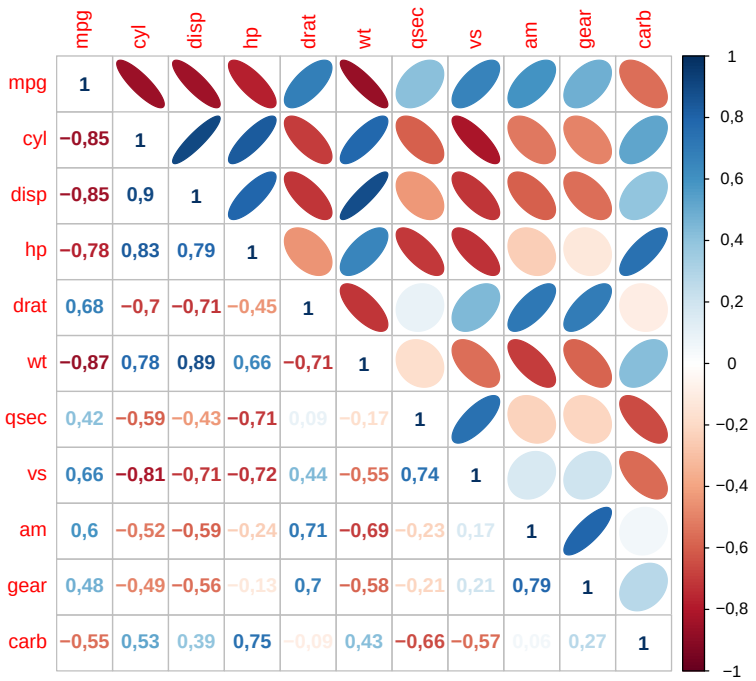
As variâncias médias explicadas pelos métodos SGPCA e PACSPCA também encontram-se na Tabela 4.6. Pode-se observar os resultados são muito semelhantes entre os dois métodos nos dois cenários e para todos os valores de n, com uma ligeira superioridade do método SGPCA que apresentou variâncias médias explicadas um pouco maiores. Diante de tudo o que foi exposto em relação aos resultados das porcentagens de acerto e das variâncias médias explicadas, o método SGPCA parece ser mais indicado que o método PACSPCA para agrupar variáveis em situações em que o tamanho amostral seja pequeno. Com base nos dois critérios avaliados nessa subseção, a medida o número de observações aumenta os dois métodos passam a apresentar resultados similares.

4.4 Exemplo 1

Na Figura 4.5 é apresentada uma representação gráfica da matriz de correlação dos dados *mtcars*. Nessa figura também é apresentada uma barra colorida com uma graduação de -1 a 1, indicando a variação das correlações, de muito fortes e negativas (vermelho escuro) a correlações muito fortes e positivas (azul escuro). Para correlações próximas de zero, a elipse tem uma cor mais clara, tendendo ao branco. De forma geral, pode-se verificar na Figura 4.5 que as variáveis do conjunto de dados *mtcars* apresentam uma estrutura de correlação ampla, desde uma correlação desprezível (0,09) entre as variáveis *drat* e *qsec* a uma correlação muito forte (0,9) entre as variáveis *cyl* e *disp*. Outro ponto a ser destacado é que esse conjunto de dados apresenta variáveis tanto com correlações positivas quanto negativas entre si. Diante do que foi exposto, esse conjunto de dados apresenta um cenário interessante para se avaliar como os métodos SGPCA e PACSPCA se comportam ao fornecerem vetores de *loadings* para dados com essas características.

Com base nos resultados obtidos pelo método PCA padrão, na Tabela 4.7 são apresentados os respectivos autovalores e as porcentagens de variância explicada por cada componente principal. Os dois primeiros componentes foram responsáveis por 84,17% da variabilidade total, sendo o PC1 responsável por 60,08% e o PC2 por 24,09% da variação dos dados. Logo, somente os dois primeiros componentes foram retidos, o que é reforçado pelo gráfico *scree plot* apresentado na Figura 4.6a.

Figura 4.5 – Representação gráfica da matriz de correlações amostrais de Pearson observadas dos dados *mtcars*.



Fonte: Do autor (2020).

A seguir são apresentadas três citações de trabalhos em que os autores utilizaram a PCA padrão para analisar os dados. Esses exemplos ilustram como a quantidade de componentes a serem retidos pode variar para diferentes aplicações. Paiva et al. (2010) utilizando dados de 11 características em relação a produção de aves de postura do Programa de Melhoramento Genético da Universidade Federal de Viçosa, verificaram que os três primeiros componentes eram suficientes para explicar 77,00% da variabilidade, dentre onze componentes principais que foram obtidos.

Com dados de 14288 animais da raça Mangalarga Marchador, Meira et al. (2013) observaram que os seis primeiros componentes principais explicaram 78,57% da variação total dos dados. Com base nos resultados obtidos, os autores recomendaram que dentre as 13 características avaliadas, somente a pontuação da marcha, a altura na garupa, o comprimento do dorso,

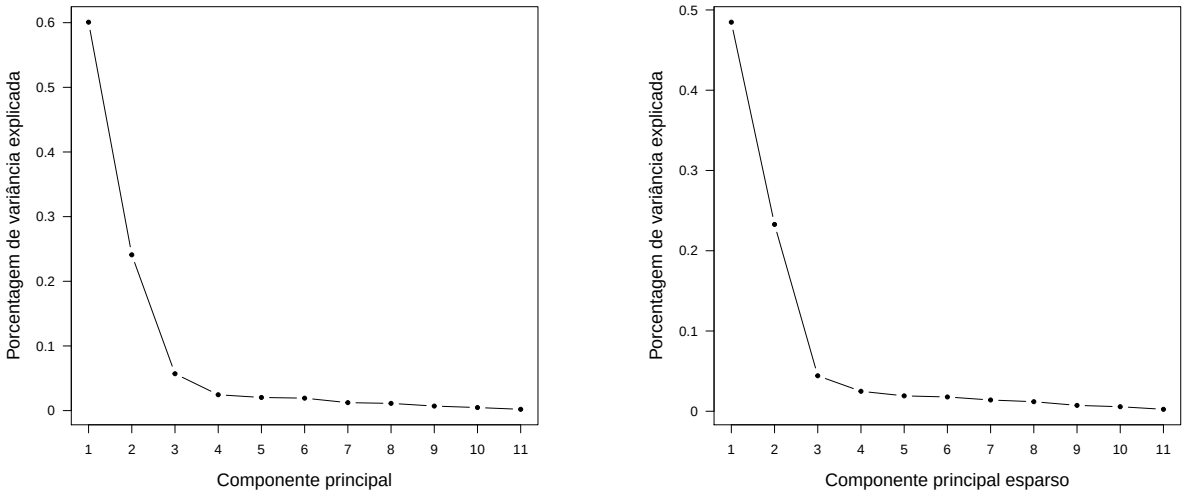
Tabela 4.7 – Autovalores, porcentagens de variância explicada e acumulada dos componentes principais obtidos pelo método PCA padrão.

Componente Principal	Autovalor	Porcentagem	Porcentagem acumulada (%)
PC1	204,86	60,08	60,08
PC2	82,16	24,09	84,17
PC3	19,44	5,70	89,87
PC4	8,36	2,45	92,32
PC5	6,93	2,03	94,35
PC6	6,56	1,92	96,27
PC7	4,19	1,23	97,50
PC8	3,81	1,12	98,62
PC9	2,39	0,70	99,32
PC10	1,61	0,47	99,79
PC11	0,68	0,20	99,99

o comprimento da garupa, a largura da cabeça e o perímetro da canela devem ser mantidas para eventuais trabalhos futuros. Com o objetivo de analisar a relação entre características de produção e proporções genotípicas de bovinos leiteiros mestiços usando a PCA padrão, Fraga et al. (2015) verificaram que dois componentes principais, que juntos explicaram 89,40% da variância total, podem ser utilizados em relação as cinco variáveis originais para o problema.

Figura 4.6 – Gráficos *Scree plot* dos 11 componentes principais obtidos pelos métodos PCA padrão e *Sparse Principal Component Analysis* (SPCA) para os dados *mtcars*.

(a) Método PCA padrão. (b) Método SPCA.



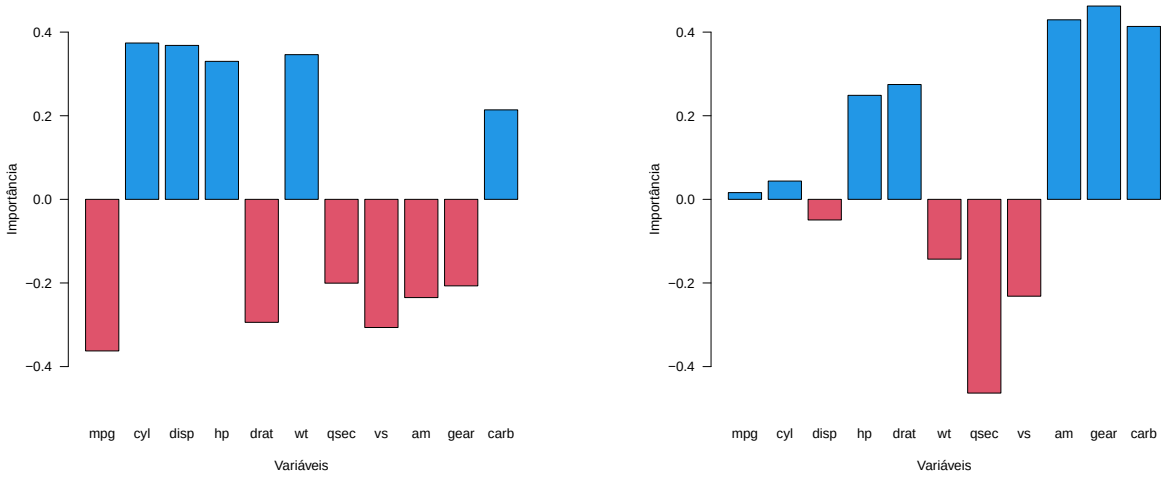
Fonte: Do autor (2020).

Uma vez que os dois primeiros componentes foram retidos no método PCA padrão, é relevante a informação de quais variáveis mais contribuíram para a variabilidade desses com-

ponentes. Isso poderia ser feito examinando-se os *loadings* de cada variável nos componentes PC1 e PC2, excluindo-se aqueles que sejam pequenos. Contudo, por ser um critério de avaliação subjetivo, esse tipo de procedimento é passível de maiores erros. Uma forma interessante de realizar esse exame é por meio de um gráfico de barras, na qual se considera os valores e os sinais (positivo ou negativo) dos *loadings* associados a cada variável nos componentes. Na Figura 4.7 pode ser observado que o PC1 consiste em um contraste entre as variáveis relacionadas à potência dos carros (*cyl*, *disp*, *hp*, *wt*, *carb*) e as varáveis relacionadas à economia (*mpg*, *drat*, *qsec*, *vs*, *am*, *gear*). Por sua vez, o PC2 não permite uma interpretação dessa natureza, uma vez que mistura variáveis relacionadas à potência e economia.

Figura 4.7 – Gráfico de barras contendo a importância (*loadings*) das variáveis no primeiro e segundo componentes obtidos pelo método PCA padrão, com base na matriz de correlações dos dados *mtcars*.

(a) Primeiro componente. (b) Segundo componente.



Fonte: Do autor (2020).

Na Tabela 4.8 são apresentados os vetores de *loadings* e as variâncias obtidas pelos componentes principais dos métodos PCA padrão, SPCA, SGPCA e PACSPCA. Em relação ao método SPCA também foram utilizados somente os dois primeiros componentes principais. Na Figura 4.6b pôde-se verificar que o acréscimo do terceiro componente esparso não contribui de forma significativa em termos da porcentagem de variância acumulada. Além disso, exigiu-se que o método fornecesse oito *loadings* não nulos para os componentes PC1 e PC2. Como salientado nos exemplos simulados, a escolha de um número maior de *loadings* não nulos torna as variâncias explicadas dos componentes também maiores. Isso fica mais evidente quando a comparação baseia-se entre as porcentagens de variância acumulada dos dois primeiros com-

ponentes do cenário 1 de simulação e do presente exemplo. Para quatro *loadings* não nulos no cenário 1 verificou-se que a porcentagem de variância explicada foi de 52,20%, enquanto que para oito *loadings* não nulos essa porcentagem subiu para 71,80% nos dois componentes principais esparsos dos dados *mtcars*.

Para os dois primeiros componentes, os métodos SGPCA e PACSPCA apresentaram os melhores desempenhos em termos de variância explicada acumulada, retendo respectivamente 76,51% e 81,13% da variância total dos dados. Novamente, os dois métodos forneceram os mesmos resultados para o componente PC1. Nesse componente foram formados dois grupos, com o valor de *loading* $-0,301$ para as variáveis (*mpg*, *drat*, *qsec*, *vs*, *am*, *gear*) e $0,301$ para as variáveis (*cyl*, *disp*, *hp*, *wt*, *carb*). Logo, os métodos SGPCA e PACSPCA não se distinguiram quanto a formação dos grupos, por fornecerem os mesmos grupos no componente principal mais importante. Porém, em termos de porcentagem de variância acumulada o método PACSPCA apresentou melhor performance, por reter uma maior porcentagem. Como o conjunto de dados *mtcars* possui uma estrutura de correlações mais ampla, a maior flexibilidade do método PACS parece fornecer uma vantagem ao método PACSPCA, em detrimento ao SGPCA.

Tabela 4.8 – Vetores de *loadings* para a formação dos componentes principais utilizando os métodos PCA, SPCA, SGPCA e PACSPCA, com base na matriz de correlações dos dados *mtcars*.

Variável	PCA		SPCA		SGPCA		PACSPCA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
mpg	-0,362	0,016	-0,390	0	-0,301	0	-0,301	0,007
cyl	0,374	0,044	0,434	0	0,301	0	0,301	0,016
disp	0,368	-0,049	0,413	0	0,301	0	0,301	-0,014
hp	0,330	0,249	0,400	0,210	0,301	0	0,301	0,222
drat	-0,294	0,275	-0,193	0,166	-0,301	0	-0,301	0,251
wt	0,346	-0,143	0,461	-0,176	0,301	0	0,301	-0,060
qsec	-0,200	-0,463	0	-0,546	-0,301	-0,517	-0,301	-0,463
vs	-0,306	-0,232	-0,279	-0,139	-0,301	0	-0,301	-0,219
am	-0,235	0,429	-0,023	0,451	-0,301	0,494	-0,301	0,454
gear	-0,207	0,462	0	0,438	-0,301	0,494	-0,301	0,463
carb	0,214	0,414	0	0,429	0,301	0,494	0,301	0,448
Variância (%)	60,08	24,09	48,50	23,30	57,32	19,19	57,32	23,81
Variância acumulada (%)	60,08	84,17	48,50	71,80	57,32	76,51	57,32	81,13

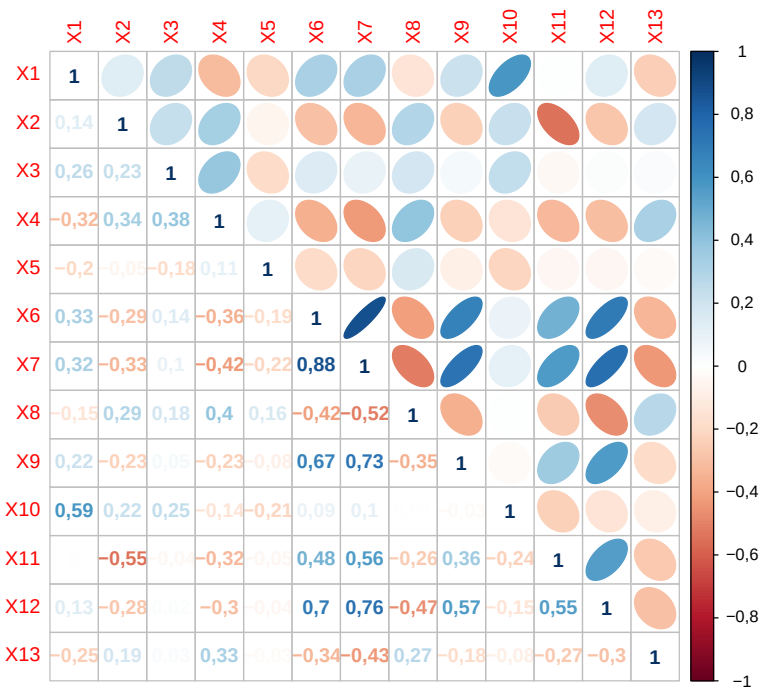
Os resultados fornecidos pelos métodos SGPCA e PACSPCA podem ser interpretados além da constituição de grupos. A análise dos *loadings* fornecidos no PC1 de ambos os métodos corrobora com a análise gráfica relacionada à importância das variáveis para esse componente,

apresentada na Figura 4.7a. Além do efeito de agrupamento, os métodos propostos conseguiram identificar as variáveis que mais contribuíram para a variabilidade do primeiro componente, por distinguir o contraste existente entre variáveis relacionadas à potência (*cyl*, *disp*, *hp*, *wt*, *carb*) e à economia (*mpg*, *drat*, *qsec*, *vs*, *am*, *gear*) dos automóveis.

4.5 Exemplo 2

Na Figura 4.8 é apresentada uma representação gráfica da matriz de correlações dos dados de avaliação de vinhos, considerando os 13 atributos avaliados. Das 78 correlações observadas entre as variáveis, positivas ou negativas, 48 são desprezíveis ($[0-0,3)$), 19 são fracas ($[0,3-0,5)$), 7 são moderadas ($[0,5-0,7)$) e 4 são fortes ($[0,7-0,9)$). Logo, a aplicação dos métodos SGPCA e PACSPCA a esse conjunto de dados reais é importante para avaliar como se comportam os novos métodos em um cenário em que mais de 85% das correlações são desprezíveis ou fracas.

Figura 4.8 – Representação gráfica da matriz de correlações amostrais de Pearson observadas dos dados de avaliação de vinhos.



Fonte: Do autor (2020).

Na Tabela 4.9 são apresentados os resultados dos 13 componentes obtidos pela combinação dos atributos constituintes para a avaliação de vinhos, considerando os autovalores e as porcentagens de explicação das variâncias. O modelo de componentes principais foi capaz de explicar 70,07% da variância total das variáveis originais até o quarto componente, com

autovalores maiores que 1. Esse último comentário foi fornecido em razão de um critério de descarte de componentes principais fornecido por Jolliffe em dois artigos nos anos de 1972 e 1973, envolvendo dados simulados (ou artificiais) e dados reais. O critério adotado por Jolliffe para descartar os componentes se baseia em eliminar componentes cuja variância é inferior a 0,7 ($\lambda_i < 0,7$). Esse não foi o caso do presente exemplo, uma vez que o menor autovalor é 14,44. Nessa situação, a escolha dos componentes a serem retidos poderia ser feita mais uma vez com base na proporção de variância acumulada mínima de 70,00% (JOLLIFFE, 2002, p. 113). Os componentes principais PC1, PC2, PC3 e PC4 foram responsáveis por reter 35,31%, 16,45%, 10,78% e por 7,53% da porcentagem de variância total dos dados, respectivamente.

Tabela 4.9 – Autovalores, porcentagens de variância explicada e acumulada dos componentes principais obtidos pelo método PCA padrão.

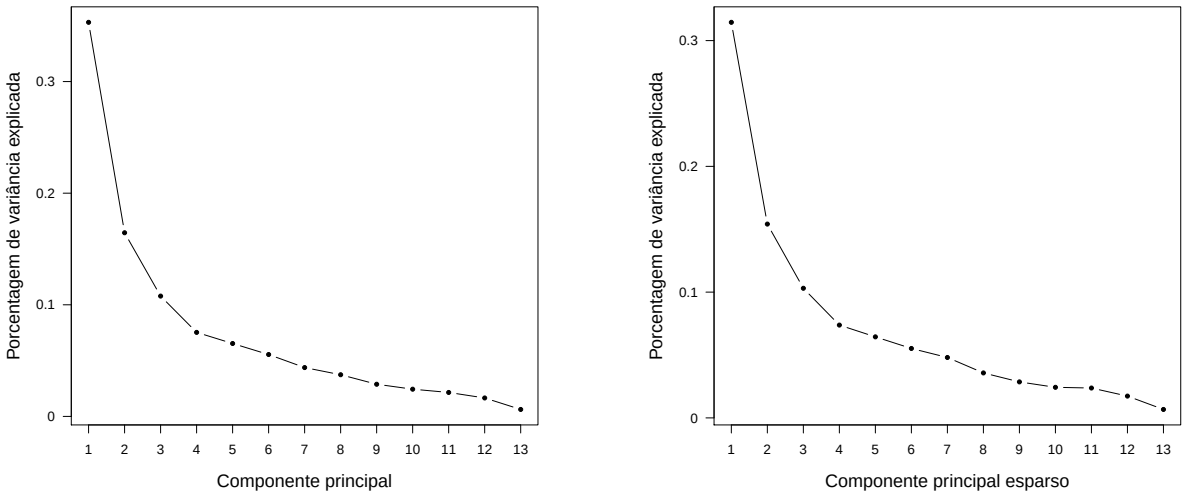
Componente Principal	Autovalor	Porcentagem	Porcentagem acumulada (%)
PC1	817,03	35,31	35,31
PC2	380,76	16,45	51,76
PC3	249,46	10,78	62,54
PC4	174,23	7,53	70,07
PC5	151,15	6,53	76,60
PC6	128,31	5,54	82,15
PC7	101,25	4,38	86,52
PC8	86,38	3,73	90,26
PC9	66,68	2,88	93,14
PC10	56,44	2,44	95,58
PC11	49,66	2,15	97,72
PC12	38,22	1,65	99,38
PC13	14,44	0,62	100,00

A análise do gráfico *scree plot* da PCA padrão (FIGURA 4.9a) sugere a necessidade de se utilizarem no mínimo quatro componentes, dependendo da porcentagem de variância que se deseje reter. Esse exemplo ilustra uma dificuldade que pode surgir em algumas situações práticas, envolvendo os componentes gerados pelo método PCA. Um número maior de componentes, associado também a um número elevado de variáveis, implica não somente em uma dificuldade relacionada a interpretação dos mesmos, mas também é um entrave a um dos principais objetivos na PCA, no que diz respeito a redução da dimensão do problema.

Na sequência de apresentação dos resultados, serão fornecidos apenas aqueles relacionados aos dois primeiros componentes de cada método. Com base na análise do gráfico de barras pode-se observar que o primeiro componente é um contraste entre dois grupos de variáveis, a saber (X_1: álcool, X_6: fenóis totais, X_7: flavonóides, X_9: proantocianinas, X_{10}: intensidade da

Figura 4.9 – Gráficos *Scree plot* dos 13 componentes principais obtidos pelos métodos PCA padrão e *Sparse Principal Component Analysis* (SPCA) para os dados de avaliação da qualidade de vinho.

(a) Método PCA padrão. (b) Método SPCA.

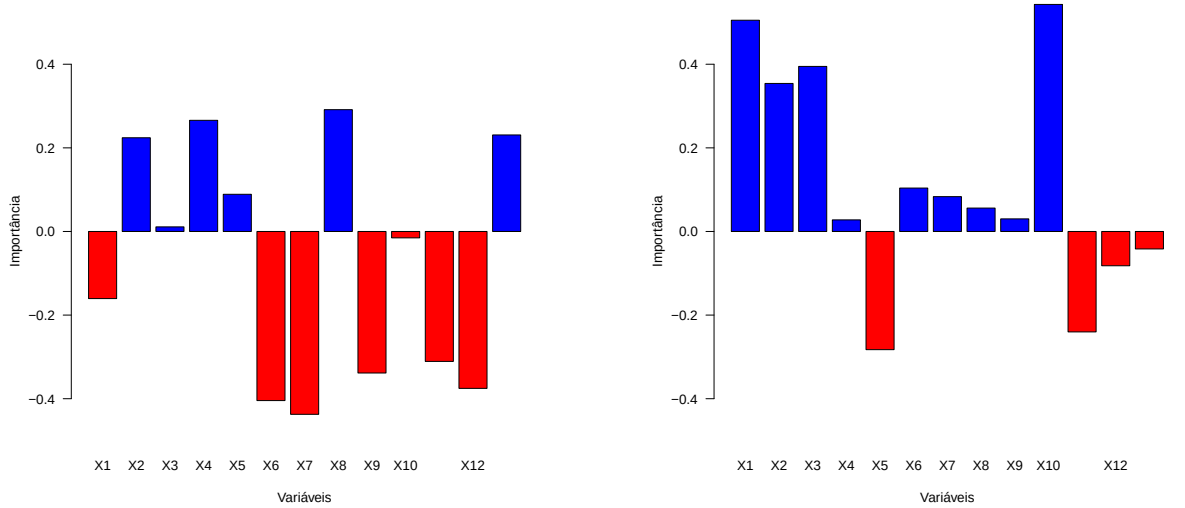


Fonte: Do autor (2020).

cor, X_{11}: matiz, X_{12}: OD280) e (X_2: ácido málico, X_3: cinza, X_4: alcalinidade das cinzas, X_5: magnésio, X_8: fenóis não flavonóides, X_{13}: prolina). Por sua vez, o segundo componente é um contraste entre os grupos (X_5: magnésio, X_{11}: matiz, X_{12}: OD280, X_{13}: prolina) e (X_1: álcool, X_2: ácido málico, X_3: cinza, X_4: alcalinidade das cinzas, X_6: fenóis totais, X_7: flavonóides, X_8: fenóis não flavonóides, X_9: proantocianinas, X_{10}: intensidade da cor).

Figura 4.10 – Gráfico de barras contendo a importância (*loadings*) das variáveis no primeiro e segundo componentes obtidos pelo método PCA padrão, com base na matriz de correlações dos dados de avaliação da qualidade de vinho.

(a) Primeiro componente. (b) Segundo componente.



Fonte: Do autor (2020).

Na Tabela 4.10 são apresentados os vetores de *loadings* e variâncias estimadas dos dois primeiros componentes para cada método. Novamente, o método SPCA apresentou a menor variância acumulada nos dois primeiros componentes (48,46%) em comparação aos métodos SGPCA (49,13%) e PACSPCA (48,85%). Para o método SPCA indicou-se que para cada componente seriam fornecidos dez *loadings* não nulos. No primeiro e mais importante componente, pode-se verificar que os *loadings* nulos foram gerados para as variáveis (X_3: cinza, X_5: magnésio, X_{10}: intensidade da cor), que possuem as menores correlações com as demais variáveis (FIGURA 4.8).

Hsu, Huang e Chen (2014) destacam que um componente crítico e desafiador na análise de dados de alta dimensão na pesquisa do câncer está relacionado a como reduzir a dimensão dos dados e como extrair variáveis relevantes. No artigo, os autores revisaram algumas abordagens da PCA esparsa, incluindo o método SPCA proposto po Zou, Hastie e Tibshirani (2006). Hsu, Huang e Chen (2014) utilizaram um estudo de simulação para comparar a PCA e as PCA's esparsas, além de aplicarem os métodos em um conjunto de dados de câncer de pulmão. No exemplo de aplicação os autores verificaram que as abordagens esparsas do método PCA selecionam menos genes, o que contribui para uma maior eficiência na classificação de risco. Os autores concluíram que, embora a aplicação de componentes principais esparsos na pesquisa do câncer ainda esteja no estágio inicial em comparação à PCA padrão, eles veem um grande potencial, especialmente em alguns tipos de câncer como o câncer de pâncreas, cujos níveis de expressão gênica são bastante homogêneos e, portanto, são muito desafiadores para o desenvolvimento de perfis genômicos.

Zhao et al. (2017) utilizaram os métodos PCA padrão e SPCA em um problema de reconstrução craniofacial. Segundo os autores, a técnica de reconstrução craniofacial auxiliada por computador tem sido amplamente utilizada em campos de investigação criminal, arqueologia, antropologia e cirurgia estética. A pesquisa de Zhao et al. (2017) foi realizada com um banco de dados de 208 tomografias computadorizadas de toda a cabeça de voluntários, pertencentes principalmente ao grupo étnico *Han*, no norte da China. A idade dos indivíduos variou de 19 a 75 anos, sendo constituído por 81 mulheres e 127 homens. As tomografias foram obtidas de um sistema de tomografia *multislice* (*Siemens Sensation16*) no Hospital *Xianyang*, localizado no oeste da China. Com base nos resultados experimentais da análise de similaridade craniofacial, os autores concluíram que os resultados dos dois métodos são idênticos em grande medida, principalmente quanto a redução da dimensão do problema e por manter os

Tabela 4.10 – Vetores de *loadings* para a formação dos componentes principais utilizando os métodos PCA, SPCA, SGPCA e PACSPCA, com base na matriz de correlações dos dados de avaliação da qualidade de vinho.

Variável	PCA		SPCA		SGPCA		PACSPCA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
X_1: Álcool	-0,161	0,506	-0,048	0,586	-0,120	0,541	-0,316	0,612
X_2: Ácido málico	0,224	0,354	0,062	0,192	0,241	0,350	0,316	0,304
X_3: Cinza	0,011	0,395	0	0,414	0	0,374	0	0,304
X_4: AC	0,266	0,028	0,219	0	0,264	0	0,316	0
X_5: Magnésio	0,089	-0,283	0	-0,216	0	-0,176	0,348	-0,179
X_6: Fenóis totais	-0,404	0,104	-0,473	0,010	-0,365	0	-0,316	0
X_7: Flavonóides	-0,437	0,083	-0,480	0	-0,594	0	-0,316	0
X_8: FNF	0,291	0,056	0,068	-0,007	0,264	0	0,316	0
X_9: Proantocianinas	-0,339	0,030	-0,389	-0,003	-0,264	0	-0,316	0
X_{10}: IC	-0,016	0,543	0	0,626	0	0,641	0	0,612
X_{11}: Matiz	-0,311	-0,240	-0,261	-0,098	-0,264	-0,060	-0,316	-0,179
X_{12}: OD280/OD315	-0,376	-0,082	-0,502	0	-0,324	0	-0,316	0
X_{13}: Prolina	0,231	-0,042	0,123	-0,032	0,241	0	0,316	0
Variância (%)	35,31	16,45	31,45	15,41	33,74	15,39	33,20	15,65
Variância acumulada (%)	35,31	51,76	31,45	46,86	33,74	49,13	33,20	48,85

AC: Alcalinidade das cinzas; FNF: fenóis não flavonóides; IC: Intensidade da cor.

principais elementos dos dados originais. Todavia, os autores ressaltam que o uso do SPCA em uma comparação de similaridade permite não apenas a comparação do grau de similaridade de dois dados craniofaciais, mas também a identificação das áreas de alta similaridade, o que é importante para melhorar o efeito da reconstrução craniofacial.

Na Tabela 4.10 também são apresentados os vetores de *loadings* dos métodos SGPCA e PACSPCA. No presente exemplo os novos métodos diferiram entre si quanto a formação de grupos. Em relação aos *loadings* fornecidos pelo método SGPCA pode-se verificar que o método formou três grupos de variáveis no primeiro componente, com *loadings* iguais a 0,241, 0,264 e -0,264 para (X_2: ácido málico, X_{13}: prolina), (X_4: alcalinidade das cinzas, X_8: fenóis não flavonóides) e (X_9: proantocianinas, X_{11}: matiz), respectivamente. Neste ponto, a igualdade dos *loadings* fornecida pelo método SGPCA não pode ser justificada em razão da presença de correlações altas entre as variáveis, pois como já destacado as correlações para esse conjunto de dados são majoritariamente desprezíveis ou fracas. Uma justificativa para os resultados talvez tenha que ser delineada por meio do conhecimento das relações que possam existir entre as variáveis. Os métodos SPCA e SGPCA concordam entre si em relação a *loadings* nulos para as variáveis (X_3: cinza, X_{10}: intensidade de cor) no primeiro componente.

O método PACSPCA forneceu dois grupos no primeiro componente, sendo os grupos (X_1: álcool, X_6: fenóis totais, X_7: flavonóides, X_9: proantocianinas, X_{11}: matiz, X_{12}: OD280) e (X_2: ácido málico, X_4: alcalinidade das cinzas, X_8: fenóis não flavonóides, X_{13}: prolina) com *loadings* iguais a -0,316 e 0,316, respectivamente. Uma justificativa que explique o agrupamento fornecido pelo método PACSPCA também não pode ser dada com base em altas correlações. Todavia, de forma semelhante ao que ocorreu nos exemplos anteriores os dois grupos formados pelo método PACSPCA possuem somente correlações positivas, o que reforça mais uma vez que o sentido das correlações (positiva ou negativa) parece desempenhar uma papel importante na formação de grupos. No segundo componente, os métodos SGPCA e PACSPCA fornecem *loadings* nulos para as mesmas variáveis. Em termos de variância acumulada explicada, o método que apresentou melhor desempenho foi o método SGPCA (49,13%), seguido pelos métodos PACSPCA (48,85%) e SPCA (46,86%).

4.6 Conclusão

Com base nos resultados obtidos nos exemplos simulados e envolvendo dados reais, pode-se concluir que:

1. Nos cenários simulados, os métodos SGPCA e PACSPCA apresentaram resultados satisfatórios em relação ao agrupamento de variáveis, identificando corretamente as variáveis que deveriam ser agrupadas.

2. Apesar da principal característica dos métodos propostos ser o agrupamento de variáveis, esses métodos podem ser utilizados para a seleção de variáveis na PCA, por também fornecerem soluções esparsas.

3. Os métodos SGPCA e PACSPCA tendem a agrupar com maior facilidade as variáveis que apresentem entre si correlações de magnitude moderada, forte e principalmente muito forte. Porém, a aplicação aos dados de seleção de vinhos permitiu que se verificasse que os novos métodos conseguem agrupar algumas variáveis mesmo em uma situação em que a maior parte das correlações são fracas ou desprezíveis. Nesse último caso, a razão para o agrupamento das variáveis deve levar em consideração o conhecimento do pesquisador sobre o fenômeno sob investigação.

4. Pôde-se constatar que o sentido das correlações, negativo ou positivo, também afeta a forma como os métodos SGPCA e PACSPCA podem ou não agrupar algumas variáveis.

5 CONSIDERAÇÕES FINAIS

Os componentes principais com *loadings* esparsos tem sido um interessante tópico de pesquisa desde os anos de 1990. Alguns métodos como o SCoTLASS e o SPCA representam grandes avanços à melhora da interpretabilidade dos componentes principais pela utilização das restrições LASSO e *Elastic Net*, respectivamente. Neste trabalho foram propostos os métodos SGPCA e PACSPCA que utilizam em suas formulações os métodos de regressão OSCAR e PACS, nessa ordem. Assim como os métodos SCoTLASS e SPCA, os novos métodos possuem a capacidade de fornecer *loadings* nulos. Todavia, os métodos SGPCA e PACSPCA generalizam os métodos SCoTLASS e SPCA no sentido de também fornecerem a propriedade de agrupamento de variáveis. Nesse sentido, os métodos propostos fornecem novas formas supervisionadas de se obterem componentes principais esparsos e com a capacidade de formação de grupos de variáveis. Do ponto de vista prático, os novos procedimentos podem ser aplicados na fase preliminar da análise dos dados em adição aos métodos PCA padrão e SPCA. Os grupos resultantes de variáveis podem ser investigados posteriormente para que se possa compreender a estrutura de agrupamento, que são pertinentes ao fenômeno em estudo. Desse modo, o pesquisador pode utilizar os resultados fornecidos para eventualmente selecionar as variáveis que sejam de seu interesse e que deverão permanecer no modelo. Diante dos resultados iniciais, pode-se afirmar que os métodos SGPCA e PACSPCA são procedimentos que possuem boas características quando comparados ao método SPCA, tais como: *i*) facilidade de implementação computacional, *ii*) maior porcentagem de variância explicada, *iii*) habilidade de selecionar as variáveis mais importantes nos componentes de maior relevância e *iv*) capacidade de agrupar as variáveis em cada componente principal constituído.

As próximas etapas neste trabalho de pesquisa consistirão em:

1. Publicar ao menos um artigo relacionado à fundamentação teórica e à aplicação dos métodos SGPCA e PACSPCA.

2. Implementar um pacote para o ambiente *R* para disponibilizar as novas propostas de PCA.

3. Utilizar outros métodos de seleção e agrupamento de variáveis para estender a teoria de componentes principais modificados, com novos métodos que possuam essas propriedades e que possam apresentar melhorias.

6 REFERÊNCIAS

BANCROFT, T. On biases in estimation due to the use of preliminary tests of significance. **Annals of Mathematical Statistics**, v. 15, n. 2, p. 190-204, 1944.

BONDELL, H. D.; REICH, B. J. Simultaneous regression shinkage, variable selection, and supervised clustering of predictors with OSCAR. **Biometrics**, v. 64, n. 1, p. 115-123, 2008.

CADIMA, J.; JOLLIFFE, I. T. Loadings and correlations in the interpretation of principal components. **Journal of Applied Statistics**, v. 22, n. 2, p. 203-214, 1995.

CONT, V. D. Francis Galton: eugenia e hereditariedade. **Scientiae Studia**, v. 6, n. 2, p. 201-218, 2008.

COSTA, L. A. Novo estimador de cumeeira de Rao com aplicação em seleção genômica. 2015. 126 p. **Tese (Doutorado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2015.

DARNELL, A. C. Harold Hotelling 1985-1973. **Statistical Science**, v. 3, n. 1, p. 57-62, 1988.

EFRON, B. *et al*. Least angle regression. **The Annals of Statistics**, v. 32, n. 2, p. 407-499, 2004.

FERREIRA, D. F. **Estatística multivariada**. 3. ed. Lavras: Editora UFLA, 2018.

FRAGA, A. B. *et al*. Multivariate analysis to evaluate genetic groups and production traits of crossbred Holstein x Zebu cows. **Tropical Animal Health and Production**, v. 48, n. 3, p. 533-538, 2015.

FRIEDAN, J. *et al*. **glmnet: Lasso and Elastic-Net regularized generalized linear models**. 2020. Disponível em: <https://cran.r-project.org/web/packages/glmnet/index.html>. Acesso em: 3 mai. 2020.

FRISCH, R. Correlation and scatter in statistical variables. **Nordic Statistical Journal**, v. 8, n. 1, p. 36-102, 1929.

GRUBER, M. H. J. **Improving efficiency by shrinkage: the James-Stein and ridge regression estimator**. 2nd ed. New York: M. Dekker, 1998.

HEYDE, C. C.; SENETA, E. **Statisticians of the centuries**. New York: Springer, 2001.

HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for non-orthogonal problems. **Technometrics**, v. 12, n. 1, p. 55-67, 1970.

HOTELING, A. H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 1, p. 417-441, 1933.

HOTELLING, A. H. The relations of the newer multivariate statistical methods to factor analysis. **British Journal of Mathematical and Statistical Psychology**, v. 10, n. 2, p. 69-79, 1957.

HSU, Y.-L.; HUANG, P.-Y.; CHEN, D.-T. Sparse principal component analysis in cancer research. **Translational Cancer Research**, v. 3, n. 3, p. 182-190, 2014.

JAMES, W.; STEIN, C. Estimation with quadratic loss. **Proceeding Berkeley Symposium on Mathematical Statistics and Probability**, v. 1, n. 1, p. 361-380, 1961.

JEFFERS, J. N. R. Two case studies in the application of principal component analysis. **Journal of the Royal Statistical Society**, v. 16, n. 3, p. 225-236, 1967.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. New Jersey: Prentice Hall, 2007.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. I: Artificial data. **Journal of the Royal Statistical Society**, v. 21, n. 2, p. 160-173, 1972.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II: Real data. **Journal of the Royal Statistical Society**, v. 22, n. 1, p. 21-31, 1973.

JOLLIFFE, I. T. Rotation of principal components: choice of normalization constraints. **Journal of Applied Statistics**, v. 22, n. 1, p. 29-35, 1995.

JOLLIFFE, I. T. **Principal component analysis**. 2nd ed. New York: Springer-Verlag, 2002.

JOLLIFFE, I. T.; TRENDAFILOV, N. T.; UDDIN, M. A modified principal component technique based on the LASSO. **Journal of Computational and Graphical Statistics**, v. 12, n. 3, p. 531-547, 2003.

KENDALL, M. G. **A course in multivariate analysis**. London: Griffin, 1957.

LEBART, L.; MORINEAU, A.; PIRON, M. **Statistique exploratoire multidimensionnelle**. Paris: Dunod, 1995.

LEVENE, H. Harold Hotelling 1985-1973. **The American Statistician**, v. 28, n. 2, p. 71-73, 1974.

MAGNELLO, M. E. Karl Pearson and the establishment of mathematical statistic. **International Statistical Review**, v. 77, n. 1, p. 3-29, 2009.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic Press, 1979.

MEIRA, C. T. *et al*. Seleção de características morfofuncionais de cavalos da raça Mangalarga Marchador por meio da análise de componentes principais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 65, n. 6, p. 1843–1848, 2013.

MONROY, L. G. D.; RIVERA, M. A. M. **Estadística multivariada: inferencia y métodos**. Bogotá: Universidad Nacional de Colombia, 2012.

MONTGOMERY, D. C; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 5th ed. New Jersey: J. Wiley, 2012.

PAIVA, A. L. C. *et al*. Análise de componentes principais em características de produção de aves de postura. **Revista Brasileira de Zootecnia**, v. 39, n. 2, p. 285–288, 2010.

PEARSON, K. On lines and planes of closet fit to systems of points in space. **Philosophical Magazine**, v. 6, n. 2, p. 559-572, 1901.

PEARSON, K. **The grammar of science**. 3rd ed. London: Adam and Charles Black, 1911.

PEREIRA, L. S. Geometria dos métodos de regressão LARS, LASSO e Elastic Net com uma aplicação em seleção genômica. 2017. 167 p. **Tese (Doutorado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2017.

PETRY, S.; TUTZ, G. Shrinkage and variable selection by polytopes. **Journal of Statistical Planing and Inference**, v. 142, n. 1, p. 48-64, 2012.

PORTER, T. M. **Karl Pearson: the scientific life in a statistical age**. New Jersey: Princeton University Press, 2004.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2020. Disponível em: <http://www.R-project.org>. Acesso em: 15 jun. 2020.

RAO, C. R. The use and interpretation of principal component analysis in applied research. **The Indian Journal of Statistics**, v. 26, n. 4, p. 329-358, 1964.

RENCHER, A. C.; SCHAALJE, G. B. **Linear models in statistics**. New Jersey: J. Wiley, 2008.

ROGGO, Y. *et al*. A review of near infrared spectroscopy and chemometrics in harmaceutical technologies. J**ournal of Pharmaceutical and Biomedical Analysis**, v. 44, n. 3, p. 683-700, 2007.

SALSBURG, D. **Uma senhora toma chá**. Rio de Janeiro: Zahar, 2009.

SALEH, A. K. Md. E.; ARASHI, M.; KIBRIA, B. M. G. **Theory of Ridge regression estimation with applications**. New Jersey: J. Wiley, 2019.

SEGAL, M.; DAHLQUIST, K.; CONKLIN, B. Regression approach for microarray data analysis. **Journal of Computacional Biology**, v. 10, n. 6, p. 961-980, 2003.

SHARMA, D. B.; BONDELL, H. D.; ZHANG, H. H. Consistent group identification and variable selection in regression with correlated predictors. **Journal of Computational and Graphical Statistics**, v. 22, n. 2, p. 319-340, 2013.

SILVEIRA, F. G. Abordagem geométrica do método dos quadrados mínimos parciais com uma aplicação a dados de seleção genômica. 2014. 176 p. **Tese (Doutorado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2014.

SMITH, W. L. Harold Hotelling 1985-1973. **The Annals of Statistics**, v. 6, n. 6, p. 1173-1183, 1978.

STEIN, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. **Proceeding Third Berkeley Symposium on Mathematical Statistics and Probability**, v. 1, n. 1, p. 197-206, 1956.

THURSTONE, L. L. Multiple factor analysis. **Psychological Review**, v. 38, n. 1, p. 406-427, 1931.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267-288, 1996.

TIKHNOV, A. N. solution of incorrectly formulated problems and the regularization method. **Soviet Mathematics**, v. 4, n. 4, p. 1035-1038, 1963.

TUTZ, G.; ULBRICHT, J. Penalized regression with correlation-based penalty. **Statistics and Computing**, v. 19, n. 1, p. 239-253, 2009.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S**. 4th ed. New York: Springer-Verlag, 2002.

WICKENS, T. D. **The geometry of multivariate statistics**. New York: Lawrence Erlbaum Associates, 1995.

WILKS, D. S. **Statistical methods in the atmospheric sciences**. 3rd ed. Cambridge: Academic Press, 2011.

WITTEN, D. M.; SHOJAIE, A.; ZHANG, F. The cluster Elastic Net for high-dimensional regression with unknown variable grouping. **Technometrics**, v. 56, n. 1, p. 112-122, 2014.

YU, H. **Bootcluster: bootstrapping estimates of clustering stability**. 2017. Disponível em: <https://cran.r-project.org/web/packages/bootcluster/index.html>. Acesso em: 3 mai. 2020.

ZHAO, J. *et al*. Craniofacial similarity analysis through sparse principal components analysis. **PLoSONE**, v. 12, n. 6, p. 1-18, 2017.

ZOU, H; HASTIE, Y. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society**, v. 67, n. 2, p. 301-320, 2005.

ZOU, H.; HASTIE, T. **elasticnet: Elastic-Net for sparse estimation and sparse PCA**. 2018. Disponível em: <https://CRAN.R-project.org/package=elasticnet>. Acesso em: 15 jan. 2019.

ZOU, H.; HASTIE, T.; TIBSHIRANI, R. Sparse principal component analysis. **Journal of Computacional and Grahical Statistics**, v. 15, n. 2, p. 265-286, 2006.

7 ANEXO

Neste Anexo são apresentados alguns conceitos e resultados que foram utilizados no decorrer do texto. Os conceitos e resultados aqui apresentados também podem ser encontrados em Rencher e Shaalje (2008) e em Monroy e Rivera (2012). Neste caso, as citações dessas referências serão usualmente omitidas.

ANEXO A - Distância

O conceito de distância é um dos mais importantes e sobre os quais muitos conceitos matemáticos foram elaborados, como convergência e espaços métricos. A estatística não é estranha ao seu uso, ainda mais para o desenvolvimento de algumas técnicas definidas e adaptadas a partir de algumas dessas distâncias. Nesta parte, é feita referência ao conceito de distância dentro de um contexto estatístico, sem a pretensão de se fazer uma apresentação rigorosa do assunto. Um dos problemas aos quais a estatística dedicou mais esforço é o estudo da variabilidade. De que se ocupariam os estatísticos se não houvesse variabilidade nos dados? Para isso, foi necessário criar maneiras de medir, usar e modelar a heterogeneidade das informações contidas nos dados ou observações.

Para um investigador, pode ser importante determinar se dois indivíduos, com certas características (variáveis), devem ser considerados próximos ou não. O interesse pode consistir na localização dos indivíduos em uma das várias populações, com base em sua proximidade a cada uma delas. Outra situação é decidir se deve ou não rejeitar uma hipótese estatística com base em sua discrepância com os dados observados (amostra). Uma das maneiras de estimar os parâmetros associados a um modelo de regressão é minimizando as distâncias, em direção da variável resposta, entre os pontos observados e a reta, curva ou superfície de regressão proposta. Essa metodologia é conhecida pelo nome de mínimos quadrados. As vezes, a qualidade de um estimador é julgada por sua distância em relação ao parâmetro. Essa distância se traduz muito comumente em viés, erro de estimativa, variância ou consistência, entre outros. A seguir, é apresentada um tipo de de distância muito útil na maioria das técnicas de estatística multivariada.

A1 - Distância de Mahalanobis

As variáveis usadas em um estudo geralmente estão em escalas de medida diferentes e são correlacionadas. Assim, por exemplo, a altura e o peso das pessoas são quantidades com unidades diferentes (m e kg), de modo que o número que representa a distância entre dois indivíduos mudará não apenas de acordo com as unidades de medida usadas, mas também em função do grau de associação entre as variáveis. Dessa maneira, se duas variáveis estão intimamente relacionadas e em dois objetos ou indivíduos são observados valores muito diferentes, eles devem ser considerados mais distantes do que se os mesmos valores tivessem sido observados em variáveis independentes. O que se quer dizer com isso é que numa análise estatística deve existir uma medida em que a associação entre as variáveis também seja levada em consideração.

A distância de Mahalanobis entre os objetos ou indivíduos $\boldsymbol{X}_h = \left(X_{h1}, X_{h2}, \ldots, X_{hp}\right)$ e $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ é definida pela seguinte forma quadrática:

$$D^2_{hi} = (\boldsymbol{X}_h - \boldsymbol{X}_i)' \boldsymbol{S}^{-1} (\boldsymbol{X}_h - \boldsymbol{X}_i), \tag{7.1}$$

em que $\boldsymbol{S}$ é a matriz de covariâncias amostrais, com $h, i = 1, 2, \ldots, n$.

Essa métrica considera o efeito das unidades de medida e também a correlação entre as variáveis. Para o caso bidimensional, a distância de Mahalanobis entre as observações h e i é dada pela seguinte expressão:

$$D^2_{hi} = \frac{1}{1-r^2} \left[\frac{(X_{h1} - X_{i1})^2}{s_1^2} + \frac{(X_{h2} - X_{i2})^2}{s_2^2} - 2r \frac{(X_{h1} - X_{i1})(X_{h2} - X_{i2})}{s_1 s_2} \right], \tag{7.2}$$

s_1^2 e s_2^2 são as variâncias das variáveis X_1 e X_2, respectivamente, e r é o coeficiente de correlação entre as duas variáveis.

Nessa expressão pode-se observar que se as variáveis não estão correlacionadas $(r = 0)$ então tem-se a distância de Karl Pearson entre as duas variáveis e, além disso, se as variáveis também possuem variâncias iguais a 1, essa distância é reduzida à distância euclidiana ao quadrado. Logo, as distâncias de Karl Pearson e euclidiana são casos especiais da distância de Mahalanobis. Observe também que o terceiro termo de (7.2), que incluí o coeficiente de correlação r, influencia a distância entre dois objetos.

A distância de Mahalanobis é frequentemente usada para medir a distância entre uma observação multivariada (individual) e a média da população de onde a observação provém. Se $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ representa um indivíduo em particular, selecionado aleatoriamente a partir de uma população com média $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ e matriz de covariâncias $\boldsymbol{\Sigma}$, então:

$$D_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{7.3}$$

é considerada como uma medida da distância entre o indivíduo $\boldsymbol{x}_i$ e o centroide $\boldsymbol{\mu}$ da população.

O valor D_i^2 pode ser considerado como um resíduo multivariado para a observação $\boldsymbol{x}_i$, onde resíduo significa a distância entre uma observação e o "centro de gravidade" de todos os dados. Se a população puder ser assumida como normal multivariada, os valores de D_i^2 serão distribuídos como uma distribuição Qui-Quadrado com p graus de liberdade. Dessa maneira, existe um instrumento útil para a detecção de valores extremos.

A distribuição Qui-Quadrado apresenta-se associada à distância de Mahalanobis. Se considerarmos um vetor aleatório composto de p variáveis aleatórias normais e independentes, isto é, $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$, em $X_j \sim N\left(\mu_j, \sigma_j^2\right)$ para $j = 1, 2, \ldots, p$ então a distância padronizada entre o vetor $\boldsymbol{X}$ e o vetor de médias $\boldsymbol{\mu}$ é dada por:

$$\sum_{j=1}^{p} \left(\frac{x_j - \mu_j}{\sigma_j} \right)^2 = (\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{D}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \sum_{j=1}^{p} z_j^2 = \chi_{(p)}^2, \tag{7.4}$$

em que $z_j \sim N(0,1)$ e $\boldsymbol{\Sigma} = \boldsymbol{D} = \text{diag}\left(\sigma_j^2\right)$. Assim, a distribuição χ^2 pode ser interpretada como a distância padronizada entre um vetor de variáveis normais independentes $\boldsymbol{X}$ e seu vetor de médias, ou também, como o comprimento (norma) de um vetor de variáveis aleatórias $N(0,1)$ e independentes. A distância euclidiana é um caso particular da distância de Mahalanobis quando $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

ANEXO B - Álgebra matricial

O Teorema 1 fornece um resultado geral para o posto (rank) do produto de duas matrizes.

Teorema 1: Para qualquer matriz $\boldsymbol{X}$, $\text{rank}[\boldsymbol{X}'\boldsymbol{X}] = \text{rank}[\boldsymbol{X}\boldsymbol{X}'] = \text{rank}[\boldsymbol{X}'] = \text{rank}[\boldsymbol{X}]$.

Teorema 2: Seja $\boldsymbol{X}$ uma matriz de dimensões $(n \times p)$. Se $\text{rank}[\boldsymbol{X}'] = p$, então $\boldsymbol{X}'\boldsymbol{X}$ é positiva definida.

Teorema 3: Uma matriz simétrica $\boldsymbol{A}$ é positiva definida se e somente se existe uma matriz não singular $\boldsymbol{P}$ tal que $\boldsymbol{A} = \boldsymbol{P}'\boldsymbol{P}$.

Corolário 3.1: Uma matriz positiva definida é não singular.

Essa sequência de resultados é útil para garantir que a matriz a matriz $\boldsymbol{X}'\boldsymbol{X}$ possui inversa única. Sabendo que a matriz $\boldsymbol{X}$ possui posto coluna completo e utilizando-se os Teoremas 1 e 2, pode-se concluir a partir do Corolário 3.1 que a matriz $\boldsymbol{X}'\boldsymbol{X}$ é não singular, ou seja, $\boldsymbol{X}'\boldsymbol{X}$ admite inversa e essa é única.

Teorema 4: Se $\boldsymbol{A}$ é uma matriz $(n \times n)$ qualquer e $\boldsymbol{C}$ é uma matriz ortogonal $(n \times n)$, então:

$$\operatorname{tr}\left[\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C}\right] = \operatorname{tr}\left[\boldsymbol{A}\right]. \tag{7.5}$$

Nos Teoremas 5 e 6 são apresentados resultados sobre derivadas de funções de vetores e matrizes que foram muito utilizados em todo o trabalho.

Teorema 5: Seja $\boldsymbol{u} = \boldsymbol{a}'\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{a}$, em que $\boldsymbol{a}' = (a_1, a_2, \ldots, a_p)$ é um vetor de constantes. Então

$$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial\left(\boldsymbol{a}'\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = \frac{\partial\left(\boldsymbol{x}'\boldsymbol{a}\right)}{\partial \boldsymbol{x}}. \tag{7.6}$$

Teorema 6: Seja $\boldsymbol{u} = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{a}$, em que $\boldsymbol{A}$ é uma matriz de constantes. Então

$$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial\left(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}. \tag{7.7}$$

Teorema 7 (Gauss-Markov): Se $E\left[\boldsymbol{y}\right] = \boldsymbol{X}\boldsymbol{\beta}$ e $\operatorname{cov}\left[\boldsymbol{y}\right] = \sigma^2\boldsymbol{I}$, os estimadores de mínimos quadrados $\hat{\beta}_j$, $j = 0, 1, \ldots, k$, tem variância mínima entre todos os estimadores lineares não viesados.

Desse teorema tem-se o seguinte corolário, que é uma extensão para a combinação linear dos $\hat{\beta}$'s.

Corolário 7.1: Se $E\left[\boldsymbol{y}\right] = \boldsymbol{X}\boldsymbol{\beta}$ e $\operatorname{cov}\left[\boldsymbol{y}\right] = \sigma^2\boldsymbol{I}$, o melhor estimador linear não viesado de $\boldsymbol{a}'\boldsymbol{\beta}$ é $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$, em que $\hat{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$.

8 APÊNDICE

Neste Apêndice são apresentados alguns resultados que foram omitidos ao longo do texto. Esses resultados são apresentados em ordem, de acordo com a seção em que o mesmo se encontra.

2.3 Matriz centrada e estimador de mínimos quadrados

Proposição 1: A matriz $\left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right]$ é uma matriz de projeção simétrica.

Prova:

Precisamos mostrar que $\left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right]$ é simétrica e idempotente. Em relação a primeira condição, a simetria é imediata, pois:

$$\left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}\right] = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{bmatrix}. \tag{8.1}$$

Agora observe que:

$$\boldsymbol{J}^2 = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}^2 = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} n & n & \dots & n \\ n & n & \dots & n \\ \vdots & \vdots & \ddots & \vdots \\ n & n & \dots & n \end{bmatrix}. \tag{8.2}$$

Dessa forma,

$$\boldsymbol{J}^2 = \begin{bmatrix} n & n & \dots & n \\ n & n & \dots & n \\ \vdots & \vdots & \ddots & \vdots \\ n & n & \dots & n \end{bmatrix} = n \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} = n\boldsymbol{J}. \tag{8.3}$$

Por último, utilizando (8.3) resulta que a matriz $\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]$ é idempotente. De fato,

$$\begin{aligned} \left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]^2 &= \left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right] \\ &= \boldsymbol{I} - 2\frac{1}{n}\boldsymbol{J} + \frac{1}{n^2}\boldsymbol{J}^2 \\ &= \boldsymbol{I} - 2\frac{1}{n}\boldsymbol{J} + \frac{1}{n^2}n\boldsymbol{J} \\ &= \boldsymbol{I} - 2\frac{1}{n}\boldsymbol{J} + \frac{1}{n}\boldsymbol{J} \\ &= \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}. \end{aligned} \tag{8.4}$$

Portanto, a matriz $\left[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right]$ é uma matriz de projeção simétrica.

■

2.8 Relação entre os componentes principais e o estimador *Ridge*

Proposição 2: A matriz inversa de $(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}' + \tau\boldsymbol{V}\boldsymbol{V}')$ é $\boldsymbol{V}(\boldsymbol{D}^2 + \tau\boldsymbol{I})^{-1}\boldsymbol{V}'$.

Prova:

A demonstração será feita por verificação, mostrando-se que o produto dessas matrizes resulta na matriz identidade. De fato,

$$\begin{aligned} \left(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}' + \tau\boldsymbol{V}\boldsymbol{V}'\right)\boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1}\boldsymbol{V}' &= \boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)\boldsymbol{V}'\boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1}\boldsymbol{V}' \\ &= \boldsymbol{V}\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)\left(\boldsymbol{D}^2 + \tau\boldsymbol{I}\right)^{-1}\boldsymbol{V}' \\ &= \boldsymbol{V}\boldsymbol{V}' \\ &= \boldsymbol{I}. \end{aligned} \tag{8.5}$$

De modo análogo, pode-se mostrar que $\boldsymbol{V}\left(\boldsymbol{D}^2+\tau\boldsymbol{I}\right)^{-1}\boldsymbol{V}'\left(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'+\tau\boldsymbol{V}\boldsymbol{V}'\right)=\boldsymbol{I}$. Portanto, a matriz inversa de $(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'+\tau\boldsymbol{V}\boldsymbol{V}')$ é $\boldsymbol{V}(\boldsymbol{D}^2+\tau\boldsymbol{I})^{-1}\boldsymbol{V}'$.

■

2.9 O método *Octagonal Shrinkage and Clustering Algorithm for Regression* (OSCAR)

Para $p=3$, o convexo K_p da penalização OSCAR é dada por:

$$K_p=\left\{\boldsymbol{\beta}\in\mathbb{R}^3;\sum_{j=1}^{3}|\beta_j|+c\sum_{j<k}^{3}\max\left\{|\beta_j|,|\beta_k|\right\}\leq t\right\}. \tag{8.6}$$

Segue para o conjunto $\sum_{j=1}^{3}|\beta_j|+c\sum_{j<k}^{3}\max\left\{|\beta_j|,|\beta_k|\right\}\leq t$, ao se tomar os máximos dois a dois vem que:

$$|\beta_1|+|\beta_2|+|\beta_3|+c\max\{|\beta_1|,|\beta_2|\}+c\max\{|\beta_1|,|\beta_3|\}+c\max\{|\beta_2|,|\beta_3|\}\leq t. \tag{8.7}$$

Aqui será demonstrado como se obter os 26 vértices da região de penalidade OSCAR quando se considera $p=3$.

i) 6 vértices considerando uma entrada não nula.

Primeiramente, podemos considerar $\beta_2=\beta_3=0$. Desse modo, de (8.7) vem que:

$$|\beta_1|+0+0+c\max\{|\beta_1|,0\}+c\max\{|\beta_1|,0\}+c\max\{0,0\}\leq t. \tag{8.8}$$

Igualando a última expressão a t vem que:

$$\begin{aligned}|\beta_1|+c|\beta_1|+c|\beta_1|=t &\Rightarrow |\beta_1|+2c|\beta_1|=t\\ &\Rightarrow (1+2c)|\beta_1|=t\\ &\Rightarrow |\beta_1|=\frac{t}{1+2c}\\ &\Rightarrow \beta_1=\pm\frac{t}{1+2c}.\end{aligned} \tag{8.9}$$

Nesse caso, tem-se 2 vértices cujas coordenadas são dadas por $\left(\pm\frac{t}{1+2c},0,0\right)$. Os demais 4 vértices podem ser obtidos de maneira análoga, sendo 2 vértices de coordenadas $\left(0,\pm\frac{t}{1+2c},0\right)$ para $\beta_1=\beta_3=0$ e 2 vértices cujas coordenadas são $\left(0,0,\pm\frac{t}{1+2c}\right)$ para $\beta_1=\beta_2=0$.

ii) 12 vértices com duas entradas não nulas e iguais (em valor absoluto).

Considere inicialmente $\beta_1 = \beta_2$ e $\beta_3 = 0$. Neste caso, novamente de (8.7) que:

$$|\beta_1| + |\beta_1| + 0 + c\max\{|\beta_1|, |\beta_1|\} + c\max\{|\beta_1|, 0\} + c\max\{|\beta_1|, 0\} \leq t. \tag{8.10}$$

Da última expressão tem-se:

$$\begin{aligned} |\beta_1| + |\beta_1| + c|\beta_1| + c|\beta_1| + c|\beta_1| + c|\beta_1| = t &\Rightarrow 2|\beta_1| + 3c|\beta_1| = t \\ &\Rightarrow (2+3c)|\beta_1| = t \\ &\Rightarrow |\beta_1| = \frac{t}{2+3c} \\ &\Rightarrow \beta_1 = \pm\frac{t}{2+3c}. \end{aligned} \tag{8.11}$$

Dessa maneira, para $\beta_1 = \beta_2$ e $\beta_3 = 0$ tem-se 4 vértices cujas coordenadas são dadas por $\left(\pm\frac{t}{2+3c}, \pm\frac{t}{2+3c}, 0\right)$. Os demais 8 vértices podem ser obtidos de forma similar, sendo 4 vértices com coordenadas $\left(\pm\frac{t}{2+3c}, 0, \pm\frac{t}{2+3c}\right)$ para $\beta_1 = \beta_3$ e $\beta_2 = 0$ e 4 vértices cujas coordenadas são $\left(0, \pm\frac{t}{2+3c}, \pm\frac{t}{2+3c}\right)$ para $\beta_2 = \beta_3$ e $\beta_1 = 0$.

iii) 8 vértices para o caso $\beta_1 = \beta_2 = \beta_3$.

Novamente considere (8.7). Uma vez que $\beta_1 = \beta_2 = \beta_3$ então:

$$|\beta_1| + |\beta_1| + |\beta_1| + c\max\{|\beta_1|, |\beta_1|\} + c\max\{|\beta_1|, |\beta_1|\} + c\max\{|\beta_1|, |\beta_1|\} \leq t. \tag{8.12}$$

Logo, tomando a igualdade para t na última expressão resulta que:

$$\begin{aligned} |\beta_1| + |\beta_1| + |\beta_1| + c|\beta_1| + c|\beta_1| + c|\beta_1| = t &\Rightarrow 3|\beta_1| + 3c|\beta_1| = t \\ &\Rightarrow (3+3c)|\beta_1| = t \\ &\Rightarrow |\beta_1| = \frac{t}{3+3c} \\ &\Rightarrow \beta_1 = \pm\frac{t}{3+3c}. \end{aligned} \tag{8.13}$$

Como $\beta_1 = \beta_2 = \beta_3$, ao se considerar todas as permutações simples tem-se os 8 vértices, cujas coordenadas são $\left(\pm\frac{t}{3+3c}, \pm\frac{t}{3+3c}, \pm\frac{t}{3+3c}\right)$.

■

2.10 O método *Pairwise Absolute Clustering and Sparsity* (PACS)

Proposição 3: Para todo $j,k = 1,2,\ldots,p$ $(j \neq k)$ segue que:

$$\max\left\{\left|\beta_j\right|,\left|\beta_k\right|\right\} = 0,5\left(\left|\beta_k - \beta_j\right| + \left|\beta_k + \beta_j\right|\right). \tag{8.14}$$

Prova:

Para provar esse resultado serão utilizadas três propriedades da função máximo:

i) $\frac{1}{c}\max\{a,b\} = \max\{ca,cb\}$, para $c \geq 0$.

ii) $\max\{a+c,b+c\} = \max\{a,b\} + c$.

iii) $\max\{a,-a\} = |a|$.

Portanto,

$$\begin{aligned}
\max\left\{\left|\beta_j\right|,\left|\beta_k\right|\right\} &= \max\left\{\left|\beta_k\right|,\left|\beta_j\right|\right\} \\
&= \frac{1}{2}\max\left\{2\left|\beta_k\right|,2\left|\beta_j\right|\right\} \\
&= \frac{1}{2}\max\left\{\left|\beta_k\right| + \left|\beta_k\right|,\left|\beta_j\right| + \left|\beta_j\right|\right\} \\
&= \frac{1}{2}\max\left\{\left|\beta_k - \beta_j\right| + \left|\beta_k + \beta_j\right|,\left|\beta_j - \beta_k\right| + \left|\beta_j + \beta_k\right|\right\} \\
&= \frac{1}{2}\max\left\{\left|\beta_k - \beta_j\right| + \left|\beta_k + \beta_j\right|,\left|\beta_j - \beta_k\right| + \left|\beta_k + \beta_j\right|\right\} \\
&= \frac{1}{2}\left(\max\left\{\left|\beta_k - \beta_j\right|,\left|\beta_j - \beta_k\right|\right\} + \left|\beta_k + \beta_j\right|\right) \\
&= \frac{1}{2}\left(\max\left\{\left|\beta_k - \beta_j\right|,-\left|\beta_k - \beta_j\right|\right\} + \left|\beta_k + \beta_j\right|\right) \\
&= \frac{1}{2}\left(\left|\left|\beta_k - \beta_j\right|\right| + \left|\beta_k + \beta_j\right|\right) \\
&= \frac{1}{2}\left(\left|\beta_k - \beta_j\right| + \left|\beta_k + \beta_j\right|\right) \\
&= 0,5\left(\left|\beta_k - \beta_j\right| + \left|\beta_k + \beta_j\right|\right).
\end{aligned} \tag{8.15}$$

■

4.1 Cenário 1

Na discussão dos resultados do cenário 1 foi mencionado que a especificação dos parâmetros na função *spca* do pacote *elasticnet* pode mudar de forma significativa os valores de *loadings* para os componentes esparsos obtidos pelo método SPCA. Por consequência, isso se reflete na variância ajustada dos componentes esparsos. Antes de exemplificar as mudanças que podem ocorrer para pequenas alterações nas especificações da função *spca*, será apresentada a sintaxe dessa função para o *R*. A sintaxe aqui apresentada é a mesma presente no manual de referência do pacote *elasticnet*.

Descrição

Usando um algoritmo de minimização alternativo para minimizar o critério SPCA.

Uso

```
spca(x, K, para, type=c("predictor","Gram"),
     sparse=c("penalty","varnum"), use.corr=FALSE, lambda=1e-6,
     max.iter=200, trace=FALSE, eps.conv=1e-3)
```

Argumentos

x Uma matriz. Pode ser a matriz de preditores ou a matriz de covariâncias/correlações da amostra.

K Número de componentes.

para Um vetor de comprimento K. Todos os elementos devem ser positivos. Se `sparse="varnum"`, os elementos são inteiros.

type Se `type="preditor"`, `x` é a matriz de preditores. Se `type="Gram"`, a função pede ao usuário para fornecer a matriz de covariâncias ou de correlações amostrais.

sparse Se `sparse="penalty"`, `para` é um vetor de parâmetros de penalidade da norma L_1. Se `sparse="varnum"`, `para` define o número de *loadings* esparsos a serem obtidos.

lambda Parâmetro de penalidade quadrático. O valor padrão é 1e-6.

use.corr Realizar a PCA com a matriz de correlações? Essa opção só é eficaz quando o argumento `type` é definido como `"data"`.

max.iter Número máximo de iterações.

trace Se `TRUE`, imprime seu progresso.

eps.conv Critério de convergência.

Para a obtenção dos vetores de *loadings* do método SPCA utilizou-se no argumento `x` a matriz de observações $\boldsymbol{X}_{\mathrm{p}}$ (dados padronizados). Para isso, no argumento `type` da função *spca* indicou-se que `type="predictor"`. No argumento `sparse` utilizou-se a opção `"varnum"`. A especificação `sparse="varnum"` permite que se indique a quantidade de *loadings* não nulos para cada um dos K=10 componentes esparsos. Zou, Hastie e Tibshirani (2006) afirmaram na discussão do exemplo sintético que o primeiro componente esparso deveria recuperar o fator V_2 usando apenas (X_5, X_6, X_7, X_8) e o segundo componente esparso deveria recuperar o fator V_1 usando somente as variáveis (X_1, X_2, X_3, X_4). Diante disso, no vetor `para` foi indicado que cada vetor de *loadings* deveria apresentar 4 *loadings* não nulos (`para=c(4, ..., 4)`).

Diante do que foi exposto, iremos apresentar mais duas novas especificações para os argumentos da função *spca*, com o intuito de exemplificar como grandes mudanças podem ocorrer nos valores dos vetores de *loadings* e nas variâncias explicadas dos componentes esparsos, para simples mudanças nos argumentos dessa função. Para isso vamos considerar três configurações, sendo que a primeira é a mesma que foi adotada no cenário 1. Essas três configurações são:

```
spca(Xp,K=10,type="predictor",sparse="varnum",para=c(4,4,4,4,4,4,4,4,4,4))

spca(Xp,K=10,type="predictor",sparse="varnum",para=c(5,4,4,4,4,4,4,4,4,4))

spca(Xp,K=10,type="predictor",sparse="varnum",para=c(6,4,4,4,4,4,4,4,4,4))
```

Pode-se observar que a única diferença entre as três configurações encontra-se na quantidade de *loadings* não nulos em `para`, em relação ao primeiro vetor de *loadings*. A quantidade de *loadings* não nulos variou de 4 a 6 nas configurações apresentadas. Na Tabela 8.1 são apresentados os resultados considerando-se somente os vetores de *loadings* esparsos para os dois primeiros componentes.

Os resultados foram apresentados conforme a ordem de saída do *R*. Pode-se observar na Tabela 8.1, para as configurações 1 e 2, que os vetores de *loadings* obtidos forneceram componentes principais esparsos em que o PC1 apresentou menor variância explicada que o PC2.

Tabela 8.1 – Vetores de *loadings* para a formação dos componentes principais utilizando o método SPCA, para três diferentes configurações da função *spca* do pacote *elasticnet*.

Variável	Config. 1		Config. 2		Config. 3	
	PC1	PC2	PC1	PC2	PC1	PC2
X_1	0	-0,496	0	-0,481	0	-0,539
X_2	0	-0,473	0	-0,492	0	-0,503
X_3	0	-0,553	0	-0,510	0	-0,508
X_4	0	-0,474	0	-0,516	0	-0,446
X_5	-0,007	0	-0,014	0	-0,443	0
X_6	-0,078	0	-0,129	0	-0,412	0
X_7	-0,035	0	0	0	-0,392	0
X_8	-0,996	0	-0,089	0	-0,404	0
X_9	0	0	-0,117	0	-0,419	0
X_{10}	0	0	-0,981	0	-0,376	0
Variância (%)	12,50	39,70	17,40	37,60	58,50	39,30
Variância acumulada (%)	12,50	52,20	17,40	55,00	58,50	97,80

Config. = Configuração.

Pode-se observar também que ocorreu um aumento na variância acumulada dos dois componentes esparsos da configuração 1 (52,20%) para os dois componentes esparsos da configuração 2 (55,00%). Isso pode ser explicado em razão do maior grau de esparsidade do PC1 da configuração 1 em relação ao PC1 da configuração 2. Todavia, o resultado mais expressivo pode ser observado para os vetores de *loadings* e para as variâncias dos componentes utilizando-se a configuração 3. Primeiramente, nota-se que os vetores de *loadings* fornecidos geraram componentes que preservam a ordem em função da magnitude (ou importância) das variâncias, em que o PC1 apresentou maior variância explicada em relação ao componente PC2. Além disso, fica evidente também como a proporção de variância acumulada dos componentes da configuração 3 (97,80%) é muito superior a proporção de variância acumulada para os dois primeiros componentes das configurações 1 e 2.

Cabe ressaltar que o intuito aqui não foi explicar as razões que levam a função *spca* a fornecer resultados tão distintos para uma pequena mudança nos valores de seus parâmetros. O objetivo foi mostrar que essas diferenças existem e que a sensibilidade da função *spca* ao fornecer os resultados salienta a necessidade de cautela na escolha dos valores de seus argumentos. Caso ocorram resultados semelhantes aos resultados das configurações 1 e 2, recomenda-se que os vetores de *loadings* sejam ordenados conforme os valores de variância explicada dos componentes com esparsidade, do maior para o menor.