UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Observações atípicas em alta dimensão

Matheus Toshio Hisatugu

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Hisatugu, Matheus Toshio
H673o Observações atípicas em alta dimensão / Matheus Toshio Hisatugu; orientador Mário de Castro. -- São Carlos, 2022.
63 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Observações atípicas em alta dimensão. 2. Análise de componentes principais. 3. Maldição da dimensionalidade. 4. Ruído heteroscedástico. 5. HeteroPCA. I. Castro, Mário de , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Matheus Toshio Hisatugu

Observações atípicas em alta dimensão

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Mário de Castro Andrade Filho

USP – São Carlos
Outubro de 2022

Matheus Toshio Hisatugu

Outliers in high dimension

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Mário de Castro Andrade Filho

USP – São Carlos
October 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Matheus Toshio Hisatugu, realizada em 15/09/2022.

Comissão Julgadora:

Prof. Dr. Mário de Castro Andrade Filho (USP)

Profa. Dra. Viviana Giampaoli (IME-USP)

Prof. Dr. Marcelo Ângelo Cirillo (UFLA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Agradeço primeiramente à minha família que sempre me apoiou desde o início, pelos conselhos e pelo suporte.

Agradeço também ao meu orientador Mário de Castro pelos conselhos e discussões sobre diversos assuntos.

Além disso, agradeço pelos amigos que fiz durante toda essa trajetória desde o início da graduação e espero que para o resto da vida.

Agradeço também à banca de qualificação e defesa pelos conselhos e pelas sugestões para a melhoria do projeto.

E não menos importante, agradeço pelos professores que tive durante todo o processo de aprendizado, todos foram importantes para o meu desenvolvimento tanto como estudante quanto como pessoa.

“The Road goes ever on and on
Down from the door where it began.
Now far ahead the Road has gone,
And I must follow, if I can,
Pursuing it with eager feet,
Until it joins some larger way
Where many paths and errands meet.
And whither then? I cannot say”
(J.R.R. Tolkien)

RESUMO

HISATUGU, M. T. **Observações atípicas em alta dimensão**. 2022. 63 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo e Departamento de Estatística, Universidade Federal de São Carlos, São Carlos – SP.

Observações atípicas e ruído heteroscedástico são duas situações muito comuns em Estatística. Atualmente, a quantidade de dados gerada é muito alta e por essa razão é possível encontrar dados de alta dimensão (número de variáveis, ou dimensão, d tão grande ou maior do que o número de observações n). Além disso, é possível que os dados possuam ruído heteroscedástico, isto é, a variância do ruído pode variar de entrada para entrada. A análise de componentes principais (ACP) é uma técnica muito utilizada que tem como principal objetivo a redução da dimensionalidade. A técnica é utilizada em diversas áreas como a Estatística, Econometria, Aprendizado de Máquina e Matemática Aplicada. Choi e Marron (2019) apresentaram uma nova noção de valores atípicos em alta dimensão que engloba outros tipos e, além disso, investigaram o comportamento dessas observações atípicas no subespaço criado pela análise de componentes principais. Grande parte das técnicas utilizadas nesse contexto são utilizadas sob a suposição de homoscedasticidade, porém, como já mencionado, sabe-se que isso nem sempre acontece. Sendo assim, Zhang, Cai e Wu (2022) propuseram um novo método chamado HeteroPCA que tem como objetivo principal remover o viés da diagonal principal da matriz de covariâncias amostral sob o qual está sujeita devido à heteroscedasticidade. Este trabalho tem como objetivo combinar o método proposto por Zhang, Cai e Wu (2022) com a metodologia proposta por Choi e Marron (2019) para encontrar um subespaço capaz de identificar a presença de observações atípicas quando o ruído heteroscedástico está presente.

Palavras-chave: Análise de componentes principais; observações atípicas em alta dimensão; maldição da dimensionalidade; ruído heteroscedástico; heteroPCA.

ABSTRACT

HISATUGU, M. T. **Outliers in high dimension**. 2022. 63 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo e Departamento de Estatística, Universidade Federal de São Carlos, São Carlos – SP.

Outliers and heteroskedastic noise are two common situations in Statistics. Nowadays the amount of generated data is very high and for this reason it is possible to find high dimensional data (the dimension d is just as large or larger than the number of observations n). Furthermore, it is possible that the data have heteroskedastic noise, which means that the noise variance can be different entrywise. Principal component analysis is a technique that aims to create a subspace with lower dimension than the original space. The technique is used in different areas such as Statistics, Econometrics, Machine Learning and Applied Mathematics. Choi and Marron (2019) introduced a new notion of high dimensional outliers that embraces other types and also investigates the behaviour of these outliers in the subspace created by the principal components analysis. Most of the techniques used in this context are based on the assumption of homoskedastic noise. However, as mentioned before, it is known that this is not always the case. Therefore, Zhang, Cai and Wu (2022) proposed a new method called HeteroPCA, which main objective is to remove the bias of the main diagonal of the sample covariance matrix due to heteroskedasticity. In this work, the main objective is to combine the method proposed by Zhang, Cai and Wu (2022) and the methodology proposed by Choi and Marron (2019) to find a subspace capable of identifying the presence of outliers when heteroskedasticity noise is present.

Keywords: Principal component analysis, high dimensional outliers, curse of dimensionality, heteroskedastic noise, heteroPCA.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Observações atípicas (ou *outliers*), em geral, são vistas como observações ruins que podem trazer prejuízo à análise estatística. As formas mais comuns de se lidar com essas observações atípicas nessas análises são associar pesos menores a essas observações ou até mesmo removê-las da base de dados. No entanto, é possível que em algumas situações as observações atípicas possuam informações importantes. Por exemplo, é possível que observações atípicas em expressão gênica estejam relacionadas com mutações, ou ainda observações atípicas em um processo de produção podem possuir informações importantes sobre algum tipo de falha.

Uma observação atípica pode ser descrita como uma observação que não se encaixa na mesma distribuição da qual a maioria dos dados provém. No entanto, essa definição não é tão simples quando se está em espaços de alta dimensão (número de variáveis, ou dimensão, d tão grande quanto ou maior do que o número de observações n) devido à maldição da dimensionalidade, isto é, o fenômeno no qual conforme a dimensão d aumenta, as observações tendem a ficar cada vez mais distantes umas das outras. Sendo assim, os métodos usuais de detecção de observações atípicas que são baseados, por exemplo, em distância (HAWKINS, 1980) não funcionam bem para dados de alta dimensão. Além disso, não existe um consenso na definição de uma observação atípica e, por essa razão, existem diferentes métodos de detecção baseados nas diferentes características de cada tipo.

Frequentemente, a teoria assintótica usual não fornece boas aproximações para dados de alta dimensão. Sendo assim, neste trabalho é adotada uma nova noção de observações atípicas em alta dimensão que engloba outros tipos. Além disso, é estudado sob quais condições as observações atípicas podem ser diferenciadas dos valores comuns assim como as condições sob as quais as observações atípicas podem ser identificados por um subespaço de baixa dimensão produzido pela análise fatorial (AF) estimada pela análise de componentes principais (ACP).

A ACP e os métodos espectrais são ferramentas muito utilizadas em diferentes campos incluindo a Estatística, Aprendizado de Máquina e Matemática Aplicada. Ambas as técnicas têm

sido estudadas e usadas em diversas aplicações.

A principal ideia da ACP é encontrar uma estrutura implícita de dimensão menor de observações com ruído. Nesse contexto, pode-se dizer que os autovetores representam as direções do subespaço e os autovalores representam a intensidade de cada autovetor. Conforme o desenvolvimento do estudo da ACP, o modelo de covariâncias *spike* (JOHNSTONE, 2001) tem sido extensivamente estudado e utilizado como base metodológica e teórica. Nas condições do modelo, tem-se que

$$\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n \overset{iid}{\sim} N_d(\boldsymbol{\mu},\boldsymbol{\Sigma}_0+\sigma^2\boldsymbol{I}_d),$$

em que $\boldsymbol{\Sigma}_0=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ é uma matriz simétrica em que $\boldsymbol{U}$ é uma matriz com os autovetores, $\boldsymbol{\Lambda}$ é uma matriz diagonal com os autovalores e $\boldsymbol{I}_d$ é a matriz identidade de dimensão d. Equivalentemente, o modelo de covariâncias *spike* pode ser escrito como

$$\boldsymbol{Y}_j=\boldsymbol{X}_j+\boldsymbol{\varepsilon}_j,\ \text{em que}\ \boldsymbol{X}_j \overset{iid}{\sim} N_d(\boldsymbol{\mu},\boldsymbol{\Sigma}_0),\quad \boldsymbol{\varepsilon}_j \overset{iid}{\sim} N_d(\boldsymbol{0},\sigma^2\boldsymbol{I}_d),\quad j=1,\ldots,n, \tag{1.1}$$

com $\boldsymbol{X}_j$ e $\boldsymbol{\varepsilon}_j$ independentes. O objetivo é recuperar $\boldsymbol{\Sigma}_0$. Seja $\hat{\boldsymbol{\Sigma}}$ a matriz de covariâncias amostral de $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$. As propriedades assintóticas dos autovalores e autovetores de $\hat{\boldsymbol{\Sigma}}$ são bem estabelecidas assim como seus estimadores baseados nessa decomposição (ANDERSON, 1963). Uma suposição importante neste caso é que os erros são homoscedásticos no sentido de que se assume que cada $\boldsymbol{\varepsilon}_j$ segue uma distribuição gaussiana esférica.

Em muitas aplicações o ruído pode ser heteroscedástico, ou seja, a variância das entradas de uma matriz de dados não é constante. É natural que o ruído heteroscedástico esteja presente, por exemplo, em bases de dados com diferentes tipos de variáveis. Além disso, para dados que são modelados pelas distribuições de Poisson, multinomial ou binomial negativa, os ruídos são naturalmente heteroscedásticos.

É comum então relaxar a suposição de homoscedasticidade na Equação 1.1 e considerar o modelo de covariâncias *spike* generalizado (BAI; YAO, 2012; YAO; ZHENG; BAI, 2015) para uma observação $\boldsymbol{Y}_j$ tal que

$$\begin{aligned}
\boldsymbol{Y}_j=\boldsymbol{X}_j+\boldsymbol{\varepsilon}_j,\qquad & \mathbb{E}(\boldsymbol{X}_j)=\boldsymbol{\mu},\qquad Cov(\boldsymbol{X}_j)=\boldsymbol{\Sigma}_0,\\
\mathbb{E}(\boldsymbol{\varepsilon}_j)=\boldsymbol{0},\qquad & Cov(\boldsymbol{\varepsilon}_j)=\operatorname{diag}(\sigma_1^2,\ldots,\sigma_d^2),\\
\boldsymbol{\varepsilon}_j=(\varepsilon_1,\ldots,\varepsilon_d)^T,\qquad & \boldsymbol{X}_j \text{ e } \boldsymbol{\varepsilon}_j \text{ são independentes.}
\end{aligned} \tag{1.2}$$

Neste caso, $Cov(\boldsymbol{X}_j)$ é de posto r $(r<d)$ e admite a decomposição em autovalores, isto é,

$$Cov(\boldsymbol{X}_j)=\boldsymbol{\Sigma}_0=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,\ \text{em que}\ \boldsymbol{U}\in\mathbb{R}^{d\times r}\ \text{e}\ \boldsymbol{\Lambda}\in\mathbb{R}^{r\times r}$$

e $\sigma_1^2,\ldots,\sigma_d^2$ são desconhecidas e não necessariamente iguais. Esse modelo também é usado em análise fatorial.

Devido à heteroscedasticidade da variância do ruído, a ACP usual pode levar a estimadores inconsistentes. Utilizar a ACP em $\{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n\}$ significa aplicar a decomposição em valores

singulares (DVS) em $\boldsymbol{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n]$, ou seja, estimar $\boldsymbol{U}$ por vetores singulares à esquerda da matriz centralizada

$$\boldsymbol{Y} - \overline{\boldsymbol{Y}}\mathbf{1}_n^T, \qquad \text{em que } \overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{Y}_j. \tag{1.3}$$

Vale notar que $\mathbb{E}(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}_0 + \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Então, quando as variâncias $\sigma_1^2, \ldots, \sigma_d^2$ são diferentes, pode-se dizer que existe um viés na diagonal principal da matriz de covariâncias amostral $\hat{\boldsymbol{\Sigma}}$. Além do modelo de covariâncias *spike* generalizado em ACP, este fenômeno aparece de forma similar em outros problemas com ruído heteroscedástico como, por exemplo, situações de alta dimensão como a remoção de ruído heteroscedástico em matrizes de posto baixo (CANDÈS; SING-LONG; TRZASKO, 2013) e ACP de Poisson (SALMON *et al.*, 2014).

Florescu e Perkins (2016) propuseram um método chamado de DVS com eliminação diagonal com o intuito de tentar lidar melhor com o viés da diagonal principal de $\hat{\boldsymbol{\Sigma}}$. A ideia é colocar zeros na diagonal principal da matriz de covariâncias amostral e, posteriormente, utilizar a decomposição em valores singulares. No entanto, não é claro se anular a diagonal principal é sempre a melhor escolha. Sendo assim, ao invés de anular as entradas da diagonal principal da matriz de covariâncias amostral, Zhang, Cai e Wu (2022) propuseram que as entradas da diagonal principal sejam iterativamente atualizadas com base nos valores fora dessa diagonal com o intuito de reduzir o viés e, possivelmente, obter uma estimação mais precisa.

Devido ao viés já mencionado, as ferramentas usuais podem não ser ideais para lidar com a ACP heteroscedástica e diversos trabalhos têm sido desenvolvidos nessas condições. Bai e Yao (2012) e Yao, Zheng e Bai (2015) estenderam a teoria do modelo de covariâncias *spike* usual para um modelo generalizado e estudaram a distribuição limite da matriz de covariâncias amostral. Além disso, o desempenho da decomposição em valores singulares e as distribuições assintóticas para os estimadores dos autovalores e autovetores nesse contexto também têm sido estudados (HONG; BALZANO; FESSLER, 2016; HONG; FESSLER; BALZANO, 2018).

As propriedades assintóticas dos autovalores e autovetores amostrais têm sido analisadas em diferentes domínios e os mais estudados são o domínio clássico, o domínio teórico das matrizes aleatórias (*random matrix theory*, ou RMT) e o domínio de alta dimensão e pequenas amostras (*high dimensional low sample size*, ou HDLSS).

O estudo assintótico no domínio clássico é aquele no qual o tamanho amostral $n \to \infty$ e a dimensão d é fixada (ou seja, $n/d \to \infty$). Já o domínio teórico das matrizes aleatórias é aquele que tanto o tamanho amostral n quanto a dimensão d tendem a infinito (ou seja, $n/d \to c$). E o domínio de alta dimensão e pequenas amostras é baseado em um limite, tal que $d \to \infty$ e a amostra de tamanho n é fixado (ou seja, $n/d \to 0$).

No domínio clássico, diferentes estudos foram feitos sob diferentes perspectivas. Por exemplo, Girshick (1939) estudou as propriedades assintóticas dos autovalores e autovetores amostrais quando todos os autovalores de $\boldsymbol{\Sigma}$ são diferentes. Já Lawley (1953) estudou a teoria assintótica dos autovetores no caso em que os menores $d-q$ autovalores de $\boldsymbol{\Sigma}$ são iguais e os demais

são diferentes. Além disso, Anderson (1963) estudou o caso particular em que $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ têm distribuição normal multivariada e fornece a distribuição assintótica dos autovalores amostrais $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$, e autovetores amostrais $\hat{\boldsymbol{U}}_1, \ldots, \hat{\boldsymbol{U}}_d$ quando $\lambda_1, \ldots, \lambda_d$ possui multiplicidades.

No domínio das matrizes aleatórias, um fato conhecido é que a distribuição espectral empírica da matriz de covariâncias amostral converge quase certamente para a distribuição de Marcenko-Pastur (MARCENKO; PASTUR, 1967) quando a matriz de covariâncias populacional é a matriz identidade e d e n tendem ao infinito de forma proporcional. Combinado ao fato de que um autovalor é uma função contínua de uma matriz, isso apoia a ideia de que a matriz de covariâncias amostral não é uma boa estimativa da matriz de covariâncias populacional quando a dimensão d é muito alta. Casos em que uma pequena porção dos autovalores é muito maior do que os demais são comuns para dados de alta dimensão.

Já no contexto de alta dimensão e pequenas amostras, diversos trabalhos foram desenvolvidos. Em particular, Jung e Marron (2009) exploraram o comportamento assintótico dos autovetores quando a grandeza dos autovalores aumenta com uma razão d^{α}. Esse estudo mostra que para $\alpha > 1$ os estimadores são consistentes no subespaço, ou seja, o subespaço gerado pelos autovetores amostrais consistentemente estima o subespaço gerado pelos autovetores populacionais e é inconsistente para $\alpha < 1$, em outras palavras, o ângulo entre cada autovetor amostral e o respectivo autovetor populacional converge para 90 graus. Além disso, Shen, Shen e Marron (2016) fornecem uma estrutura geral da consistência da ACP que conecta os resultados existentes de diferentes domínios com exceção de alguns casos.

Neste trabalho, é explorado o comportamento de observações atípicas em alta dimensão via representação geométrica no domínio de alta dimensão e pequenas amostras e também a teoria assintótica dos autovalores e autovetores em conjunto de dados com poucas observações atípicas dentro da estrutura estudada por Shen, Shen e Marron (2016). Um grande interesse está na estimação consistente das direções atípicas na qual um pequeno números de observações atípicas se encontra. É fornecido para cada cenário uma condição que permite a consistência individual da ACP ou a consistência do subespaço gerado.

Sendo assim, o principal objetivo desse trabalho é utilizar o método proposto por Zhang, Cai e Wu (2022) para estimar a matriz de covariâncias em um contexto em que os dados têm ruído heteroscedástico e alta dimensão e, em seguida, aplicar a metodologia apresentada por Choi e Marron (2019) para verificar se é possível identificar a presença de observações atípicas em um subespaço gerado pela ACP.

O trabalho segue da seguinte forma: No Capítulo 2 é apresentado um possível modelo gerador de observações atípicas em alta dimensão (CHOI; MARRON, 2019) assim como três definições de observações atípicas que são casos particulares do modelo apresentado e algumas propriedades. No Capítulo 3, o método HeteroPCA (ZHANG; CAI; WU, 2022) é apresentado juntamente com resultados teóricos. As simulações que verificam alguns resultados apresentados nos capítulos anteriores, assim como a situação conjunta, isto é, identificação da presença de

observações atípicas em alta dimensão na presença de ruído heteroscedástico, estão no Capítulo 4. No Capítulo 5 é feita uma aplicação em dados reais. E, por último, o Capítulo 6 conclui o trabalho com algumas observações e próximos passos.

CAPÍTULO

2

OBSERVAÇÕES ATÍPICAS EM ALTA DIMENSÃO

Neste capítulo é apresentado o modelo gerador de observações atípicas proposto por Choi e Marron (2019), além de uma nova noção de observações atípicas em alta dimensão. Também são apresentadas a representação geométrica e a consistência da ACP para diferentes cenários.

2.1 Modelo

Como já mencionado, pode-se interpretar os autovetores como direções de um subespaço e os autovalores como a intensidade de cada autovetor. Pode-se dizer ainda que dentre as direções desse subespaço existem dois tipos. Sendo o primeiro tipo a direção das observações atípicas (direções atípicas) que pode levar às observações atípicas em alta dimensão e o segundo tipo a direção das observações típicas (direções base ou direções principais) cuja variação é compartilhada dentre todas as observações incluindo as atípicas. O modelo proposto por Choi e Marron (2019) incorpora as duas direções e é apresentado em três partes.

Primeira parte Seja $\boldsymbol{X}_j$ um vetor aleatório com distribuição normal multivariada de dimensão d, ou seja, $\boldsymbol{X}_j \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_0)$. A decomposição espectral da matriz de covariâncias populacional é

$$\boldsymbol{\Sigma}_0 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,$$

em que $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d]$ contém os autovetores ortonormais de $\boldsymbol{\Sigma}_0$ nas colunas e $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ é uma matriz diagonal com os correspondentes autovalores. Então, o vetor $\boldsymbol{X}_j$ pode ser escrito como

$$\boldsymbol{X}_j = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{W},$$

em que $\boldsymbol{Z} \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$ e $\boldsymbol{W} \sim N_d(\mathbf{0}, \boldsymbol{\Lambda})$. Ou seja, $\boldsymbol{X}_j$ é uma combinação linear de $\boldsymbol{U}_i$ com

coeficiente $w_i \sim N(0, \lambda_i)$, isto é,

$$\boldsymbol{X}_j = \sum_{i=1}^{d} w_i \boldsymbol{U}_i, \qquad w_i \sim N(0, \lambda_i). \tag{2.1}$$

Os valores de w_i podem ser vistos como a intensidade da direção $\boldsymbol{U}_i$. Intuitivamente, se $\boldsymbol{X}_j$ tem coeficiente w_i "grande" para algum i, então a direção $\boldsymbol{U}_i$ é uma direção importante de variação da distribuição que modela $\boldsymbol{X}_j$, enquanto que valores de w_i próximos de 0 quer dizer que a direção $\boldsymbol{U}_i$ não é tão relevante.

Segunda parte Com base na intuição anterior, uma observação atípica pode ser vista como uma observação que segue em algumas direções que a maioria dos dados não segue. Seja $\boldsymbol{U}_i^*$ a i-ésima direção e seu respectivo coeficiente aleatório w_i^*. Então, observações atípicas que seguem essa direção $\boldsymbol{U}_i^*$ têm grandes valores w_i^* enquanto que as outras observações têm pequenos valores de w_i^*. Para modelar essa variação de w_i^* pode-se utilizar uma distribuição com mistura de escala com duas variâncias diferentes, $\tau_{i,2} \gg \tau_{i,1} > 0$, isto é,

$$w_i^* \sim \begin{cases} \sqrt{\tau_{i,1}} z_i, & \text{com probabilidade } 1 - p_i, \\ \sqrt{\tau_{i,2}} z_i, & \text{com probabilidade } p_i, \end{cases} \tag{2.2}$$

em que z_i são variáveis aleatórias independentes e identicamente distribuídas (iid) com média 0 e variância 1 e $0 \leq p_i \leq 1$, com $p_i \approx 0$. A primeira parcela da distribuição de mistura com menor variância, $\tau_{i,1}$, descreve o comportamento da maioria das observações com pouca variação na direção $\boldsymbol{U}_i^*$. A segunda parcela da distribuição com a maior variância, $\tau_{i,2}$, descreve o comportamento das observações atípicas, e assume-se que p_i é pequeno, por exemplo, menor do que 0,05. Esse modelo de mistura representa um mecanismo gerador de observações atípicas.

Terceira parte Um novo modelo que engloba um pequeno conjunto de observações atípicas com base na Equação 2.1 conjuntamente com a distribuição da Equação 2.2 para modelos não necessariamente gaussianos. Seja $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$ uma matriz de observações cujas colunas são observações independentes com distribuição d-dimensional com um pequeno número de vetores tais que as respectivas intensidades w_i são diferentes da maioria. Seja $\{\boldsymbol{U}_i\}_{1 \leq i \leq d}$ um conjunto de vetores ortogonais tal que alguns destes sejam responsáveis pelas possíveis observações atípicas. Assim como na primeira parte, um vetor $\boldsymbol{X}_j$ pode ser expressado como uma combinação linear dos vetores de direção ortonormais, $\{\boldsymbol{U}_i\}_{1 \leq i \leq d}$, cujos coeficientes são variáveis aleatórias independentes com diferentes distribuições de mistura, isto é,

$$\boldsymbol{X}_j = \sum_{i=1}^{d} w_{ij} \boldsymbol{U}_i, \qquad w_{ij} \sim \begin{cases} \sqrt{\tau_{i,1}} z_{ij}, & \text{com probabilidade } 1 - p_i, \\ \sqrt{\tau_{i,2}} z_{ij}, & \text{com probabilidade } p_i, \end{cases} \tag{2.3}$$

tal que

$$\mathbb{E}(\boldsymbol{X}_j) = \boldsymbol{0} \qquad \text{e} \qquad Cov(\boldsymbol{X}_j) = Cov(\boldsymbol{U}\boldsymbol{w}_j) = \boldsymbol{U} Cov(\boldsymbol{w}_j) \boldsymbol{U}^T,$$

em que $\boldsymbol{w}_j = (w_{1j}, \ldots, w_{dj})^T$ e z_{ij} são variáveis iid com média 0, variância 1 e quarto momento limitado. Portanto, as variáveis aleatórias $\{w_{ij}\}_{1 \leq j \leq n}$ com $p_i > 0$ modelam como a direção

$\boldsymbol{U}_i$ vista como uma componente atípica pode gerar observações atípicas. Além disso, para as direções principais usa-se $p_i = 0$, pois isso permite uma flexibilidade para incluir a forma de descrever a variação das direções como na Equação 2.1. Para distinguir as duas componentes, denota-se I_{in} como o conjunto de índices que corresponde às componentes típicas e I_{out} como o conjunto de índices que corresponde às componentes atípicas. Ou seja, $I_{out} = \{1 \leq i \leq d | p_i > 0\}$ e $I_{in} = \{1 \leq i \leq d | p_i = 0\} = \{1, \ldots, d\} \backslash I_{out}$.

Nas condições do modelo na Equação 2.3, é permitido que observações atípicas tenham características ou ruídos similares às observações consideradas típicas, ou seja, com exceção das variáveis que diferenciam as observações em atípicas e típicas, as demais variáveis tem comportamento similares. O modelo também permite que uma observação atípica esteja associado com várias componentes atípicas e isso permite uma flexibilidade na modelagem da natureza das observações atípicas. Um vetor do modelo dado pela Equação 2.3 pode ser visto como um vetor aleatório de uma distribuição de mistura cujas componentes têm estruturas de covariâncias diferentes.

Como mencionado anteriormente, não existe um consenso na definição de uma observação atípica. Diferentes técnicas têm diferentes objetivos, cada uma com base em sua definição. Neste caso, Choi e Marron (2019) apresentam três tipos de observações atípicas que são comumente usados em diversas aplicações e são casos particulares do modelo proposto na Equação 2.3 e estão descritos a seguir.

- Observações atípicas em variáveis específicas: São aquelas que são diferentes da maioria das observações em apenas uma variável. Assumindo que o vetor tenha d variáveis, cada observação pode ser modelada como na Equação 2.3 com $\boldsymbol{U}_i = \boldsymbol{e}_i$ para $i = 1, \ldots, d$, em que $\boldsymbol{e}_i$ é um vetor cuja i-ésima entrada é 1 e as demais 0 (base canônica). Então, uma observação atípica $\boldsymbol{X}_j$ na m-ésima variável pode ser descrito pela Equação 2.3 tal que $\tau_{m,2} > \tau_{m,1}$ e uma proporção p_m ($p_m \approx 0$).

- Observações atípicas com mistura de escala: São aquelas que apresentam um comportamento bem diferente em todas as variáveis simultaneamente e consequentemente estão mais dispersos que a maioria dos dados. Seja

$$\boldsymbol{X}_j \sim \begin{cases} N_d(\boldsymbol{0}, \sigma_1^2 \boldsymbol{\Sigma}), & \text{com probabilidade } 1-p, \\ N_d(\boldsymbol{0}, \sigma_2^2 \boldsymbol{\Sigma}), & \text{com probabilidade } p, \end{cases}$$

  em que $p \approx 0$ e $\sigma_2^2 \gg \sigma_1^2$. Esse modelo de mistura de escala é um caso especial do modelo dado pela Equação 2.3 com $p_i = p, \tau_{i,1} = \sigma_1^2, \tau_{i,2} = \sigma_2^2$ para todo $i = 1, \ldots, d$, em que $\{\boldsymbol{U}_i\}$ é um conjunto de vetores ortogonais como, por exemplo, os autovetores de $\boldsymbol{\Sigma}$. Além disso, I_{out} inclui todos os índices e então $I_{in} = \varnothing$. Então,

$$\boldsymbol{X}_j = \begin{cases} \sum_{i=1}^d w_{ij} \boldsymbol{U}_i, & w_{ij} = \sigma_1 z_{ij}, \text{ com probabilidade } 1-p, \\ \sum_{i=1}^d w_{ij} \boldsymbol{U}_i, & w_{ij} = \sigma_2 z_{ij}, \text{ com probabilidade } p. \end{cases}$$

- Observações atípicas deslocadas: São aquelas que são deslocadas em uma direção comum. Neste caso, as observações atípicas possuem grande parte da sua variação igual à maior parte dos dados, mas apresentam um padrão anormalmente alto ou baixo, que tipicamente é explicado pelo vetor de médias $\boldsymbol{\mu}$. Sejam $\boldsymbol{X}_j$ vetores aleatórios independentes da forma $a_j\boldsymbol{\mu}+\boldsymbol{Z}_j$, em que

$$\boldsymbol{Z}_j \sim N_d(\mathbf{0},\boldsymbol{\Sigma}) \qquad \text{e} \qquad a_j \sim \begin{cases} N(0,\sigma_1^2), & \text{com probabilidade } 1-p, \\ N(0,\sigma_2^2), & \text{com probabilidade } p, \end{cases}$$

com $\boldsymbol{Z}_j$ e a_j independentes. Assumindo que $\sigma_1 < \sigma_2$ e para $p \approx 0$, a variável a_j descreve como uma pequena porção dos dados pode estar deslocada. Definindo um vetor, por exemplo, $\boldsymbol{U}_1$, para ser o vetor de médias normalizado, isto é, $\boldsymbol{U}_1 = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, e os outros vetores serem ortogonais entre si. Então, as variâncias de $\boldsymbol{U}_1$ para observações típicas e atípicas são, respectivamente, $\sigma_1^2\|\boldsymbol{\mu}\|^2+\boldsymbol{U}_1^T\boldsymbol{\Sigma}\boldsymbol{U}_1$ e $\sigma_2^2\|\boldsymbol{\mu}\|^2+\boldsymbol{U}_1^T\boldsymbol{\Sigma}\boldsymbol{U}_1$. E portanto, cada observação pode ser modelada por

$$\boldsymbol{X}_j = \sum_{i=1}^{d} w_{ij}\boldsymbol{U}_i, \qquad w_{1j} \sim \begin{cases} N(0,\sigma_1^2\|\boldsymbol{\mu}\|^2+\boldsymbol{U}_1^T\boldsymbol{\Sigma}\boldsymbol{U}_1), & \text{com probabilidade } 1-p, \\ N(0,\sigma_2^2\|\boldsymbol{\mu}\|^2+\boldsymbol{U}_1^T\boldsymbol{\Sigma}\boldsymbol{U}_1), & \text{com probabilidade } p, \end{cases}$$

e os outros valores de $w_{ij} \sim N(0,\boldsymbol{U}_i^T\boldsymbol{\Sigma}\boldsymbol{U}_i)$ para $i = 2,\ldots,d$.

2.2 Representação geométrica

Devido à maldição da dimensionalidade, conforme a dimensão d aumenta, as observações tendem a ficar cada vez mais distantes umas das outras. Em outras palavras, a distância é altamente dominada por ruído resultando em um conjunto de dados em que as observações atípicas são mais difíceis de serem identificadas. Um caso em que algumas observações atípicas estão próximas umas das outras foi estudada por Zhou e Marron (2016). Além disso, Hall, Marron e Neeman (2005) estudaram a representação geométrica de observações atípicas em um contexto de alta dimensão e pequenas amostras.

Intuitivamente, se $\tau_{i,2}$ em alguma componente atípica é maior do que $\tau_{i,1}$, então as observações atípicas são mais propensas a estarem separadas da concentração dos dados. Em contrapartida, se $\tau_{i,2}$ não é muito diferente de $\tau_{i,1}$, espera-se que as observações atípicas sejam mais difíceis de serem identificadas. Uma observação interessante em dados de alta dimensão é que se uma observação atípica possui uma grande porção de direções atípicas, a dimensão d incentiva a separabilidade das observações atípicas das observações típicas mesmo quando $\tau_{i,2}$ não é muito grande. Por outro lado, se a observação atípica possui um número pequeno de direções atípicas, a dimensão d não incentiva essa separabilidade, mesmo para $\tau_{i,2}$ relativamente grande.

Considera-se um cenário simples em que os dados provêm da Equação 2.3 com $\tau_{i,1} = \sigma^2$ e $\tau_{i,2} = \tau^{(d)}$ para todo i sob a suposição de normalidade. A variação para as componentes atípicas

é indexada por d, $\tau^{(d)}$, como indicação de mudança conforme a dimensão. O modelo pode ser expressado como

$$w_{ij} = \begin{cases} \sigma z_{ij}, & \text{com probabilidade } 1-p_i, \\ \tau^{(d)} z_{ij}, & \text{com probabilidade } p_i, \end{cases} \tag{2.4}$$

para $i \in I_{out}$ e $w_{ij} = \sigma z_{ij}$ para $i \notin I_{out}$.

Considere uma observação típica $\boldsymbol{X}_j$ dada pela Equação 2.4 que pode ser expressada como $\boldsymbol{X}_j = \sum_{i=1}^{d} \sigma z_{ij} \boldsymbol{U}_i$, em que $\boldsymbol{U}_i$ são os autovetores ortonormais. Conforme a dimensão d aumenta, pela lei dos grandes números segue que a distância euclidiana ao quadrado dividida por d converge quase certamente para a constante σ^2, ou seja,

$$\frac{1}{d}\|\boldsymbol{X}_j\|^2 = \frac{1}{d}\sum_{i=1}^{d} \sigma^2 z_{ij}^2 \to \sigma^2. \tag{2.5}$$

Então, pode-se dizer que uma observação típica $\boldsymbol{X}_j$ se encontra aproximadamente na superfície de uma esfera d-dimensional de raio $(\sigma^2 d)^{1/2}$. De forma análoga, pode-se obter um comportamento limitante na distância entre pares de observações típicas. A distância entre duas observações típicas $\boldsymbol{X}_j$ e $\boldsymbol{X}_k$ é aproximadamente igual a $(2\sigma^2 d)^{1/2}$, ou seja, a convergência é quase certa em

$$\frac{1}{d}\|\boldsymbol{X}_j - \boldsymbol{X}_k\|^2 = \frac{1}{d}\sum_{i=1}^{d} \sigma^2 (z_{ij} - z_{ik})^2 \to 2\sigma^2. \tag{2.6}$$

Similarmente, pode-se explorar o comportamento de observações atípicas em alta dimensão. Uma observação atípica $\boldsymbol{X}_{j'}$ pode ser expressado como

$$\boldsymbol{X}_{j'} = \sum_{i \in I_{out}^{j'}} \sqrt{\tau^{(d)}} z_{ij'} \boldsymbol{U}_i + \sum_{i \notin I_{out}^{j'}} \sigma z_{ij'} \boldsymbol{U}_i,$$

em que $I_{out}^{j'}$ é o conjunto de índices para as componentes atípicas relacionadas a $X_{j'}$. Seja $K_{out}^{j'} = |I_{out}^{j'}|$ a cardinalidade do conjunto $I_{out}^{j'}$ para cada d e $p_{out}^{j'} = \lim_{d\to\infty} \frac{K_{j'}^{(d)}}{d}$ a fração de componentes atípicas. A distância de $Y_{j'}$ das outras observações depende do nível de $K_{j'}^{(d)}$, ou seja, para $p_{out}^{j'} > 0$, $p_{out}^{j'} = 0$ com $K_{j'}^{(d)} \to \infty$ e $p_{out}^{j'} = 0$ com $K_{j'}^{(d)}$ fixo. Cada caso requer diferentes níveis de $\tau^{(d)}$ que são descritos a seguir.

Considera-se o primeiro caso em que $p_{out}^{j'} > 0$ com $\tau = \lim_{d\to\infty} \tau^{(d)}$. Aplicando-se a lei dos grandes números ao quadrado da sua distância em relação à origem dividido por d, temos

$$\begin{aligned} \frac{1}{d}\|\boldsymbol{X}_{j'}\|^2 &= \frac{1}{d}\sum_{i \in I_{out}^{j'}} \tau^{(d)} z_{ij'}^2 + \frac{1}{d}\sum_{i \notin I_{out}^{j'}} \sigma^2 z_{ij'}^2 \\ &= \frac{K_{j'}^{(d)}}{d}\frac{1}{K_{j'}^{(d)}}\sum_{i \in I_{out}^{j'}} \tau^{(d)} z_{ij'}^2 + \frac{d - K_{j'}^{(d)}}{d}\frac{1}{d - K_{j'}^{(d)}}\sum_{i \notin I_{out}^{j'}} \sigma^2 z_{ij'}^2 \\ &\to p_{out}^{j'}\tau + (1 - p_{out}^{j'})\sigma^2, \end{aligned} \tag{2.7}$$

quase certamente conforme $d \to \infty$. Isso implica que um valor atípico $\boldsymbol{X}_{j'}$ está aproximadamente a uma distância de $\{\sigma^2 d + p_{out}^{j'}(\tau - \sigma^2)d\}^{1/2}$ da origem. Além disso, a distância entre um valor atípico $\boldsymbol{X}_{j'}$ e um valor típico $\boldsymbol{X}_j$ dividido por $d^{1/2}$ converge quase certamente para $(p_{out}^{j'}(\tau - \sigma^2) + 2\sigma^2)^{1/2}$ conforme $d \to \infty$, isto é,

$$\begin{aligned}\frac{1}{d}\|\boldsymbol{X}_j - \boldsymbol{X}_{j'}\|^2 &= \frac{1}{d}\sum_{i \in I_{out}^{j'}} (\sigma z_{ij} - \sqrt{\tau^{(d)}} z_{ij'})^2 + \frac{1}{d}\sigma^2 (z_{ij} - z_{ij'})^2 \\ &\to p_{out}^{j'}(\tau - \sigma^2) + 2\sigma^2. \end{aligned} \tag{2.8}$$

Sendo assim, valores grandes de $p_{out}^{j'}$ ou τ ajudam a melhor separar as observações atípicas $\boldsymbol{X}_{j'}$ dos valores típicos assumindo que $\tau > \sigma^2$ e $p_{out} > 0$. Em particular, essa propriedade geométrica mostra que mesmo quando τ não é muito maior do que σ^2, para $p_{out}^{j'}$ suficientemente grande a separabilidade ainda é boa para altas dimensões, mas isso tende a não ser o caso para dimensões baixas (FILZMOSER; MARONNA; WERNER, 2008).

O tipo de observações atípicas com escala de mistura na Seção 2.1 é um caso particular do apresentado anteriormente em que $\sigma_1^2 = \sigma^2$ e $\sigma_2^2 = \tau$. Particularmente neste caso, as observações atípicas têm $p_{out} = 1$, e a combinação da Equação 2.5 com a Equação 2.7 conduz a duas esferas d-variadas com diferentes raios: uma esfera com raio $(\sigma^2 d)^{1/2}$ cujas observações típicas se aproximam da superfície, e outra esfera de raio $(\tau d)^{1/2}$ para as observações atípicas.

Foi visto como a dimensão d encoraja a separação geométrica de um valor atípico $\boldsymbol{X}_j$ se $p_{out}^{j'} > 0$. No entanto, isso não acontece quando $p_{out}^{j'} = 0$, pois os termos $p_{out}^{j'}\tau$ e $p_{out}^{j'}(\tau - \sigma^2)$ na Equação 2.7 e na Equação 2.8, respectivamente, se anulam para dimensões altas, desencorajando a separação. Nesta situação, é necessário que $\tau^{(d)}$ seja muito maior do que σ^2 para modelar a separação de forma aproximada. Dessa forma, permite-se que $\tau^{(d)}$ aumente conforme a dimensão d aumenta. Como temos dois casos em que $p_{out}^{j'} = 0$, considera-se primeiramente o caso em que $K_{j'}^{(d)}$ cresce. Um valor atípico é modelado com base na ideia de que

$$\frac{K_{j'}^{(d)}\tau^{(d)}}{d} \to r_{j'} \qquad \text{quando } d \to \infty. \tag{2.9}$$

Então, pode-se mostrar que $\frac{1}{d}\|\boldsymbol{X}_{j'}\|^2 \to r_{j'} + \sigma^2$ e $\frac{1}{d}\|\boldsymbol{X}_j - \boldsymbol{X}_{j'}\|^2 \to r_{j'} + 2\sigma^2$ quando $d \to \infty$. Isso indica que $r_{j'}$ tem um papel importante na separação geométrica entre $\boldsymbol{X}_{j'}$ e valores típicos. Se $r_{j'}$ é muito pequeno, e em particular igual a 0, então as observações na amostra, incluindo as observações atípicas, se comportam assintoticamente como um conjunto de dados sem observações atípicas. Por outro lado, se $r_{j'}$ é suficientemente grande, o valor atípico $\boldsymbol{X}_{j'}$ tende a se separar da superfície da esfera da qual as observações típicas se encontram.

No caso de um número limitado de direções atípicas, isto é $K_{j'}^{(d)} = K_{j'}$, utiliza-se a convergência em distribuição. Tem-se então que

$$\frac{1}{d}\|\boldsymbol{X}_{j'}\|^2 \xrightarrow{D} \frac{1}{K_{j'}}\sum_{i \in I_{out}^{j'}} r_{j'} z_{ij'}^2 + \sigma^2 \quad \text{e} \quad \frac{1}{d}\|\boldsymbol{X}_j - \boldsymbol{X}_{j'}\|^2 \xrightarrow{D} \frac{1}{K_{j'}}\sum_{i \in I_{out}^{j'}} r_{j'} z_{ij'}^2 + 2\sigma^2. \tag{2.10}$$

Analogamente, $r_{j'}$ determina a separabilidade de um valor atípico das observações típicas. Mas vale mencionar que $r_{j'}$ depende apenas de $\tau^{(d)}$, pois $K_{j'}$ foi fixado.

Foi visto como a transição do fenômeno em que as observações atípicas estão próximas à superfície de uma esfera de alta dimensão para longe dessa esfera acontece. Os resultados indicam que existem dois fatores que afetam essa transição sendo que o primeiro é a proporção de componentes atípicas envolvidas em um valor atípico (Equação 2.7) e o segundo é a intensidade dessas direções atípicas (Equação 2.10).

2.3 Consistência da análise de componentes principais

Em um modelo de covariâncias *spike* (Equação 1.2) assume-se que um número fixo de autovalores populacionais é muito maior do que os demais. Isso fornece um sentido importante no qual valores grandes dos autovalores populacionais são consistentemente estimados pela ACP sob algumas condições que dependem dos diferentes domínios assintóticos (SHEN; SHEN; MARRON, 2016). Seja K o número de diferentes componentes entre as matrizes de covariâncias nas componentes de mistura. Define-se por $\{\boldsymbol{U}_i\}_{1\leq i\leq K}$ como direções de pico e $\{\boldsymbol{U}_i\}_{K+1\leq i\leq d}$ como direções base. As componentes base geralmente são consideradas ruídos. Modificando a definição da Seção 2.1, seja o conjunto de índices para as componentes de pico das observações atípicas e as componentes de base, respectivamente, denotadas por

$$I_{out} = \{1 \leq i \leq K | p_i > 0\} \qquad \text{e} \qquad I_{in} = \{1,\ldots,K\}\backslash I_{out},$$

em que p_i é a proporção de observações atípicas. Ou seja, $\{\boldsymbol{U}_i\}_{i\in I_{out}}$ é o conjunto de direções atípicas e $\{\boldsymbol{U}_i\}_{i\in I_{in}}$ é o conjunto de direções principais. A variação de cada direção $\boldsymbol{U}_i$ pode ser representada por

$$\lambda_i = (1-p_i)\tau_{i,1} + p_i\tau_{i,2}$$

pela distribuição de mistura da Equação 2.3 e tais λ_i são os autovalores correspondentes às direções $\boldsymbol{U}_i$. Isso se dá devido ao fato de que a matriz de covariâncias de $\boldsymbol{X}_j$ pode ser escrita como na Equação 2.3. E pela independência de $\{w_{ij}\}_{1\leq i\leq d}$, $Cov(\boldsymbol{w}_j)$ é uma matriz diagonal cujas entradas são

$$\mathbb{V}(w_{ij}) = (1-p_i)\tau_{i,1} + p_i\tau_{i,2},$$

e portanto os valores λ_i são os autovalores de $\boldsymbol{\Sigma}_0$.

Sejam $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ observações como na Equação 2.3 com K componentes de pico como descrito anteriormente. Seja a matriz de covariâncias amostral

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^T = \frac{1}{n-1}\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^T$$

e sua decomposição dada por

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^T \text{ com } \hat{\boldsymbol{U}} = [\hat{\boldsymbol{U}}_1,\ldots,\hat{\boldsymbol{U}}_d] \text{ e } \hat{\Lambda} = \text{diag}(\hat{\lambda}_1,\ldots,\hat{\lambda}_d),$$

em que $\{(\hat{\lambda}_i, \hat{\boldsymbol{U}}_i : i = 1, \ldots, d)\}$ são os pares de autovalores e autovetores de $\hat{\boldsymbol{\Sigma}}$ tais que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_d$. Nesta seção, propriedades assintóticas de $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ e $\hat{\boldsymbol{U}}_1, \ldots, \hat{\boldsymbol{U}}_d$ são analisadas sob uma estrutura mais geral desenvolvida por Shen, Shen e Marron (2016). Essa estrutura inclui diversos domínios como casos particulares já estudados e permite entender conexões entre os vários domínios. Além disso, são fornecidos resultados assintóticos para dados com distribuição de mistura como na Equação 2.3 e permite entender o comportamento de observações atípicas em alta dimensão.

Considera-se três casos em que o tamanho da amostra n aumenta, a dimensão d aumenta e a intensidade aumenta. Como indicação do aumento da intensidade, sejam λ_i, $\tau_{i,1}$ e $\tau_{i,2}$ sequências indexadas por n, ou seja, $\lambda_i^{(n)}$, $\tau_{i,1}^{(n)}$ e $\tau_{i,2}^{(n)}$. Considere as $M+1$ camadas em que os primeiros K autovalores, $\{\lambda_i^{(n)}\}_{1 \leq i \leq K}$, estão agrupados de forma que q_m autovalores pertençam à m-ésima camada em que $\sum_{m=1}^{M} q_m = K$ e os autovalores restantes estão agrupados na $(M+1)$-ésima camada. Define-se ainda que $q_0 = 0, q_{M+1} = d - K$, e as somas parciais $p_m = \sum_{k=0}^{m} q_k$. Então, o conjunto de índices dos autovalores na m-ésima camada pode ser escrito como

$$H_m = \{p_{m-1} + 1, p_{m-1} + 2, \ldots, p_{m-1} + q_m\} \qquad \text{para } m = 1, \ldots, M+1.$$

As suposições a seguir fornecem condições para as variâncias $\tau_{i,1}^{(n)}$ e $\tau_{i,2}^{(n)}$. Diferentes condições são assumidas para as direções de pico principais, direções de pico atípicas e para o ruído, que ajudam a distinguir componentes de pico das componentes base. Existem dois tipos de ruído no modelo, sendo um deles para todas as observações correspondente às componentes base no modelo e o outro apresenta alta intensidade para poucas observações, mas é ruído para a maioria. Esse segundo tipo de ruído é modelado por uma pequena variação nas componentes atípicas. Duas suposições mostram as variâncias para os dois tipos de ruído (CHOI; MARRON, 2019).

Suposição 1. $\lim_{n \to \infty} \tau_{i,1}^{(n)} = \lim_{n \to \infty} \tau_{i,2}^{(n)} = c_\lambda$ para $i \in H_{M+1}$.

Suposição 2. $\lim_{n \to \infty} \tau_{i,1}^{(n)} = c_\lambda$ para $i \in I_{out}$.

Em um modelo de covariâncias *spike*, a intensidade do ruído é descrita nas componentes base e os autovalores correspondentes são assumidos constantes para modelar ruído branco. Isso ajuda os autovalores correspondentes ao ruído a terem algumas propriedades assintóticas. Por exemplo, a distribuição dos autovalores converge para uma distribuição conhecida, como a lei de Marcenko-Pastur ou a lei semi circular, e os autovalores extremos (menor e maior autovalores) são consistentes para alguns valores ou seguem assintoticamente a distribuição de Tracy-Widom (MARCENKO; PASTUR, 1967). A Suposição 1 descreve o comportamento assintótico da intensidade do ruído para as direções base $\{\boldsymbol{U}_i\}_{i \in H_{M+1}}$. E para uma dimensão alta d, os autovalores $\{\lambda_i^{(n)}\}_{i > K}$ são iguais a c_λ. A Suposição 2 descreve a variância do ruído $(\tau_{i,1}^{(n)})$ para as componentes atípicas. Como as componentes de pico atípicas são apenas ruídos para a

maioria das observações, a mesma variação pode ser assumida para as componentes base. Sendo assim, a variância do ruído para as direções de pico atípicas também é assintoticamente igual a c_λ. Isso conecta o modelo atípico com o modelo nulo, ou seja, o caso sem as componentes de pico atípicas, tal que as componentes atípicas se juntam às componentes de ruído base.

A intensidade de cada componente de pico é determinada pela variação que cada componente está envolvida, que é equivalente ao seu correspondente autovalor. Para uma amostra de tamanho n grande, os autovalores $\lambda_i^{(n)}$, são simplesmente $\tau_{i,1}^{(n)}$ para $i \in I_{in}$, enquanto que para I_{out}, os autovalores são $p_i\tau_{i,2}^{(n)}$, pois a variação de $\tau_{i,2}^{(n)}$ domina a variação de $\tau_{i,1}^{(n)}$. A consistência da ACP depende fortemente da magnitude dos autovalores, que são especificados nas seguintes suposições (CHOI; MARRON, 2019). Sejam $\delta_m^{(n)}$ para $m = 1,\ldots,M$ sequências de valores constantes indexadas por n.

Suposição 3. $\lim_{n\to\infty} \frac{\tau_{i,1}^{(n)}}{\delta_m^{(n)}} = 1$ para $i \in H_m \cap I_{in}$ e $\lim_{n\to\infty} \frac{\tau_{i,2}^{(n)}}{p_i\delta_m^{(n)}} = 1$ para $i \in H_m \cap I_{out}$, $m = 1,\ldots,M$.

Suposição 4. Conforme $n \to \infty$, $\delta_1^{(n)} \succ \delta_2^{(n)} \succ \ldots \succ \delta_M^{(n)} \succ \delta_{K+1}^{(n)}$ em que $a_n \succ b_n$ significa $\lim_{n\to\infty} \frac{a_n}{b_n} > 1$.

A Suposição 3 permite que as componentes na mesma camada tenham autovalores assintóticos equivalentes. A Suposição 4 assume que diferentes camadas possuem diferentes coeficientes limitantes, que permite a caracterização dos M subespaços gerados pelas direções em cada camada.

Apesar de assumir uma distribuição de mistura para a estrutura das observações de forma que essas observações não são iid, ainda assim tem-se resultados assintóticos parecidos como os obtidos por Shen, Shen e Marron (2016). Isso é possível, pois apesar de as observações não serem de distribuições idênticas, é permitido que os autovetores do modelo na Equação 2.3 sejam os mesmos. Então, tem-se uma matriz de covariâncias tal que um modelo de covariâncias *spike* possa ser empregado mesmo quando as observações seguem distribuições diferentes.

De forma geral, a intensidade das componentes e o aumento do tamanho amostral n incentivam a consistência da ACP enquanto que o aumento da dimensão d não incentiva a consistência. Quando a intensidade na m-ésima camada, com n crescente, prevalece sobre a dimensão d, ou seja, $\frac{d}{n\delta_m^{(n)}} \to 0$, segue que os estimadores dos autovetores são consistentes no subespaço na m-ésima camada, assim como os estimadores dos autovalores. O Teorema 1 apresentado por Choi e Marron (2019) demonstra tal comportamento assintótico em diferentes cenários.

Teorema 1. Nas condições da Suposição 1 à Suposição 4, tem-se que

(a) se $\frac{d}{n\delta_M^{(n)}} \to 0$, então

(i) para $i \leq K$ $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to 1$, $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ para $i \in I_{in}$ e $\lambda_i^{(n)} = p_i \tau_{i,2}^{(n)}$ para $i \in I_{out}$;

(ii) para $i > K$ e uma constante c,

- se $0 < c < \infty$, $c_\lambda (1 - \sqrt{c})^2 \leq \hat{\lambda}_{n \wedge d} \leq \hat{\lambda}_1 \leq c_\lambda (1 + \sqrt{c})^2$ assintoticamente;
- se $c = \infty$, $\frac{n\hat{\lambda}_i}{d} \to c_\lambda$;
- se $c = 0$, $\hat{\lambda}_i \to c_\lambda$.

(b) se $\frac{d}{n\delta_h^{(n)}} \to 0$ em que $1 \leq h \leq M$ e $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$, então

(i) para $i \leq p_h$, $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to 1$ em que $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ para $i \in I_{in}$ e $\lambda_i^{(n)} = \tau_{i,2}^{(n)}$ para $i \in I_{out}$;

(ii) para $i > p_h$, $\frac{n\hat{\lambda}_i}{d} \to c_\lambda$.

O Teorema 1 considera dois cenários: (a) Quando as intensidades das componentes são fortes e (b) quando a intensidade é forte até a h-ésima camada e as demais são influenciados pelo aumento da dimensão, isto é, $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$. Os diferentes cenários levam a diferentes situações assintóticas: O Teorema 1 (a) considera os três casos para o limite de c, isto é, $0 < c < \infty$, $c = \infty$ e $c = 0$, enquanto o Teorema 1 (b) considera somente o caso $c = \infty$. Isso se deve à condição $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$ de (b) que em conjunto com a Suposição 4 desconsideram os casos $c < \infty$, pois $\frac{d}{n\lambda_{K+1}^{(n)}} \to \infty$ somente acontece quando $\frac{d}{n} \to \infty$. Em ambos os casos, se a intensidade em uma camada é forte o suficiente tal que $\frac{d}{n\delta^{(n)}} \to 0$, então os autovalores amostrais correspondentes à camada estimam os autovalores populacionais de forma consistente. Por outro lado, se os sinais de pico não são tão fortes, então os correspondentes autovalores amostrais tendem a ser "absorvidos" pela pequena massa de autovalores.

Intuitivamente, apesar de a componente atípica ser intensa, sua influência é mais fraca do que a verdadeira, pois perde poder devido à pequena probabilidade de participação. A Suposição 3 mostra essa intuição e dá uma condição tal que a intensidade da i-ésima componente atípica deve ser $1/p_i$ vezes maior do que a intensidade da componente típica na mesma camada para compensar essa perda de poder. Com base nessa suposição, o Teorema 1 demonstra que assintoticamente à intensidade da componente atípica seria igual à intensidade da componente base na mesma camada. Em particular, o autovalor amostral de uma componente atípica converge para a variância dominante $(\tau_{i,2}^{(n)})$ multiplicada pela correspondente proporção (p_i) no modelo de mistura da Equação 2.3. Sendo assim, as verdadeiras intensidades das componentes atípicas podem ser aproximadamente estimadas dividindo os autovalores correspondentes pela proporção $(\approx p_i)$ das observações atípicas relevantes.

Apesar de o Teorema 1 sugerir que alguns autovalores amostrais grandes podem estimar de forma consistente as verdadeiras intensidades, isso não é suficiente para dizer que as correspondentes direções das componentes típicas constroem um subespaço útil para a detecção de observações atípicas. Em dados de alta dimensão, procurar por observações atípicas

em subespaços de dimensão muito menor em que as observações atípicas são distinguíveis é vantajoso. O subespaço pode fornecer informações do porquê uma observação se destaca e também até que ponto essa observação é um valor atípico. Isso é quase impossível quando se usa a dimensão total (d) devido à grande quantidade de ruído. Uma vez que o subespaço é encontrado, um método de detecção convencional apropriado pode ser aplicado às observações de baixa dimensão aproximada. A prova do Teorema 1 pode ser encontrada em Choi e Marron (2019) juntamente com outras propriedades.

CAPÍTULO

3

ANÁLISE DE COMPONENTES PRINCIPAIS E RUÍDO HETEROSCEDÁSTICO

Neste capítulo será apresentado o algoritmo do método HeteroPCA criado por Zhang, Cai e Wu (2022) e algumas de suas propriedades.

3.1 Conceitos preliminares

No decorrer do capítulo serão utilizadas algumas notações que serão apresentadas a seguir. Para qualquer sequência de valores positivos $\{a_k\}$ e $\{b_k\}$, denota-se como $a \lesssim b$ e $b \gtrsim a$ se existe uma constante $C > 0$ tal que $a_k \leq Cb_k$ para todo k. Tem-se também que $a \asymp b$ se $a \lesssim b$ e $a \gtrsim b$ ambas são verdadeiras. Para qualquer matriz real $\boldsymbol{M}$ de dimensão $d_1 \times d_2$, seja $\lambda_k(\boldsymbol{M})$ o k-ésimo maior valor singular. Então, a decomposição em valores singulares de $\boldsymbol{M}$ pode ser escrita como

$$\boldsymbol{M} = \sum_{k=1}^{d_1 \wedge d_2} \lambda_k(\boldsymbol{M}) \boldsymbol{u}_k \boldsymbol{v}_k^T,$$

em que $d_1 \wedge d_2 = \min(d_1, d_2)$, $\boldsymbol{u}_k$ e $\boldsymbol{v}_k$ são o k-ésimo vetor à esquerda e à direita, respectivamente. Seja também $\mathrm{DVS}_r(\boldsymbol{M}) = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$ os principais r vetores singulares à esquerda e $QR(\boldsymbol{M})$ a parte Q da decomposição QR de $\boldsymbol{M}$. Além disso, a norma espectral matricial

$$\|\boldsymbol{M}\| = \sup_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{M}\boldsymbol{u}\|_2 = \lambda_1(\boldsymbol{M})$$

em que $\|\cdot\|_2$ é a norma L2 euclidiana.

Seja ainda $\mathbf{0}_{m \times n}$ uma matriz cujas entradas têm valor 0 de dimensão $m \times n$ e $\mathbf{1}_{m \times n}$ uma matriz de dimensão $m \times n$ cujas entradas têm valor 1. Além disso, sejam $\mathbf{0}_m$ e $\mathbf{1}_m$ os vetores com entradas iguais a 0 e 1, respectivamente. Denota-se ainda por

$$\mathbb{O}_{d,r} = \{\boldsymbol{U} \in \mathbb{R}^{d \times r} : \boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}_d\}$$

como o conjunto de todas as matrizes de dimensão $d \times r$ com colunas ortonormais. Para qualquer $\boldsymbol{U} \in \mathbb{O}_{d,r}$, denota-se por $\boldsymbol{U}_\perp \in \mathbb{O}_{d,d-r}$ como o complemento ortogonal tal que $[\boldsymbol{U}\boldsymbol{U}_\perp] \in \mathbb{R}^{d\times d}$ é uma matriz ortogonal.

Motivado por uma suposição muito utilizada na completitude matricial, condição de incoerência (CANDÈS; RECHT, 2009), define-se como constante de incoerência de $\boldsymbol{U} \in \mathbb{O}_{d,r}$ como

$$I(\boldsymbol{U}) = \frac{d}{r} \max_{i\in\{1,\ldots,d\}} \|\boldsymbol{e}_i^T \boldsymbol{U}\|_2^2. \tag{3.1}$$

Tem-se ainda que a distância $\operatorname{sen}\Theta$ é usada para medir a distância entre subespaços. Especificamente para quaisquer $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{O}_{d,r}$, define-se

$$\|\operatorname{sen}\Theta(\boldsymbol{U}_1,\boldsymbol{U}_2)\| = \|\boldsymbol{U}_{1\perp}^T \boldsymbol{U}_2\| = \|\boldsymbol{U}_{2\perp}^T \boldsymbol{U}_1\|.$$

Para qualquer matriz quadrada $\boldsymbol{A}$, seja $\Delta(\boldsymbol{A})$ a matriz $\boldsymbol{A}$ com os valores da diagonal principal iguais a 0 e seja $D(\boldsymbol{A})$ a matriz $\boldsymbol{A}$ com os valores iguais a 0 com exceção da diagonal principal. Ou seja, $\boldsymbol{A} = \Delta(\boldsymbol{A}) + D(\boldsymbol{A})$.

3.2 ACP heteroscedástica

Seja $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ uma amostra cujas observações são independentes e identicamente distribuídas seguindo o modelo de covariâncias *spike* da Equação 1.2. Para estimar uma matriz de covariâncias $\boldsymbol{\Sigma}_0$, o estimador mais natural é $\hat{\boldsymbol{U}} = \mathrm{DVS}_r(\hat{\boldsymbol{\Sigma}})$, ou seja, o subespaço composto pelos primeiros r vetores singulares à esquerda de $\hat{\boldsymbol{\Sigma}}$. Esse estimador pode ser visto como um problema de otimização (GOLUB; HOFFMAN; STEWART, 1987), tal que

$$\hat{\boldsymbol{U}} = \mathrm{DVS}_r(\tilde{\boldsymbol{\Sigma}}), \quad \text{em que } \tilde{\boldsymbol{\Sigma}} = \underset{\tilde{\boldsymbol{\Sigma}}: r(\tilde{\boldsymbol{\Sigma}})\leq r}{\arg\min} \|\tilde{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}\| \text{ e } r(\tilde{\boldsymbol{\Sigma}}) \text{ é o posto da matriz } \tilde{\boldsymbol{\Sigma}}. \tag{3.2}$$

Uma importante variação no teorema de Davis e Kahan (1970) dado por Yu, Wang e Samworth (2015) é tal que

$$\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\| \lesssim \frac{\|\hat{\boldsymbol{\Sigma}} - (\boldsymbol{\Sigma}_0 + \beta\boldsymbol{I}_d)\|}{\lambda_r(\boldsymbol{\Lambda})} \wedge 1, \tag{3.3}$$

para qualquer escalar $\beta \geq 0$. Como visto no Capítulo 1, $\mathbb{E}(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}_0 + \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Sendo assim, a diferença $\hat{\boldsymbol{\Sigma}} - (\boldsymbol{\Sigma}_0 + \beta\boldsymbol{I}_d)$ não é homogênea se os valores $\sigma_1^2, \ldots, \sigma_d^2$ são diferentes. Em outras palavras, os valores da diagonal principal podem ser muito diferentes entre si.

Para obter estimativas mais robustas de $\boldsymbol{U}$, é proposto o seguinte procedimento computacional. Se a diferença $\hat{\boldsymbol{\Sigma}} - (\boldsymbol{\Sigma}_0 + \beta\boldsymbol{I}_d)$ tem amplitude alta na diagonal principal, ignora-se as entradas da diagonal principal na Equação 3.2 e considera-se

$$\hat{\boldsymbol{U}} = \mathrm{DVS}_r(\hat{\boldsymbol{M}}), \quad \text{em que} \quad \hat{\boldsymbol{M}} = \underset{\hat{\boldsymbol{M}}: r(\hat{\boldsymbol{M}})\leq r}{\arg\min} \|\Delta(\hat{\boldsymbol{M}} - \hat{\boldsymbol{\Sigma}})\|. \tag{3.4}$$

Como a Equação 3.4 é não convexa, considera-se o seguinte algoritmo:

Algoritmo 1 – Algoritmo HeteroPCA

Fornecer $\hat{\boldsymbol{\Sigma}}$, posto r, número de iterações T.
Passo 1 Inicializar atribuindo valor 0 à diagonal principal de $\hat{\boldsymbol{\Sigma}}$: $\boldsymbol{N}^{(0)} = \Delta(\hat{\boldsymbol{\Sigma}})$.
Passo 2 Para $t = 0, \ldots, T$ aplicar a decomposição em valores singulares em $\boldsymbol{N}^{(t)}$ e definir $\tilde{\boldsymbol{N}}^{(t)}$ como a melhor aproximação de posto r:

$$\boldsymbol{N}^{(t)} = \boldsymbol{U}^{(t)}\boldsymbol{\Sigma}^{(t)}(\boldsymbol{V}^{(t)})^T = \sum_{i=1}^{d} \lambda_i^{(t)} \boldsymbol{u}_i^{(t)} (\boldsymbol{v}_i^{(t)})^T, \quad \text{em que } \lambda_1^{(t)} \geq \lambda_2^{(t)} \geq \cdots \lambda_r^{(t)} \geq 0,$$

$$\tilde{\boldsymbol{N}}^{(t)} = \sum_{i=1}^{r} \lambda_i^{(t)} \boldsymbol{u}_i^{(t)} (\boldsymbol{v}_i^{(t)})^T.$$

Passo 3 Atualizar $\boldsymbol{N}^{(t+1)} = D(\tilde{\boldsymbol{N}}^{(t)}) + \Delta(\boldsymbol{N}^{(t)})$, isto é, trocar as entradas da diagonal principal de $\boldsymbol{N}^{(t)}$ pelas entradas de $\tilde{\boldsymbol{N}}^{(t)}$.

$$\boldsymbol{N}_{ij}^{(t+1)} = \begin{cases} \boldsymbol{N}_{ij}^{(t)} = \tilde{\boldsymbol{N}}_{ij}^{(t)}, & \text{se } i = j; \\ \hat{\boldsymbol{\Sigma}}_{ij}, & \text{se } i \neq j. \end{cases}$$

Passo 4 Repetir os passos 2 e 3 até a convergência ou até que o número de passos máximo seja atingido.

3.3 Propriedades

Sejam

$$\sigma_{max}^2 = \max_{i \in \{1,\ldots,d\}} \sigma_i^2 \qquad \text{e} \qquad \sigma_{sum}^2 = \sum_{i=1}^{d} \sigma_i^2.$$

Para o Algoritmo 1 apresentado anteriormente, um limite superior para a ACP heteroscedástica dado por Zhang, Cai e Wu (2022) é dado pelo seguinte teorema.

Teorema 2. Considere o modelo de covariâncias *spike* generalizado na Equação 1.2, em que $\boldsymbol{X}$ e $\boldsymbol{\varepsilon}$ são sub-gaussianas, isto é,

$$\max_{q \geq 1, \|\boldsymbol{v}\|_2 = 1} q^{-\frac{1}{2}} \left(\mathbb{E}(|\boldsymbol{v}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{U}^T \boldsymbol{X}|^q) \right)^{\frac{1}{q}} \leq C$$

$$\text{e} \quad \max_{q \geq 1,\ i=1,\ldots,d} q^{-\frac{1}{2}} \left[\mathbb{E}\left(\left| \frac{\varepsilon_i}{\sigma_i} \right|^q \right) \right]^{\frac{1}{q}} \leq C.$$

Sejam $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ observações iid dadas pela Equação 1.2. Assuma que $n \geq Cr, \sigma_{sum}^2/\lambda_r(\boldsymbol{\Lambda}) \geq \exp(-Cn)$ e $\|\boldsymbol{\Lambda}\|/\lambda_r(\boldsymbol{\Lambda}) \leq C$ para uma constate $C > 0$. Então, existe uma constante $c_I > 0$ tal que a constante de incoerência (Equação 3.1) satisfaz $I(\boldsymbol{U}) \leq c_I \dfrac{d}{r}$. Então, a saída $\hat{\boldsymbol{U}}$ do algoritmo aplicada à matriz de covariâncias $\hat{\boldsymbol{\Sigma}}$ satisfaz

$$\mathbb{E}\Big(\| \operatorname{sen} \Theta(\hat{\boldsymbol{U}}, \boldsymbol{U}) \| \Big) \lesssim \frac{C}{\sqrt{n}} \left(\frac{\sigma_{sum} + r^{\frac{1}{2}} \sigma_{max}}{\lambda_r^{\frac{1}{2}}(\boldsymbol{\Lambda})} + \frac{\sigma_{sum} \sigma_{max}}{\lambda_r(\boldsymbol{\Lambda})} \right) \wedge 1 \tag{3.5}$$

em que $\| \operatorname{sen} \Theta(\cdot, \cdot) \|$ é a distância $\operatorname{sen} \Theta$ entre dois subespaços.

Observação 1. Seja $\tilde{d} = \sigma_{sum}^2/\sigma_{max}^2$. O limite superior (Equação 3.5) pode ser reescrito como

$$\mathbb{E}\Big(\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\|\Big) \lesssim \left(\sqrt{\frac{\tilde{d}\vee r}{n}}\frac{\sigma_{max}}{\lambda_r^{\frac{1}{2}}(\boldsymbol{\Lambda})} + \sqrt{\frac{\tilde{d}}{n}}\frac{\sigma_{max}^2}{\lambda_r(\boldsymbol{\Lambda})}\right)\wedge 1. \tag{3.6}$$

em que $\tilde{d}\vee r = \max(\tilde{d}, r)$.

Considere a ACP homoscedástica em que $\sigma_1^2 = \ldots = \sigma_d^2 = \sigma_{max}^2$. Um caso particular do Teorema 2 é tal que

$$\mathbb{E}\Big(\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\|\Big) \lesssim \sqrt{\frac{d}{n}}\left(\frac{\sigma_{max}}{\lambda_r^{\frac{1}{2}}(\boldsymbol{\Lambda})} + \frac{\sigma_{max}^2}{\lambda_r(\boldsymbol{\Lambda})}\right)\wedge 1, \tag{3.7}$$

E comparando a Equação 3.6 com a Equação 3.7, tem-se que a média entre $\tilde{d}\vee r$ e $\tilde{d}$ pode ser vista como a "dimensão efetiva" para a ACP heteroscedástica.

Considere a seguinte classe de matrizes de covariâncias *spike* generalizadas:

$$\begin{aligned}\mathscr{F}_{d,n,r,C}(\sigma_{sum},\sigma_{max},\nu) = \Bigg\{&\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T + \boldsymbol{D}:\\ &\boldsymbol{D}\text{ é diagonal não negativa},\quad \Sigma_i \boldsymbol{D}_{ii} \le \sigma_{sum}^2,\quad \max_i \boldsymbol{D}_{ii} \le \sigma_{max}^2,\\ &\boldsymbol{U}\in\mathbb{O}_{d,r},\quad I(\boldsymbol{U}) \le c_I\frac{d}{r},\quad \frac{\|\boldsymbol{\Lambda}\|}{\lambda_r(\boldsymbol{\Lambda})}\le C,\quad \lambda_r(\boldsymbol{\Lambda})\ge\nu\Bigg\}.\end{aligned} \tag{3.8}$$

Obtemos então um limite inferior para a ACP heteroscedástica para as matrizes de covariâncias em $\mathscr{F}_{d,n,r}(\sigma_{sum},\sigma_{max},\nu)$ no Teorema 3 enunciado em Zhang, Cai e Wu (2022).

Teorema 3. Supondo que $\sqrt{d}\sigma_{max} \ge \sigma_{sum} \ge \sigma_{max} > 0$. Existe uma constante $C > 0$, tal que se $d \ge Cr$, tem-se o seguinte limite inferior

$$\inf_{\hat{\boldsymbol{U}}}\sup_{\mathscr{F}_{d,n,r}(\sigma_{sum},\sigma_{max},\nu)}\mathbb{E}\Big(\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\|\Big) \gtrsim \frac{1}{\sqrt{n}}\left(\frac{\sigma_{sum}+r^{\frac{1}{2}}\sigma_{max}}{\nu^{\frac{1}{2}}} + \frac{\sigma_{sum}\sigma_{max}}{\nu}\right)\wedge 1. \tag{3.9}$$

Observação 2. Combinando o Teorema 2 com o Teorema 3, o Algoritmo 1 proposto atinge a taxa de erro estimada

$$\inf_{\hat{\boldsymbol{U}}}\sup_{\mathscr{F}_{d,n,r}(\sigma_{sum},\sigma_{max},\nu)}\mathbb{E}\Big(\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\|\Big) \asymp \frac{1}{\sqrt{n}}\left(\frac{\sigma_{sum}+r^{\frac{1}{2}}\sigma_{max}}{\nu^{\frac{1}{2}}} + \frac{\sigma_{sum}\sigma_{max}}{\nu}\right)\wedge 1.$$

Observação 3. Note que $\mathbb{E}(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T + \operatorname{diag}(\sigma_1,\ldots,\sigma_d)$ e $\mathbb{E}(\Delta(\hat{\boldsymbol{\Sigma}})) = \Delta(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)$. Então, quando o ruído heteroscedástico está presente, estimar $\boldsymbol{U}$ com base em $\hat{\boldsymbol{\Sigma}}$ ou em $\hat{\boldsymbol{\Sigma}}$ com as entradas da diagonal principal substituídas por 0 pode não ser ideal, pois não necessariamente as esperanças são iguais. É possível mostrar que tanto a DVS quanto a DVS com eliminação diagonal

$$\hat{\boldsymbol{U}}^{\text{DVS}} = \text{DVS}_r(\hat{\boldsymbol{\Sigma}}),\quad \hat{\boldsymbol{U}}^{DD} = \text{DVS}_r(\Delta(\hat{\boldsymbol{\Sigma}}))$$

podem ser inconsistentes mesmo quando o tamanho amostral n aumenta e a dimensão d é fixa pelo argumento do limite inferior a seguir proposto por Zhang, Cai e Wu (2022).

Proposição 1. Considere o modelo de covariâncias na Equação 1.2. Suponha que $\boldsymbol{\Sigma}_0 = \tilde{\boldsymbol{U}}\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}^T$, em que $\tilde{\boldsymbol{U}} = [\boldsymbol{U}\boldsymbol{U}_\perp]$ e $\boldsymbol{U}$ é o conjunto dos primeiros r vetores singulares. Assume-se que $\boldsymbol{X}$ e $\boldsymbol{\varepsilon}$ são sub-gaussianas. Assume-se que $n \geq Cr$, $n \wedge p \geq C(\sigma_{sum}^2/\sigma_r(\boldsymbol{\Lambda})$ e $\|\boldsymbol{\Lambda}\|/\lambda_r(\boldsymbol{\Lambda}) \leq C$ para alguma constante $C > 0$. Então, existe uma constante $c_I > 0$ tal que a constante de incoerência (Equação 3.1) $I(\boldsymbol{U}) = \max_i \frac{d}{r}\|\boldsymbol{e}_i^T\boldsymbol{U}\|_2^2$ satisfaz $I(\boldsymbol{U}) \leq c_I\frac{d}{r}$, então a saída $\hat{\boldsymbol{U}}$ do algoritmo para com uma matriz $\hat{\boldsymbol{\Sigma}}$ de posto r satisfaz

$$\mathbb{E}\Big(\|\operatorname{sen}\Theta(\hat{\boldsymbol{U}},\boldsymbol{U})\|\Big) \lesssim \left(\frac{\sigma_{sum}+\sqrt{r}\sigma_{max}}{n^{\frac{1}{2}}\lambda_r^{\frac{1}{2}}(\boldsymbol{\Lambda})} + \frac{\sigma_{sum}\sigma_{max}}{n^{\frac{1}{2}}\lambda_r(\boldsymbol{\Lambda})} + \frac{[(nd)^{\frac{1}{2}}+d]\lambda_{r+1}^{\frac{1}{2}}(\boldsymbol{\Lambda})}{n\lambda_r^{\frac{1}{2}}(\boldsymbol{\Lambda})} + \frac{\lambda_{r+1}(\boldsymbol{\Lambda})}{\lambda_r(\boldsymbol{\Lambda})}\right) \wedge 1.$$

A Proposição 1 mostra que se existe uma diferença expressiva entre $\lambda_r(\boldsymbol{\Sigma}_0)$ e $\lambda_{r+1}(\boldsymbol{\Sigma}_0)$, então o método HeteroPCA pode estimar bem a matriz $\boldsymbol{U}$.

As provas dos Teoremas 2 e 3 estão em Zhang, Cai e Wu (2022). Mais propriedades e outras situações em que o ruído heteroscedástico está presente são exploradas com mais detalhes no trabalho citado.

CAPÍTULO

4

RESULTADOS

Neste capítulo, serão apresentadas simulações em três situações. A primeira situação está relacionada com o Capítulo 2 e o objetivo é encontrar um subespaço gerado pela ACP tal que seja possível identificar a presença de observações atípicas. Na segunda situação explora-se o desempenho do método HeteroPCA apresentado no Capítulo 3, cujo objetivo é remover o viés da diagonal principal da matriz de covariâncias amostral. E na última situação explora-se a junção das situações anteriores, isto é, quando existem observações atípicas em um contexto de alta dimensão e ruído heteroscedástico, utiliza-se o método HeteroPCA e, posteriormente, procura-se identificar um subespaço capaz de identificar a presença de observações atípicas. Todas as simulações foram feitas utilizando o software R (R Core Team, 2020). Os gráficos foram feitos utilizando o pacote *ggplot* (WICKHAM, 2016).

4.1 Estudos de simulação das observações atípicas em alta dimensão

Nesta seção, dois estudos de simulação foram realizados nas condições do modelo da Equação 2.3. Neste caso, explora-se a situação em que se tem nove componentes base e apenas uma componente atípica. Sendo assim, o primeiro estudo de simulação destaca a situação em que observações atípicas estão presentes e a componente atípica é capturada pelos primeiros fatores, mas nenhuma delas representa a componente atípica de forma expressiva. Já o segundo estudo de simulação mostra como se comportam os fatores sem a presença das observações atípicas.

Foram geradas $n = 200$ observações com dimensão $d = 3000$ com base no modelo da Equação 2.3. Cada observação foi gerada de forma que $\{z_{i,j}\}_{1\leq i\leq d, 1\leq j\leq n}$ tem distribuição $N(0,1)$ e a base canônica, $\{\boldsymbol{e}_i\}_{1\leq i\leq d}$, foi usada como autovetores populacionais, $\{\boldsymbol{U}_i\}_{1\leq i\leq d}$ com $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_9$ sendo as direções principais e $\boldsymbol{e}_{10}$ sendo a direção atípica. Para as direções

principais, as variâncias foram assumidas tais que $\tau_{i,1} = 3000, 1000, 100, 90, 80, 70, 60, 50, 40$ para $i = 1, \ldots, 9$ respectivamente e, para a direção atípica $\boldsymbol{e}_{10}$, foi assumido que $\tau_{10,1} = 2000$ e $\tau_{10,2} = 1$, e além disso a proporção de observações atípicas escolhida foi de $p_{10} = 0,02$. Para as outras direções $\{\boldsymbol{U}_i\}_{11 \leq i \leq d}$ que correspondem a ruídos, assume-se que $\tau_{i,1} = 1$ e $p_i = 0$ e então gerou-se $\{\boldsymbol{X}_{i,j}\}_{1 \leq i \leq d, 1 \leq j \leq n}$. A simulação gerou seis observações atípicas.

Para esse conjunto de dados, foi calculada a matriz de covariâncias amostral e foi aplicada a ACP obtendo-se os autovalores e autovetores amostrais. Pode-se então examinar a contribuição de cada autovetor amostral em relação às direções de pico verdadeiras tomando o quadrado das entradas. A soma dos quadrados de cada entrada em cada autovetor é igual a 1 e então o quadrado da j-ésima entrada de $\hat{\boldsymbol{U}}_i$, $\hat{u}_{i,j}^2$, pode ser visto como a proporção explicada por $\boldsymbol{e}_i$ na direção $\hat{\boldsymbol{U}}_i$. A Tabela 1 mostra o quadrado das 12 primeiras entradas (nas linhas) dos primeiros 11 autovetores amostrais $\hat{\boldsymbol{U}}_1, \ldots, \hat{\boldsymbol{U}}_{11}$ (nas colunas). Além disso, as duas últimas linhas indicam os autovalores amostrais $\hat{\lambda}_i$ e o ângulo entre a verdadeira direção atípica $\boldsymbol{e}_{10}$ e $\hat{\boldsymbol{U}}_i, i = 1, \ldots, 11$. Se o maior valor, $\hat{u}_{i,j}^2$, destacado em cada $\hat{\boldsymbol{U}}_i$ é próximo de 1 e as outras entradas são próximas de 0, então a direção $\hat{\boldsymbol{U}}_i$ é uma boa estimativa para $\boldsymbol{e}_i$. Por exemplo, a primeira entrada de $\hat{\boldsymbol{U}}_1$ é aproximadamente igual a 1 e as outras entradas são aproximadamente iguais a 0, então a direção $\boldsymbol{e}_1$ é bem estimada por $\hat{\boldsymbol{U}}_1$. De forma similar, $\hat{\boldsymbol{U}}_2$ é uma boa estimativa de $\boldsymbol{e}_2$.

Tabela 1 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado. O maior valor em cada coluna está em destaque assim como a direção atípica (linha 10). Essa linha indica o quanto cada $\hat{\boldsymbol{U}}_i$ explica a direção atípica. As duas últimas linhas mostram os autovalores correspondentes aos autovetores $\{\hat{\boldsymbol{U}}_i\}_{1 \leq i \leq 11}$ e o ângulo entre $\hat{\boldsymbol{U}}_i$ e a verdadeira direção atípica $\boldsymbol{e}_{10}$.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	**0,994**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
2	0,000	**0,984**	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,000
3	0,000	0,000	**0,806**	0,007	0,022	0,005	0,003	0,004	0,025	0,003	0,000
4	0,000	0,000	0,023	**0,595**	0,121	0,060	0,002	0,004	0,012	0,034	0,000
5	0,000	0,000	0,018	0,089	**0,585**	0,031	0,016	0,073	0,010	0,007	0,000
6	0,000	0,000	0,005	0,081	0,034	**0,392**	0,276	0,032	0,000	0,000	0,000
7	0,000	0,001	0,016	0,010	0,018	0,087	**0,277**	0,077	**0,299**	0,001	0,000
8	0,000	0,000	0,002	0,011	0,004	0,009	0,038	**0,441**	0,237	0,026	0,000
9	0,000	0,000	0,008	0,044	0,002	0,017	0,001	0,018	0,001	**0,598**	0,000
10	0,000	0,000	0,005	0,025	0,047	0,221	0,198	0,131	0,162	0,005	0,000
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	3347,13	1094,48	121,36	110,59	96,93	84,84	79,56	72,56	60,79	48,98	23,21
Ângulo	90,43	89,30	93,85	80,96	102,57	118,06	63,60	68,78	66,29	85,83	89,43

Fonte: Elaborada pelo autor.

No entanto, pode-se perceber que nenhum dos autovetores $\hat{\boldsymbol{U}}_3, \ldots, \hat{\boldsymbol{U}}_{10}$ possui entradas próximas de 1. Ao invés disso, possuem diversas entradas diferentes de 0, indicando que cada uma dessas direções tem alguma correlação com diversas componentes. Vale destacar que

para cada linha $j = 3,\ldots,10$, a soma dos quadrados das j-ésimas entradas de $\hat{\boldsymbol{U}}_3,\ldots,\hat{\boldsymbol{U}}_{10}$ é próxima de 1. Isso mostra que cada autovetor verdadeiro $\boldsymbol{e}_3,\ldots,\boldsymbol{e}_{10}$ pode ser estimado por uma combinação linear de $\hat{\boldsymbol{U}}_3,\ldots,\hat{\boldsymbol{U}}_{10}$ ao invés de uma direção individual.

Em particular, é importante notar que nenhum dos 10 primeiros autovetores fornece boas estimativas para a direção atípica $\boldsymbol{e}_{10}$ (linha 10). Os ângulos na última linha também mostram que nenhum desses autovetores amostrais está próximo de $\boldsymbol{e}_{10}$ (isto é, ângulo próximo de 0). No entanto, a soma dos quadrados das 10 entradas nos 10 primeiros fatores, $\sum_{i=1}^{10}\hat{u}_{i,10}^2$, é aproximadamente igual a 0,79, indicando que $\boldsymbol{e}_{10}$ pode ser bem capturada pelo subespaço gerado pelos 10 primeiros autovalores amostrais. Sendo assim, apesar de a utilização dos fatores de forma individual não ser tão efetiva, esse subespaço mantém informações sobre as observações atípicas e, portanto, pode ser usado para identificar tais valores.

A Tabela 2 mostra os resultados da segunda simulação em que não existem observações atípicas. Neste caso, nota-se que o segundo autovalor e autovetor indicam de forma expressiva a direção considerada atípica na primeira situação (linha 10). Nota-se também que os três primeiros autovetores com entradas próximas a 1 são aqueles que representam as maiores variâncias e estimam bem as respectivas direções principais ($\boldsymbol{e}_1,\boldsymbol{e}_{10},\boldsymbol{e}_2$).

Tabela 2 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado. O maior valor em cada coluna está em destaque. As duas últimas linhas mostram os autovalores correspondentes aos autovetores $\{\hat{\boldsymbol{U}}_i\}_{1\leq i\leq 11}$ e o ângulo entre $\hat{\boldsymbol{U}}_i$ e a direção $\boldsymbol{e}_{10}$.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	**0,976**	0,010	0,009	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
2	0,009	0,000	**0,974**	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,000
3	0,000	0,000	0,000	0,031	**0,755**	0,001	0,010	0,056	0,001	0,000	0,000
4	0,000	0,001	0,000	**0,713**	0,038	0,052	0,002	0,004	0,044	0,009	0,000
5	0,000	0,000	0,000	0,075	0,001	**0,566**	0,043	0,113	0,003	0,019	0,000
6	0,001	0,000	0,000	0,004	0,049	0,097	0,000	**0,456**	0,140	0,034	0,000
7	0,000	0,000	0,001	0,001	0,006	0,079	**0,577**	0,012	0,107	0,000	0,001
8	0,000	0,000	0,001	0,048	0,002	0,032	0,145	0,055	**0,422**	0,060	0,000
9	0,000	0,000	0,000	0,000	0,008	0,001	0,007	0,090	0,022	**0,595**	0,000
10	0,010	**0,982**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,002
12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	3760,49	2236,33	1054,24	122,51	109,77	91,60	74,32	71,48	60,12	55,69	23,14
Ângulo	95,60	7,70	90,70	89,37	90,64	91,13	90,86	90,36	91,02	90,06	89,92

Fonte: Elaborada pelo autor.

4.2 Simulação do método HeteroPCA

Nesta seção foram realizados estudos de simulação para ilustrar o mérito do procedimento em estimar subespaços quando existe ruído heteroscedástico. As simulações são baseadas em médias de 500 experimentos independentes. As médias da distância $\operatorname{sen}\Theta$ são indicadas por

marcações e os desvios padrão são dados pelas barras verticais. Vale ressaltar que esses estudos são réplicas de alguns dos estudos presentes em Zhang, Cai e Wu (2022).

Primeiro, considera-se a ACP sob o modelo de covariâncias *spike* da Equação 1.2. Para diferentes valores de $d = 15, n = 60, 120, 180, \ldots, 600$ e $r = 3$ e 5, foi gerada uma matriz $\boldsymbol{U}_0$ de dimensão $d \times r$ com entradas iid normais padrão, $w_1, \ldots, w_d \overset{iid}{\sim} U(0,1)$ e $\sigma_1, \ldots, \sigma_d \overset{iid}{\sim} U(0,1)$. O intuito dos vetores $\boldsymbol{w}$ e $\boldsymbol{\sigma}$ é adicionar heteroscedasticidade às observações. Seja $\boldsymbol{U} = QR(\text{diag}(\boldsymbol{w})\boldsymbol{U}_0) \in \mathbb{O}_{d,r}$ e $\boldsymbol{\Sigma}_0 = \boldsymbol{U}\,\text{diag}\,(1,\ldots,d)\boldsymbol{U}^T \in \mathbb{R}^{d\times d}$. O objetivo é recuperar $\boldsymbol{U}$ com base nas observações iid $\{\boldsymbol{Y}_j = \boldsymbol{X}_j + \boldsymbol{\varepsilon}_j\}_{j=1}^n$, em que $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} N_d(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \overset{iid}{\sim} N_d(\boldsymbol{0}, \text{diag}(\sigma_1^2, \ldots, \sigma_d^2))$ e $d = 15$. Implementou-se os métodos HeteroPCA, DVS e DVS com eliminação diagonal e foram representados graficamente a média dos erros e o desvio padrão da distância senΘ. O resultado da simulação pode ser visto na Figura 1 no caso em que $r = 3$ e na Figura 2 para o caso em que $r = 5$.

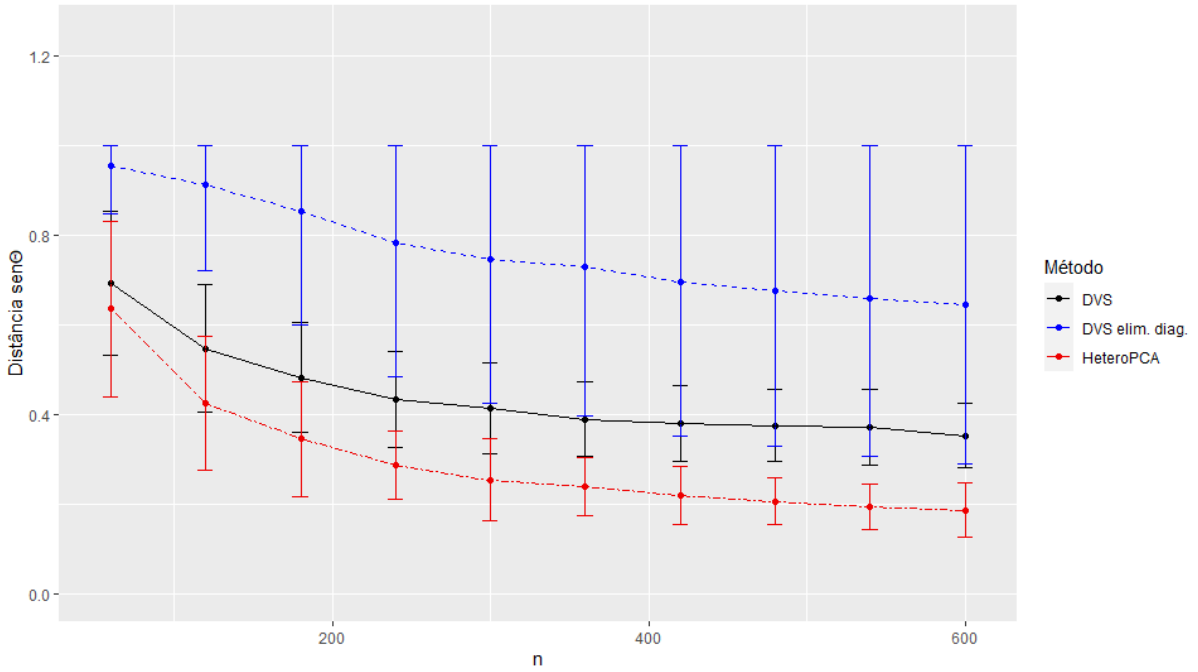

Figura 1 – Média e desvio padrão da distância senΘ para os métodos DVS, DVS com eliminação diagonal (elim. diag.) e HeteroPCA em função do tamanho da amostra com $d = 15$ e $r = 3$.

Fonte: Elaborada pelo autor.

Pode-se perceber que o estimador HeteroPCA tem um desempenho melhor que os outros métodos. A DVS, apesar de ter um comportamento parecido com o método HeteroPCA, possui o erro de estimação maior em ambas as situações (Figuras 1 e 2). Além disso, a DVS com eliminação diagonal tem um comportamento instável nos dois cenários.

Em seguida, é estudado como o grau de heteroscedasticidade afeta o desempenho dos métodos. Sejam

$$v_1, \ldots, v_d \overset{iid}{\sim} U(0,1) \quad \text{e} \quad \sigma_k^2 = \frac{0,1\; d\; v_k^\alpha}{\sum_{i=1}^d v_i^\alpha}, \quad k = 1, \ldots, d.$$

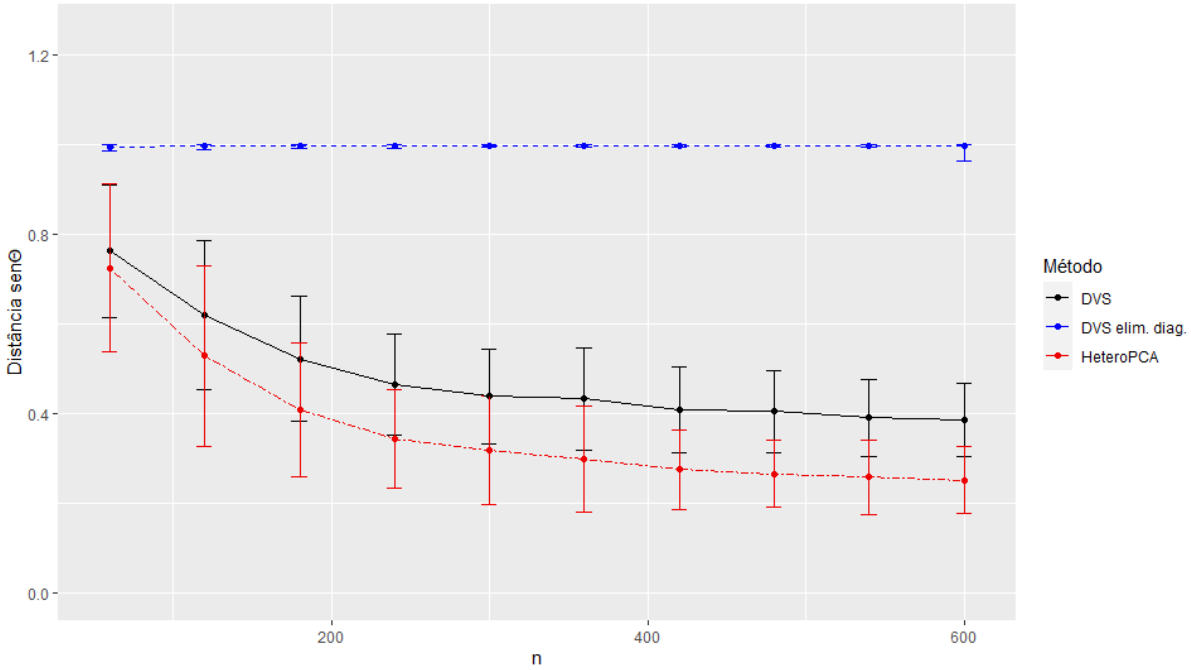

Figura 2 – Média e desvio padrão da distância sen Θ para os métodos DVS, DVS com eliminação diagonal (elim. diag.) e HeteroPCA em função do tamanho da amostra com $d = 15$ e $r = 5$.

Fonte: Elaborada pelo autor.

Neste caso, $\sigma^2_{sum} = \sigma^2_1 + \cdots + \sigma^2_d$ é sempre igual a $0,1d$ e α caracteriza o grau de heteroscedasticidade. Então, quanto maior for o valor de α, maior será o desbalanceamento da distribuição de $(\sigma_1, \ldots, \sigma_d)$. Em contrapartida, se $\alpha = 0$, então $\sigma_1 = \cdots = \sigma_d$ e tem-se o cenário homoscedástico. Gerou-se $\boldsymbol{U}, \boldsymbol{\Sigma}_0$ e $\{\boldsymbol{Y}_j, \boldsymbol{X}_j, \boldsymbol{\varepsilon}_j\}_{j=1}^n$ da mesma forma que no primeiro caso. A estimativa dos erros de $\boldsymbol{U}$ pode ser vista na Figura 3 para diferentes valores de $\alpha = 0, 1, 2, \ldots, 10$, $n = 30$ e $d = 50$. Na Figura 4 é apresentada a estimativa dos erros para os mesmos valores de α, mas com $n = 400$ e $d = 200$.

Os resultados novamente sugerem que o desempenho da DVS com eliminação diagonal é instável nos diferentes cenários. Além disso, quando $\alpha = 0$, ou seja, no caso homoscedástico, o desempenho do método HeteroPCA e da DVS são parecidos, indicando que o método HeteroPCA é geral. Porém, conforme α aumenta, o erro de estimação do método HeteroPCA aumenta de forma menor quando comparado à DVS.

4.3 Simulação de observações atípicas em alta dimensão e ruído heteroscedástico

A partir das simulações apresentadas na Seção 4.2, pode-se dizer que o método HeteroPCA tem resultados aceitáveis para estimar a matriz de covariâncias quando os dados possuem ruído heteroscedástico. Além disso, na Seção 4.1 é possível notar que a AF estimada pela ACP é

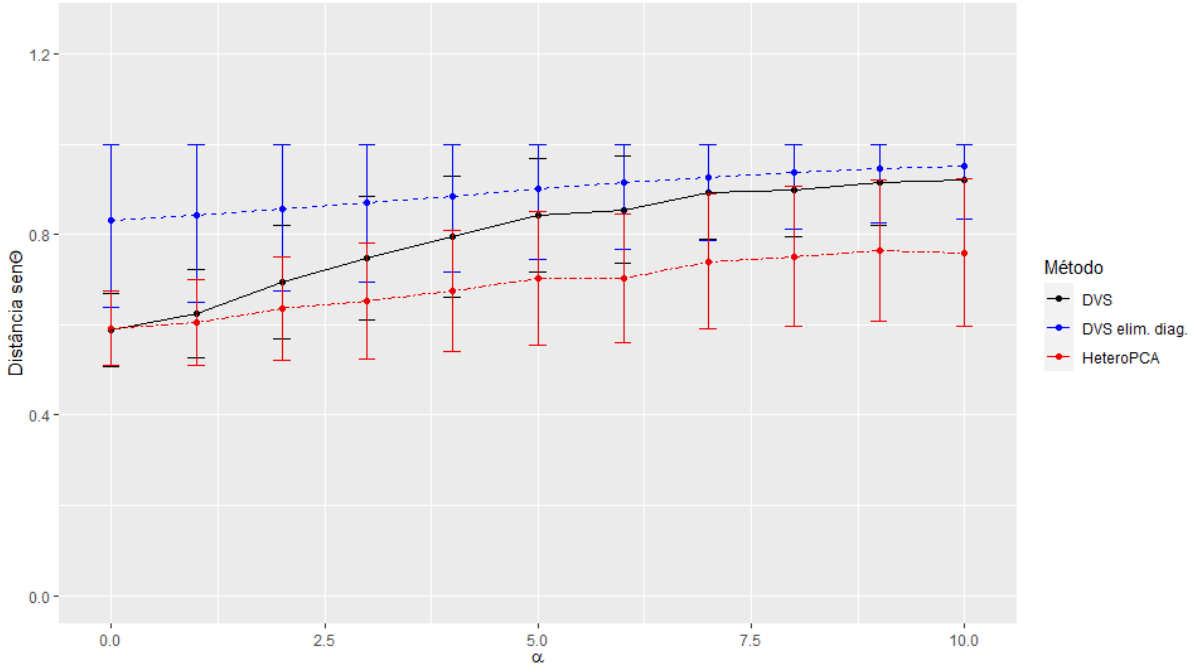

Figura 3 – Média e desvio padrão da distância sen Θ para os métodos DVS, DVS com eliminação diagonal (elim. diag.) e HeteroPCA em função do nível de heteroscedasticidade α com $n = 30, d = 50$ e $r = 5$.

Fonte: Elaborada pelo autor.

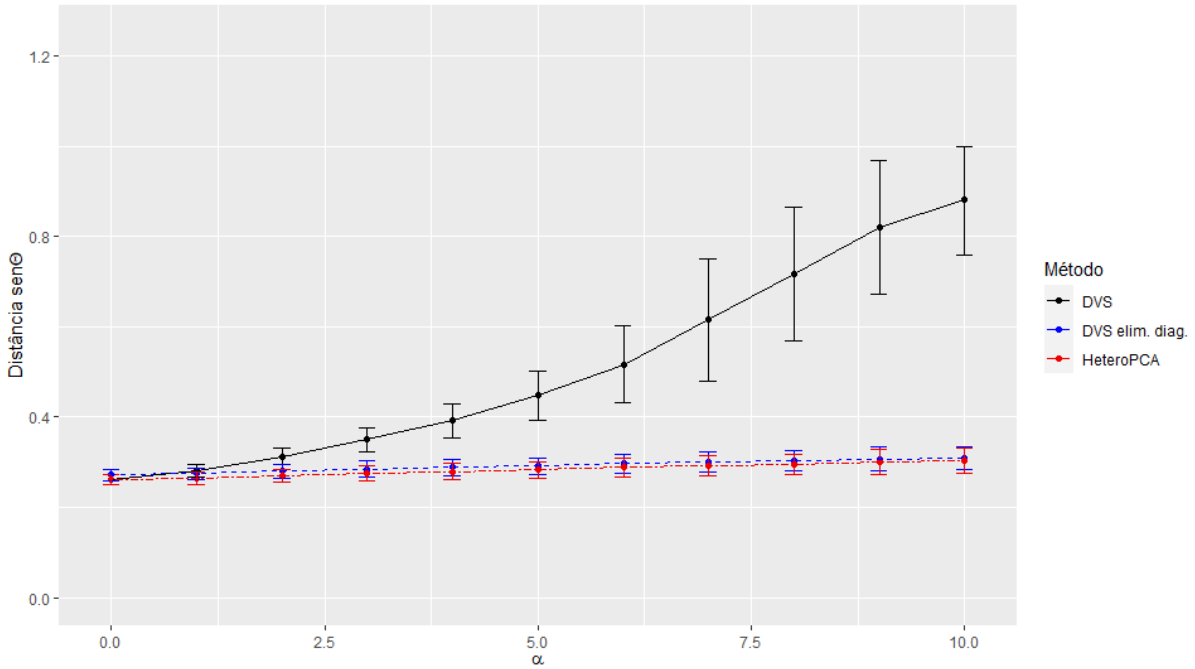

Figura 4 – Média e desvio padrão da distância sen Θ para os métodos DVS, DVS com eliminação diagonal (elim. diag.) e HeteroPCA em função do nível de heteroscedasticidade α com $n = 400, d = 200$ e $r = 5$.

Fonte: Elaborada pelo autor.

capaz de identificar um subespaço que mantém informações sobre as observações atípicas. Sendo assim, nesta seção propõe-se a junção das metodologias e o intuito das simulações apresentadas é verificar se é possível encontrar um subespaço capaz de identificar a presença de observações atípicas em alta dimensão na presença do ruído heteroscedástico. Cada simulação difere pela geração das entradas da matriz $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n}$, sendo que no primeiro estudo de simulação essa matriz é gerada como na Seção 4.1, isto é, tem distribuição $N(0,1)$. No segundo estudo de simulação, as entradas da matriz $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n}$ tem distribuição t_5. E no último estudo de simulação, as entradas da matriz $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n}$ tem distribuição normal contaminada.

4.3.1 Simulação 1

As observações foram geradas como na Seção 4.1, isto é, $\boldsymbol{X}$ é uma matriz cujas entradas $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n} \sim N(0,1)$, $n = 200$, $d = 2000$ e com direções principais tais que $\tau_{i,1} = 3000, 1000, 100, 90, 80,\ 70, 60, 50, 40$ para $i = 1, \ldots, 9$ e a direção atípica com $\tau_{10,1} = 2000, \tau_{10,2} = 1$ e a proporção de observações atípicas é $p_{10} = 0,10$. Além disso, seja o vetor $\boldsymbol{\sigma} \sim U(0,1)$ e consequentemente $\boldsymbol{\varepsilon}$ como na Seção 4.2 para induzir a heteroscedasticidade. Portanto, assim como na Equação 1.2, tem-se que $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$. Aplicou-se então o Algoritmo 1 para obter uma aproximação da matriz de covariâncias para dados com ruído heteroscedástico e posteriormente foi utilizada a ACP para encontrar um subespaço de dimensão baixa.

A Tabela 3 apresenta os autovetores amostrais nas colunas cujas entradas estão elevadas ao quadrado. Neste caso, a simulação gerou 20 observações atípicas e na aproximação da matriz de covariâncias com o método HeteroPCA utilizou-se $r = 10$. Além disso, a soma dos quadrados das 10 entradas nas 10 primeiras componentes, $\sum_{i=1}^{10} \hat{u}_{i,10}^2$, é aproximadamente $0,8976$. Isso indica que a direção $\boldsymbol{e}_{10}$ pode ser bem capturada por esse subespaço, porém, como mencionado na Seção 4.1, nenhum dos autovetores fornece uma boa estimativa para a direção atípica de forma individual.

4.3.2 Simulação 2

As observações foram simuladas como na Subseção 4.3.1, porém, neste caso, $\boldsymbol{X}$ é uma matriz cujas entradas $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n}$ seguem uma distribuição t de Student com 5 graus de liberdade, mantendo-se os mesmos valores para os parâmetros e seguindo o mesmo procedimento da Subseção 4.3.1.

A Tabela 4 apresenta os autovetores amostrais nas colunas cujas entradas estão elevadas ao quadrado. Neste caso, a simulação gerou 15 observações atípicas e na aproximação da matriz de covariâncias com o método HeteroPCA utilizou-se $r = 10$. Além disso, a soma dos quadrados das 10 entradas nas 10 primeiras componentes, $\sum_{i=1}^{10} \hat{u}_{i,10}^2$, é aproximadamente $0,9036$. Isso indica que a direção $\boldsymbol{e}_{10}$ pode ser bem capturada por esse subespaço, porém, como mencionado na Seção 4.1, nenhum dos autovetores fornece uma boa estimativa para a direção atípica de

Tabela 3 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado. O maior valor em cada coluna está em destaque assim como a direção atípica (linha 10). Essa linha indica o quanto cada $\hat{\boldsymbol{U}}_i$ explica a direção atípica. As duas últimas linhas mostram os autovalores correspondentes aos autovetores $\{\hat{\boldsymbol{U}}_i\}_{1\leq i\leq 11}$ e o ângulo entre $\hat{\boldsymbol{U}}_i$ e a verdadeira direção atípica $\boldsymbol{e}_{10}$.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	**0,981**	0,010	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000
2	0,010	**0,971**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3	0,000	0,000	0,002	**0,301**	0,359	0,062	0,009	0,057	0,002	0,011	0,000
4	0,001	0,000	0,019	0,286	**0,419**	0,061	0,000	0,005	0,022	0,001	0,000
5	0,000	0,002	0,000	0,092	0,014	0,164	**0,221**	0,126	0,109	0,008	0,000
6	0,000	0,000	0,003	0,069	0,005	**0,381**	0,300	0,000	0,012	0,003	0,000
7	0,000	0,000	0,002	0,022	0,003	0,087	0,038	**0,460**	0,099	0,013	0,000
8	0,000	0,000	0,007	0,043	0,002	0,021	0,107	0,047	**0,368**	0,113	0,000
9	0,000	0,000	0,007	0,003	0,002	0,001	0,068	0,009	0,086	**0,499**	0,000
10	0,000	0,000	**0,865**	0,005	0,003	0,012	0,001	0,001	0,010	0,001	0,000
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,001	0,000	0,001
12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,004
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	2557,02	1098,75	216,23	114,56	106,10	95,64	79,72	76,06	72,85	58,75	0,000
Ângulo	89,66	90,44	158,29	94,10	87,06	96,31	88,70	92,40	84,34	87,98	90,00

Fonte: Elaborada pelo autor.

forma individual.

4.3.3 Simulação 3

Nesta terceira simulação, $\boldsymbol{X}$ é uma matriz cujas entradas $\{\boldsymbol{X}_{ij}\}_{1\leq i\leq d, 1\leq j\leq n}$ seguem uma distribuição normal contaminada (TUKEY, 1960). Essa distribuição é uma combinação de duas distribuições normais de forma que a sua função densidade pode ser escrita da seguinte forma:

$$f(x) = (1-\beta)\varphi(x;\mu,\sigma) + \beta\varphi(x;\mu,\lambda\sigma),$$

em que $\varphi(x;\mu,\sigma)$ é a função densidade da distribuição normal com média μ e desvio padrão σ, β é a taxa de contaminação (usualmente $0 < \beta \leq 0,1$) e $\lambda > 1$ é o parâmetro que determina o desvio padrão da componente com maior variação. Neste caso, $\sigma = 1$, $\beta = 0,1$ e $\lambda = 50$ e os demais parâmetros são os mesmos das simulações anteriores.

A Tabela 5 apresenta os autovetores amostrais nas colunas cujas entradas estão elevadas ao quadrado. Neste caso, a simulação gerou 21 observações atípicas e na aproximação da matriz de covariâncias com o método HeteroPCA utilizou-se $r = 10$. Além disso, a soma dos quadrados das 10 entradas nas 10 primeiras componentes, $\sum_{i=1}^{10} \hat{u}_{i,10}^2$, é aproximadamente $0,8242$. Isso indica que a direção $\boldsymbol{e}_{10}$ pode ser bem capturada por esse subespaço, porém, como mencionado na Seção 4.1, nenhum dos autovetores fornece uma boa estimativa para a direção atípica de forma individual.

Tabela 4 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado. O maior valor em cada coluna está em destaque assim como a direção atípica (linha 10). Essa linha indica o quanto cada $\hat{\boldsymbol{U}}_i$ explica a direção atípica. As duas últimas linhas mostram os autovalores correspondentes aos autovetores $\{\hat{\boldsymbol{U}}_i\}_{1\leq i\leq 11}$ e o ângulo entre $\hat{\boldsymbol{U}}_i$ e a verdadeira direção atípica $\boldsymbol{e}_{10}$.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	**0,994**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
2	0,000	**0,978**	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,001	0,000
3	0,000	0,000	0,011	0,361	0,105	**0,369**	0,001	0,000	0,001	0,001	0,000
4	0,000	0,000	0,037	0,097	0,072	0,186	**0,420**	0,008	0,006	0,002	0,000
5	0,000	0,000	0,005	0,339	**0,424**	0,044	0,029	0,000	0,010	0,004	0,000
6	0,000	0,000	0,002	0,024	0,104	0,124	0,107	**0,389**	0,006	0,008	0,000
7	0,000	0,001	0,006	0,121	0,073	0,221	**0,227**	0,117	0,000	0,000	0,000
8	0,000	0,000	0,001	0,014	0,005	0,030	0,019	0,082	**0,543**	0,008	0,000
9	0,000	0,000	0,000	0,001	0,005	0,002	0,002	0,012	0,003	**0,607**	0,003
10	0,001	0,000	**0,850**	0,034	0,005	0,003	0,009	0,000	0,003	0,000	0,000
11	0,000	0,000	0,000	0,000	0,000	0,001	0,001	0,001	0,002	0,001	0,002
12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,017
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	5220,03	1317,88	280,02	208,94	171,91	150,30	134,14	98,92	89,79	76,19	0,00
Ângulo	91,36	90,24	22,99	79,30	85,96	93,11	95,37	89,95	93,19	89,87	89,63

Fonte: Elaborada pelo autor.

Tabela 5 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado. O maior valor em cada coluna está em destaque assim como a direção atípica (linha 10). Essa linha indica o quanto cada $\hat{\boldsymbol{U}}_i$ explica a direção atípica. As duas últimas linhas mostram os autovalores correspondentes aos autovetores $\{\hat{\boldsymbol{U}}_i\}_{1\leq i\leq 11}$ e o ângulo entre $\hat{\boldsymbol{U}}_i$ e a verdadeira direção atípica $\boldsymbol{e}_{10}$.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	**0,983**	0,010	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
2	0,009	**0,963**	0,001	0,002	0,000	0,000	0,000	0,009	0,000	0,000	0,000
3	0,000	0,000	0,433	**0,421**	0,003	0,032	0,025	0,000	0,000	0,000	0,000
4	0,000	0,000	0,010	0,034	0,010	**0,804**	0,012	0,000	0,003	0,000	0,000
5	0,002	0,003	**0,476**	0,405	0,000	0,001	0,023	0,001	0,003	0,001	0,001
6	0,000	0,000	0,000	0,000	0,091	0,000	0,079	0,000	**0,645**	0,000	0,001
7	0,001	0,001	0,001	0,036	0,006	0,029	**0,661**	0,011	0,090	0,000	0,002
8	0,000	0,001	0,002	0,007	**0,776**	0,005	0,028	0,000	0,061	0,000	0,000
9	0,000	0,000	0,002	0,001	0,000	0,000	0,000	0,000	0,00	**0,380**	0,001
10	0,000	0,009	0,001	0,004	0,000	0,001	0,009	**0,799**	0,001	0,000	0,001
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,002
12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	577952	187968	36224	27184	23269	20233	16560	13542	12987	4921	0,000
Ângulo	90,23	95,56	91,43	93,52	90,39	91,84	84,42	26,81	88,35	89,83	90,30

Fonte: Elaborada pelo autor.

CAPÍTULO

5

APLICAÇÃO

Neste capítulo o principal objetivo é aplicar o método proposto (Seção 4.3) a um conjunto de dados reais e verificar a sua efetividade. Os dados utilizados nesta aplicação são de expressão gênica (microarranjo de DNA) de câncer no sistema nervoso (ZHU; ONG; DASH, 2007). Nesse conjunto de dados, tem-se 60 observações divididas em um grupo controle (21 observações) e um grupo não controle (39 observações) e 7129 genes. Em um primeiro momento, foi feito um gráfico para verificar a variabilidade de cada gene (ou variável) de acordo com as respectivas classificações (controle e não controle). Na Figura 5 tem-se o desvio padrão para as observações do grupo "Controle" (círculos vermelhos) e do grupo "Não controle" (triângulos azuis). Em destaque, as variáveis (ou genes) 19 e 6396 são possíveis candidatas a serem características que diferenciam os grupos, pois a diferença entre o desvio padrão para cada uma dessas variáveis é expressiva.

Em seguida, o método proposto na Seção 4.3 foi aplicado para verificar se é possível determinar um subespaço capaz de identificar a presença de observações atípicas. Sendo assim, em um primeiro momento foi aplicado o método sem as observações consideradas atípicas. A Tabela 6 mostra o resultado obtido e pode-se observar que não se tem uma componente que explica a componente (ou direção) 19 ou 6396. Além disso, a soma aproximada dos valores nessas linhas para os 10 primeiros fatores são, respectivamente, 0,306 e 0,106. Ou seja, não parece que essas direções se destacam quando as observações atípicas não estão presentes.

No entanto, ao aplicar a metodologia ao conjunto de dados com as 21 observações do grupo "Controle" e 15 observações do grupo "Não controle" tem-se uma mudança nas linhas 19 e 6396. A Tabela 7 mostra o resultado e obtido e pode-se notar que, apesar de também não existir uma única componente (ou direção) que explique grande parte (valores próximos de 1) dessa variabilidade, tem-se que a soma aproximada dos valores para os 10 primeiros fatores são, respectivamente, 0,431 e 0,711. Pode-se dizer então que observações atípicas não são tão "influenciadas" pela componente 19, mas parecem ser mais "influencidas" pela componente

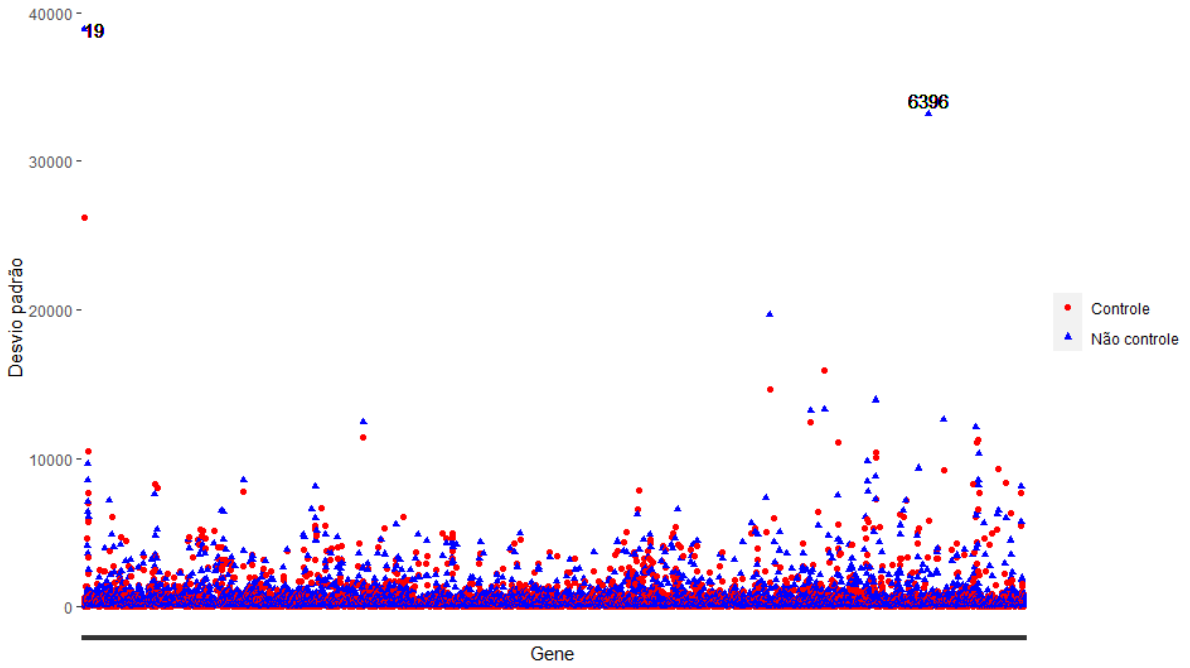

Figura 5 – Desvio padrão das variáveis do conjunto de dados separados pelos grupos definidos como "Controle" (círculo vermelho) e "Não controle" (triângulo azul).

Fonte: Elaborada pelo autor.

Tabela 6 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado para o conjunto de dados em que observações atípicas não estão presentes.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
18	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,051
19	0,223	0,004	0,006	0,002	0,038	0,005	0,003	0,002	0,000	0,005	0,093
20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,008
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6395	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
6396	0,001	0,006	0,002	0,000	0,004	0,000	0,023	0,002	0,015	0,050	0,000
6397	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

6396.

Tabela 7 – Primeiros 11 autovetores amostrais (colunas) com entradas ao quadrado para o conjunto de dados em que 15 observações atípicas estão presentes.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
18	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,227
19	0,306	0,044	0,003	0,001	0,009	0,002	0,006	0,029	0,003	0,029	0,006
20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6395	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
6396	0,182	0,028	0,159	0,252	0,015	0,016	0,012	0,021	0,004	0,020	0,030
6397	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

CAPÍTULO

6

DISCUSSÃO

Neste trabalho foi visto como a ACP é capaz de produzir um subespaço de dimensão baixa tal que as informações sobre os valores atípicos são mantidas. Apesar disso, existe uma limitação quanto ao uso individual das direções principais, isto é, utilizar apenas uma componente como indicativo de presença de valores atípicos pode não ser tão efetivo e consequentemente pode levar a conclusões inadequadas.

Além disso, foi apresentado um novo método chamado HeteroPCA com o intuito de lidar com dados que possuem ruído heteroscedástico. O método busca lidar com o viés na diagonal principal da matriz de covariâncias amostral causado pela heteroscedasticidade. Além disso, foram apresentadas duas situações em que o método HeteroPCA se destaca, sendo a primeira um caso simples (Seção 4.2, Figura 1) e a segunda levando em consideração graus de heteroscedasticidade (Seção 4.2, Figura 2).

Na Seção 4.3, foi proposto a junção do método HeteroPCA (Algoritmo 1) para aproximar a matriz de covariâncias para dados de alta dimensão que possuem ruído heteroscedástico com a AF estimada pela ACP para encontrar um subespaço capaz de identificar a presença de observações atípicas. Os resultados apresentados parecem ser satisfatórios para as diferentes situações exploradas. Sendo assim, pode-se dizer que essa junção tem potencial para ser utilizado em situações em que se deseja identificar possíveis componentes que influenciam observações atípicas em um contexto de alta dimensão e pequenas amostras com ruído heteroscedástico.

Na aplicação a dados reais, a metodologia proposta (Seção 4.3) foi aplicada na falta e na presença de observações consideradas atípicas. Foi possível notar um comportamento análogo ao das simulações, tal que, apesar de um fator por si só não explicar a possível componente que diferencia as observações atípicas, uma combinação pode ser capaz de identificar tal componente.

Sendo assim, a metodologia pode ser aplicada em situações análogas, isto é, quando o número de observações n é muito menor do que a dimensão d, como uma primeira análise, identificando possíveis variáveis e subespaços que diferenciam as possíveis observações atípicas.

Uma vez que essas variáveis e subespaços foram encontrados, pode-se utilizar essas informações como entradas para outros métodos de detecção de observações atípicas.

REFERÊNCIAS

ANDERSON, T. W. Asymptotic theory for principal component analysis. **Ann. Math. Statist.**, v. 34, p. 122–148, 1963. ISSN 0003-4851. Disponível em: <https://doi.org/10.1214/aoms/1177704248>. Citado nas páginas 22 e 24.

BAI, Z.; YAO, J. On sample eigenvalues in a generalized spiked population model. **J. Multivariate Anal.**, v. 106, p. 167–177, 2012. ISSN 0047-259X. Disponível em: <https://doi.org/10.1016/j.jmva.2011.10.009>. Citado nas páginas 22 e 23.

CANDÈS, E. J.; RECHT, B. Exact matrix completion via convex optimization. **Found. Comput. Math.**, v. 9, n. 6, p. 717–772, 2009. ISSN 1615-3375. Disponível em: <https://doi.org/10.1007/s10208-009-9045-5>. Citado na página 40.

CANDÈS, E. J.; SING-LONG, C. A.; TRZASKO, J. D. Unbiased risk estimates for singular value thresholding and spectral estimators. **IEEE Trans. Signal Process.**, v. 61, n. 19, p. 4643–4657, 2013. ISSN 1053-587X. Disponível em: <https://doi.org/10.1109/TSP.2013.2270464>. Citado na página 23.

CHOI, H. Y.; MARRON, J. Theory of high-dimensional outliers. **arXiv preprint arXiv:1909.02139**, 2019. Citado nas páginas 9, 11, 24, 27, 29, 34, 35 e 37.

DAVIS, C.; KAHAN, W. M. The rotation of eigenvectors by a perturbation. III. **SIAM J. Numer. Anal.**, v. 7, p. 1–46, 1970. ISSN 0036-1429. Disponível em: <https://doi.org/10.1137/0707001>. Citado na página 40.

FILZMOSER, P.; MARONNA, R.; WERNER, M. Outlier identification in high dimensions. **Comput. Statist. Data Anal.**, v. 52, n. 3, p. 1694–1711, 2008. ISSN 0167-9473. Disponível em: <https://doi.org/10.1016/j.csda.2007.05.018>. Citado na página 32.

FLORESCU, L.; PERKINS, W. Spectral thresholds in the bipartite stochastic block model. In: **Conference on Learning Theory**. [S.l.: s.n.], 2016. p. 943–959. Citado na página 23.

GIRSHICK, M. A. On the sampling theory of roots of determinantal equations. **Ann. Math. Statistics**, v. 10, p. 203–224, 1939. ISSN 0003-4851. Disponível em: <https://doi.org/10.1214/aoms/1177732180>. Citado na página 23.

GOLUB, G. H.; HOFFMAN, A.; STEWART, G. W. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. **Linear Algebra Appl.**, v. 88/89, p. 317–327, 1987. ISSN 0024-3795. Disponível em: <https://doi.org/10.1016/0024-3795(87)90114-5>. Citado na página 40.

HALL, P.; MARRON, J. S.; NEEMAN, A. Geometric representation of high dimension, low sample size data. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 67, n. 3, p. 427–444, 2005. Citado na página 30.

HAWKINS, D. M. **Identification of outliers**. [S.l.]: Chapman & Hall, London-New York, 1980. x+188 p. Monographs on Applied Probability and Statistics. ISBN 0-412-21900-X. Citado na página 21.

HONG, D.; BALZANO, L.; FESSLER, J. A. Towards a theoretical analysis of PCA for heteroscedastic data. In: **2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)**. Monticello, IL, USA: IEEE, 2016. p. 496–503. ISBN 9781509045501. Disponível em: <http://ieeexplore.ieee.org/document/7852272/>. Citado na página 23.

HONG, D.; FESSLER, J. A.; BALZANO, L. Optimally weighted pca for high-dimensional heteroscedastic data. **arXiv preprint arXiv:1810.12862**, 2018. Citado na página 23.

JOHNSTONE, I. M. On the distribution of the largest eigenvalue in principal components analysis. **Ann. Statist.**, v. 29, n. 2, p. 295–327, 2001. ISSN 0090-5364. Disponível em: <https://doi.org/10.1214/aos/1009210544>. Citado na página 22.

JUNG, S.; MARRON, J. S. PCA consistency in high dimension, low sample size context. **Ann. Statist.**, v. 37, n. 6B, p. 4104–4130, 2009. ISSN 0090-5364. Disponível em: <https://doi.org/10.1214/09-AOS709>. Citado na página 24.

LAWLEY, D. N. A modified method of estimation in factor analysis and some large sample results. In: **Uppsala Symposium on Psychological Factor Analysis, 17–19 March 1953**. [S.l.]: Ejnar Munksgaard, Copenhagen; Almqvist and Wiskell, Stockholm, 1953. p. 35–42. Citado na página 23.

MARCENKO, V. A.; PASTUR, L. A. DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. **Mathematics of the USSR-Sbornik**, v. 1, n. 4, p. 457–483, abr. 1967. ISSN 0025-5734. Disponível em: <http://stacks.iop.org/0025-5734/1/i=4/a=A01?key=crossref.1d0b803ddab02373cb6b0690a61e734a>. Citado nas páginas 24 e 34.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 45.

SALMON, J.; HARMANY, Z.; DELEDALLE, C.-A.; WILLETT, R. Poisson noise reduction with non-local PCA. **J. Math. Imaging Vision**, v. 48, n. 2, p. 279–294, 2014. ISSN 0924-9907. Disponível em: <https://doi.org/10.1007/s10851-013-0435-6>. Citado na página 23.

SHEN, D.; SHEN, H.; MARRON, J. S. A general framework for consistency of principal component analysis. **J. Mach. Learn. Res.**, v. 17, p. Paper No. 150, 34, 2016. ISSN 1532-4435. Disponível em: <https://doi.org/10.1145/2926791>. Citado nas páginas 24, 33, 34 e 35.

TUKEY, J. W. A survey of sampling from contaminated distributions. **Contributions to probability and statistics**, Stanford University Press, p. 448–485, 1960. Citado na página 52.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. Citado na página 45.

YAO, J.; ZHENG, S.; BAI, Z. **Large sample covariance matrices and high-dimensional data analysis**. [s.n.], 2015. OCLC: 908060156. ISBN 9781107588080. Disponível em: <http://site.ebrary.com/id/11038045>. Citado nas páginas 22 e 23.

YU, Y.; WANG, T.; SAMWORTH, R. J. A useful variant of the Davis-Kahan theorem for statisticians. **Biometrika**, v. 102, n. 2, p. 315–323, 2015. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/asv008>. Citado na página 40.

ZHANG, A. R.; CAI, T. T.; WU, Y. Heteroskedastic PCA: Algorithm, optimality, and applications. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 50, n. 1, p. 53 – 80, 2022. Disponível em: <https://doi.org/10.1214/21-AOS2074>. Citado nas páginas 9, 11, 23, 24, 39, 41, 42, 43 e 48.

ZHOU, Y.-H.; MARRON, J. Visualization of robust l1pca. **Stat**, Wiley Online Library, v. 5, n. 1, p. 173–184, 2016. Citado na página 30.

ZHU, Z.; ONG, Y.-S.; DASH, M. Markov blanket-embedded genetic algorithm for gene selection. **Pattern Recognition**, v. 40, n. 11, p. 3236–3248, 2007. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320307000945>. Citado na página 55.


ICMC
USP
SÃO CARLOS