JULIANA VIEIRA GOMES

MÉTODOS DE ESTIMAÇÃO DO DESVIO-PADRÃO PARA A PADRONIZAÇÃO DE VARIÁVEIS NA ANÁLISE DE COMPONENTES PRINCIPAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
G633m
2018

Gomes, Juliana Vieira, 1994-
Métodos de estimação do desvio-padrão para a padronização de variáveis na análise de componentes principais : . / Juliana Vieira Gomes. – Viçosa, MG, 2018.
ix, 48f. : il. ; 29 cm.

Orientador: José Ivo Ribeiro Júnior.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.46-48.

1. Matrizes (Matemática). 2. Desvios-padrão.
I. Universidade Federal de Viçosa. Departamento de Estatística. Mestrado em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 512.9434

JULIANA VIEIRA GOMES

MÉTODOS DE ESTIMAÇÃO DO DESVIO-PADRÃO PARA A PADRONIZAÇÃO DE VARIÁVEIS NA ANÁLISE DE COMPONENTES PRINCIPAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 26 de fevereiro de 2018.

Alexandre Navarro da Silva

Paulo César Emiliano

José Ivo Ribeiro Junior
(Orientador)

AGRADECIMENTOS

A elaboração deste trabalho não teria sido possível sem a colaboração, estímulo e empenho de diversas pessoas. Gostaria de expressar toda a minha gratidão e apreço a todos aqueles que, direta ou indiretamente, contribuíram para que esta tarefa se tornasse uma realidade. A todos quero manifestar os meus sinceros agradecimentos.

Em primeiro lugar, agradeço acima de tudo, a Deus, por iluminar meus caminhos ofertando a força necessária para a conclusão de mais um objetivo, para a realização de mais um sonho.

A Capes e a Fapemig, pela concessão da bolsa de estudos, durante o período de realização deste trabalho.

Ao meu pai, minha mãe, minhas irmãs e amigos tenho um agradecimento muito especial porque acreditaram em mim desde o primeiro instante. Sou quem sou porque vocês estiveram e estão sempre ao meu lado. Agradeço também ao meu namorado Samuel, por todo o seu amor, carinho, admiração, e pela presença incansável com que me apoiou ao longo do período de elaboração desse trabalho.

Não posso deixar de agradecer ao meu orientador, Professor Doutor José Ivo, por toda a paciência, empenho com que sempre me orientou neste trabalho. Muito obrigada por me ter corrigido quando necessário sem nunca me desmotivar.

A todos, obrigada por permitirem que esta conquista seja uma realidade.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

GOMES, Juliana Vieira, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Métodos de estimação do desvio-padrão para a padronização de variáveis na análise de componentes principais.** Orientador: José Ivo Ribeiro Júnior.

Este trabalho propôs avaliar a eficiência de diferentes matrizes de covariâncias sobre as estimativas dos componentes principais (CP), de acordo com diferentes métodos de estimação do desvio-padrão utilizado na padronização da variável. Além disso, procurou também, determinar a importância relativa de cada variável aleatória avaliada, normal ou não, que fez parte da composição do CP. A estimação do desvio-padrão foi feita de acordo com quatro métodos: desvio-padrão total, Lenth, Juan e Pena e Dong. Para isso, foram simulados 60 conjuntos de dados compostos por quatro variáveis aleatórias com 10.000 observações cada, com três diferentes graus de correlação, dois tipos de médias, dois tipos de variâncias e cinco percentuais de *outliers*. Os *outliers* foram adicionados com o intuito de quebrar a aleatoriedade das variáveis. De acordo com os resultados, o fator mais importante em afetar a qualidade da estimativa do desvio-padrão foi a proporção de *outliers*. Nesse sentido, o melhor método de estimação foi o de Lenth para até 2% de *outliers*. A matriz que forneceu os melhores resultados para a análise dos CPs, foi a que utilizou a estimativa do desvio-padrão obtida pelo método do desvio-padrão total, na ausência de *outliers*, com variâncias iguais e diferentes. Já para o conjunto de dados com *outliers* e variâncias iguais e diferentes, a matriz baseada no método de Lenth forneceu resultados mias satisfatórios para a análise de CPs.

ABSTRACT

GOMES, Juliana Vieira, M.Sc., Universidade Federal de Viçosa, February, 2018. **Estimation methods of the standard deviation for the standardization of variables on principal components analysis.** Adviser: José Ivo Ribeiro Júnior.

This study proposed to evaluate the efficiency of different covariance matrices on the principal components (PC) estimates, according to different estimation methods of the standard deviation used in the standardization of the variable. Furthermore, this project also sought to determine the relative importance of each random variable evaluated, whether normal or not, that composed the PC. The estimation of the standard deviation was done according to four methods: total standard deviation, Lenth, Juan and Pena and Dong. For this purpose, 60 datasets were simulated containing four random variables with 10.000 observations each with three different degrees of correlation, two types of mean, two types of variances and five percentages of outliers. The outliers were added with the aim to break the randomness of the variables. According to the results, the most import fact that affects the quality of the standard deviation estimation was the proportion of outliers. In this regard, the best method of estimation was the Lenth one for up to 2% of outliers. The matrix that provided the best results for the PCs analysis was the one that utilized the estimative of the standard deviation obtained by the total standard deviation method, in the absence of outliers, with equal and different variances. As for the dataset with outliers using equal and different variances, the matrix obtained through the Lenth method provided more satisfactory results for the PCs analysis.

1 INTRODUÇÃO

Atualmente, é evidente a facilidade e rapidez com que se analisa uma base com dados de grande dimensão e complexidade, a partir de muitos modelos estatísticos univariados e multivariados.

Consequentemente, durante os últimos anos tem-se verificado um crescimento substancial na produção e no armazenamento de dados que, em grande parte, são inviáveis de serem analisados através de métodos manuais e tradicionais.

De acordo com Reis (1997), a estatística multivariada trata de um conjunto de métodos estatísticos que permite a análise simultânea de medidas múltiplas para cada indivíduo ou objeto em análise. A necessidade de analisar as variáveis simultaneamente implica em procurar métodos que permitam reduzir a dimensionalidade sem perda significativa da informação contida nos dados.

De acordo com Kubrusly (2001), para se estabelecer um índice que possibilite ordenar um conjunto de n objetos, segundo critério definido por um conjunto de m variáveis adequadas, é necessário escolher os pesos ou ponderações das variáveis de tal forma que traduzam a informação contida na variável. Para a construção de um índice como combinação linear de variáveis, é desejável que este contenha o máximo de informação fornecida pelo conjunto de variáveis selecionadas. Um método que cria combinações lineares com máxima variância é a análise de componentes principais (CP) (SANDANIELO, 2008).

Ele é um método utilizado em diversas áreas do conhecimento, como por exemplo: agronomia, fitotecnia, zootecnia, ecologia, biologia, psicologia, medicina, engenharia florestal, etc. Foi introduzido por Karl Pearson em 1901 e fundamentado no artigo de Hotelling de 1933. Seu principal objetivo é de explicar a estrutura de variâncias e covariâncias de uma matriz composta por n indivíduos e p variáveis (MINGOTI, 2007). Neste sentido, o que se deseja é gerar combinações lineares das p variáveis que maximizam a variância total dessa matriz.

A análise dos CPs consiste em um método multivariado que trabalha em uma estrutura de covariâncias da matriz nxp, cujos p componentes

principais CP_1, CP_2, ..., CP_p são não correlacionados e estão nas direções das maiores variabilidades, a fim de descrever variação dos dados (MANLY, 2008).

Para obter os CP's, é necessário realizar a decomposição da matriz de covariâncias entre as p variáveis ($\sum$). Em geral, os pesquisadores preferem trabalhar com as variáveis em escala padronizada, ou seja, transformadas com médias e variâncias iguais a zero e um, respectivamente, que equivale à matriz de covariâncias das variáveis padronizadas ou à matriz de correlações das variáveis originais (ρ). Todavia, o uso ou não das transformações nos dados fornece diferentes estimativas dos CP's, o que pode alterar a classificação final dos elementos amostrais, uma vez que eles não são invariantes à mudança de escala (JOHNSON e WICHERN, 2002).

A análise de componentes principais tem como principais vantagens: retirar a multicolinearidade das variáveis, pois permite transformar um conjunto de variáveis originais intercorrelacionadas em um novo conjunto de variáveis não correlacionadas. Além disso, reduz muitas variáveis a eixos que representam algumas variáveis, sendo estes eixos perpendiculares (ortogonais) explicando a variação dos dados de forma decrescente e independente (HONGYU, 2015).

As desvantagens são: a sensibilidade a outliers, não recomendada quando se tem duplas ausências (muitos zeros na matriz) e dados ausentes. (HONGYU, 2015). A invariância à escala, também é uma desvantagem da análise dos CPs. A falta da invariância à escala ocorre devido à possibilidade de utilizar a matriz de covariâncias com base em diferentes tipos de variáveis. Consequentemente, a escolha menos adequada da escala (original ou padronizada) poderá proporcionar interpretações baseadas nos CPs menos adequadas.

Para a análise exploratória dos CPs, não há a exigência da pressuposição de que a matriz nxp das variáveis originais seja composta por variáveis aleatórias, e nem necessariamente, por variáveis aleatórias normais. Desse modo, diferentes combinações nessa matriz, de variáveis não aleatórias (não seguem distribuição de probabilidades), de aleatórias normais e de outros tipos de variáveis aleatórias poderão ocorrer. Isso pode implicar em estudos que necessitam em conhecer as diferentes importâncias relativas das variáveis avaliadas.

De acordo com Ferreira (2009), o desenvolvimento dos CPs não requer pressuposições de normalidade multivariada, mas possuem interpretações úteis em termos da constante elipsoide de densidade, se a normalidade existir. Diz ainda Jolliffe (2002), que as variáveis podem ser aleatórias contínuas e distribuídas normalmente ou não.

Quando se utiliza a matriz ρ, todas as variáveis se tornam igualmente importantes, o que pode gerar perda de informação, caso elas não se comportarem de tal maneira. Campana et al (2010) apresentaram uma proposta de matriz para a análise dos CPs, com base nos coeficientes de variação, a fim de, se necessário, possibilitar diferentes variabilidades relativas para as variáveis analisadas.

Na linha desse estudo, o presente trabalho também busca, por meio de uma estimativa da variação aleatória presente em cada uma das variáveis originais, proporcionar também, uma matriz a ser usada na análise dos CPs que não tenha os problemas de influência das escalas (matriz $\sum$) e nem das igualdades das importâncias relativas (matriz ρ).

2 OBJETIVOS

Geral

Verificar os efeitos de diferentes métodos para a estimativa da variabilidade aleatória de cada variável avaliada, independentemente se ela é aleatória normal ou não, sobre as estimativas dos componentes principais.

Específicos

Avaliar os efeitos dos métodos em detectar os valores aleatórios presentes no conjunto de dados, principalmente, quando as variáveis não são aleatórias.

Determinar a importância relativa de cada variável avaliada, aleatória normal ou não, que faz parte da composição do CP.

Avaliar as eficiências de diferentes matrizes de covariâncias sobre as estimativas dos componentes principais, de acordo com a estimativa do desvio-padrão, proveniente de um determinado método de estimação.

Divulgar métodos promissores para estabelecer novas maneiras de obter variáveis padronizadas para a análise dos CPs.

3 REVISÃO DE LITERATURA

3.1 Componentes Principais

A análise de componentes principais é um método que foi idealizado, de início, por Karl Pearson, em 1901, e fundamentado, em 1933, por Hotelling (HOTELLING, 1933). Ela constitui um método multivariado, e seu objetivo está relacionado com a explicação de uma estrutura de covariâncias de uma matriz, composta de n elementos amostrais e p variáveis, aleatórias ou não, para a obtenção dos componentes principais.

Cada componente principal (CP) é uma combinação linear das variáveis originais ou padronizadas, independentes ou não correlacionados entre si e estimados com o propósito de preservar, em ordem de estimação, o máximo da informação em termos da variação total contida nos dados.

A informação contida nas p variáveis observadas é substituída pela informação contida nos k (k < p) CPs não correlacionados. O ideal seria a distribuição de probabilidades da matriz em estudo (**Y)** ser normal p-variada e os CPs, além de não correlacionados (independentes) apresentarem distribuição normal k-variada. Contudo, para a utilização dos CPs, não é necessário que o conjunto de variáveis analisadas seja aleatório e nem que tenha distribuição normal, o que na prática, se transforma em uma vantagem. Por outro lado, implica em diferenças nas relações entre as variáveis e os CPs (MINGOTI, 2007).

Algebricamente, os CPs representam combinações lineares de p variáveis originais $Y_1, Y_2, \dots, Y_P$ ou de p variáveis padronizadas $Z_1, Z_2, \dots, Z_P$. De forma geométrica, essas combinações lineares representam a escolha de novos eixos coordenados, os quais são obtidos por rotações do sistema de eixos originais, representados pelas p variáveis, sendo que esses novos eixos representam as direções de máxima variabilidade.

Os CPs dependem somente da matriz de covariâncias $\sum$ ou da matriz de correlações ρ, independentemente da aleatoriedade ou não das variáveis estudadas e dos seus graus de relação.

3.2 Matriz de Covariâncias

Considere as variáveis $Y_1, Y_2, \dots, Y_P$ com vetor px1 de médias $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_p]'$ e matriz (pxp) de covariâncias $\sum$. Denota-se $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P$, como os autovalores da matriz $\sum$ e com os respectivos autovetores normalizados $\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_p}$, satisfazendo as seguintes condições:

(i) $\boldsymbol{e_j'e_{j'}} = 0, j \neq j'$;

(ii) $\boldsymbol{e_j'e_j} = 1, j = 1, 2, \dots, k, \dots, p$;

(iii) $\sum \boldsymbol{e_j} = \lambda_j \boldsymbol{e_j},\ j = 1, 2, \dots, k, \dots, p$, em que:

$\boldsymbol{e_j} = [e_{j1}, e_{j2}, \dots, e_{jp}]'$.

Os componentes principais CP_1, CP_2, ..., CP_k, ..., CP_p, são as combinações dadas por:

$$CP_1 = \boldsymbol{e_1'}\,\boldsymbol{Y} = e_{11}Y_1 + e_{12}Y_2 + \cdots + e_{1p}Y_p;$$

$$CP_2 = \boldsymbol{e_2'}\,\boldsymbol{Y} = e_{21}Y_1 + e_{22}Y_2 + \cdots + e_{2p}Y_p;$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots$$

$$CP_{\mathrm{k}} = \boldsymbol{e_k'}\,\boldsymbol{Y} = e_{k1}Y_1 + e_{k2}Y_2 + \cdots + e_{kp}Y_p;$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots$$

$$CP_{\mathrm{p}} = \boldsymbol{e_p'}\,\boldsymbol{Y} = e_{p1}Y_1 + e_{p2}Y_2 + \cdots + e_{pp}Y_p, \text{ em que:}$$

$\boldsymbol{Y} = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_p \end{bmatrix}$ = matriz (nxp) dos valores das variáveis originais.

Pode-se verificar que $Var(CP_{\mathrm{j}}) = \boldsymbol{e_j'}\sum \boldsymbol{e_j} = \lambda_j,\ j = 1, 2, \dots, p$ e $Cov(CP_{\mathrm{j}}, CP_{j'}) = \boldsymbol{e_j'}\sum \boldsymbol{e_{j'}} = 0,\ j \neq j'$. Isso indica que os componentes principais são não correlacionados e têm variâncias iguais aos autovalores de $\sum$.

Segundo Reis (1997), para maximizar uma função de várias variáveis, deve-se utilizar o método dos multiplicadores de Lagrange, que resulta na equação característica de $\sum$:

$$\left[\sum - \lambda_j \boldsymbol{I}\right]\boldsymbol{a_j} = \boldsymbol{0}, \text{ em que:}$$

$$\sum = \begin{bmatrix} \sigma_{Y_1Y_1} & \sigma_{Y_1Y_2} & \dots & \sigma_{Y_1Y_p} \\ \sigma_{Y_1Y_2} & \sigma_{Y_2Y_2} & \dots & \sigma_{Y_2Y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Y_1Y_p} & \sigma_{Y_2Y_p} & \dots & \sigma_{Y_pY_p} \end{bmatrix}$$

$\boldsymbol{e_j} = \frac{1}{||\boldsymbol{a_j}||}\boldsymbol{a_j}$;

$\boldsymbol{a_j} = [a_{j1} \quad a_{j2} \quad \dots \quad a_{jp}]$ = autovetor não normalizado de $\sum$;

$\boldsymbol{e_j} = [e_{j1} \quad e_{j2} \quad \dots \quad e_{jp}]$ = autovetor normalizado de $\sum$.

Para que essa equação tenha solução quando $\boldsymbol{a_j}$ não for nulo, a matriz $[\sum - \lambda_j \boldsymbol{I}]$ deve ser singular, ou seja, seu determinante deve ser nulo. Então, existirá uma solução não nula para $\boldsymbol{a_j}$ se, e somente se, λ_j for um autovalor de $\sum$.

Uma vez que a maximização ocorre do primeiro componente principal CP_1, em ordem decrescente, λ_1 é o autovalor a ser escolhido. De forma análoga, obtém-se o segundo componente. Considerando-se que $\boldsymbol{e_1' e_2} = 0$, CP_1 e CP_2 devem ser não correlacionados, isto é:

$$Cov(CP_1, CP_2) = Cov(\boldsymbol{e_1' Y}, \boldsymbol{e_2' Y}) = \boldsymbol{e_1' \sum e_2} = 0.$$

Dessa maneira, o segundo componente principal CP_2 é obtido de $\boldsymbol{e_2}$, autovetor normalizado associado ao segundo maior autovalor. Em termos gerais, o j-ésimo componente principal será obtido a partir do j-ésimo autovetor normalizado associado ao j-ésimo maior autovalor de $\sum$ (REIS, 1997).

Utilizando-se algumas propriedades matriciais, tem-se que:

$$\sum_{w=1}^{p} Var(Y_w) = \sum_{j=1}^{p} Var(CP_j);$$

$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_P$, em que:

σ_{ww} = variância da variável Y_w, para w = 1, 2, ..., p variáveis e j = 1, 2, ..., k, ..., p CPs.

De acordo com o teorema da decomposição espectral, pode-se expressar a inversa de uma matriz quadrada em termos de seus autovalores e autovetores, isto é, $\sum = \boldsymbol{P \Lambda P'}$, em que $\boldsymbol{\Lambda}$ é a matriz diagonal de autovalores e $\mathbf{P} = [\boldsymbol{e_1, e_2, \ldots, e_p}]$. Dessa forma, tem-se:

$$tr(\boldsymbol{P \Lambda P'}) = tr(\boldsymbol{\Lambda P' P}) = tr(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + \cdots + \lambda_P.$$

Assim, a soma das variâncias das variáveis originais é igual à soma das variâncias de todos os CPs, pois:

$$\sum_{w=1}^{p} Var(Y_w) = tr(\sum) = tr(\boldsymbol{\Lambda}) = \sum_{j=1}^{p} Var(CP_j) = \sum_{j=1}^{p} \lambda_j.$$

Conclui-se então que o k-ésimo CP explica, do total da variância das variáveis originais:

$$\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} \times 100.$$

E ainda, que os primeiros k CPs explicam, da variância total:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100.$$

Em geral, para interpretar os dados com sucesso, basta escolher os primeiros k CPs que envolvam pelo menos 70% da variância total. Isto é, basta escolher CP_1, CP_2, ..., CP_k, tal que:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100 \geq 70\%, \text{em que k} < \text{p}.$$

Conforme Kayser e Tenke (2003), o número máximo de componentes principais que podem ser extraídos de uma matriz de dados de casos por variáveis é determinado pelo menor número de linhas (casos) ou colunas (variáveis), embora menos componentes possam ser suficientes para explicar completamente a variância de uma matriz linearmente dependente.

Cada elemento do autovetor $\boldsymbol{e_j'} = [e_{j1}, \dots, e_{jk}, \dots, e_{jp}]$ também pode informar sobre a importância de cada variável para o j-ésimo CP, por meio de suas magnitudes absolutas. Porém, esses elementos são influenciados pela escala das variáveis. Para contornar este problema, pode-se utilizar uma medida de associação que não depende da escala das variáveis originais, que é o coeficiente de correlação entre CP_j e Y_w, como segue (FERREIRA, 2009):

$$\rho_{CP_j Y_w} = \frac{e_{jw}\sqrt{\lambda_j}}{\sqrt{\sigma_{ww}}}.$$

3.3 Matriz de Correlações

O método de obtenção dos CPs através da matriz de covariâncias $\sum$ não é útil quando existem grandes diferenças nas escalas das variáveis analisadas. Uma saída neste caso é proceder à transformação nos dados das variáveis originais. A transformação mais usada é a padronização da média e do desvio padrão das variáveis. Isso é o equivalente a obterem-se os CPs através da matriz de correlações ρ das variáveis originais Y_1, Y_2, ..., Y_p.

Seja $Z_w = \frac{Y_w - \mu_w}{\sqrt{\sigma_{Y_w Y_w}}}$ a variável Y_w padronizada.

Os CP's baseados nas variáveis padronizadas Zs são:

$$CP_{\text{j}} = e_{j1}Z_1 + e_{j2}Z_2 + \cdots + e_{jp}Z_p, \text{para j} = 1, 2, \dots, \text{k}, \dots, \text{p}.$$

Dessa vez, os p autovalores são estimados por meio do determinante da expressão baseada na matriz (pxp) das correlações $\boldsymbol{\rho}$ entre as variáveis

Ys, que é a mesma matriz (pxp) das covariâncias entre as variáveis padronizadas Zs:

$[\boldsymbol{\rho} - \lambda_j \boldsymbol{I}] = \mathbf{0}$, em que:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \sigma_{z_1 z_2} & \dots & \sigma_{z_1 z_p} \\ \sigma_{z_1 z_2} & 1 & \dots & \sigma_{z_2 z_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{z_1 z_p} & \sigma_{z_2 z_p} & \dots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho_{Y_1 Y_2} & \dots & \rho_{Y_1 Y_p} \\ \rho_{Y_2 Y_1} & 1 & \dots & \rho_{Y_2 Y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{Y_p Y_1} & \rho_{Y_p Y_2} & \dots & 1 \end{bmatrix}.$$

Ainda que os autovalores sejam diferentes quando baseados em $\sum$ e ρ, por estarem captando informações diferentes para ambos os casos, têm-se (JOHNSON; WICHERN, 2002; FERREIRA, 2009):

$$Var(CP_j) = \lambda_j, com\, Var(CP_1) \geq Var(CP_2) \geq \ldots \geq Var(CP_p);$$

$$Cov(CP_1, CP_2) = Cov(CP_1, CP_p) = \cdots = Cov(CP_{p-1}, CP_p) = 0.$$

Dessa forma, o CP_1 contém mais informações sobre os dados do que o CP_2, que não contém informações do CP_1, e assim sucessivamente.

Nesse caso, a importância ou influência que cada variável Y_w exerce sobre o CP_j, é dada pela correlação entre cada variável padronizada Z_w e o componente CP_j, que está sendo interpretado:

$$\rho_{CP_j, Z_w} = e_{jw}\sqrt{\lambda_j}.$$

3.4 Estimação do Desvio-padrão

Se $Y_1, Y_2, \ldots, Y_p$ constitui um vetor de variáveis, a escala, aproximadamente, de -3 a 3, para todas as variáveis padronizadas, contempla quase que toda a variação presente nas variáveis originais Ys, independentemente das respectivas magnitudes, quando a padronização é realizada a partir do desvio-padrão estimado em função de todas as observações de cada Y, separadamente.

Isso significa nessa situação, que a análise dos CPs incorpora apenas as correlações entre os Ys.

Por outro lado, se o vetor das variáveis não for composto somente por variáveis aleatórias, a variabilidade, não aleatória e não incorporada pela matriz $\boldsymbol{\rho}$, passará a ser importante.

Na análise dos CPs, é preferível utilizar a matriz $\boldsymbol{\rho}$ do que a matriz $\sum$. Como os dois resultados são diferentes, dado que os CPs não são invariantes

à escala, na maioria das vezes prefere-se dar a mesma importância a todas as variáveis Ys do que incorporar na análise as diferentes e respectivas importâncias relativas dessas variáveis, quando existirem.

Como já mencionado, a matriz $\boldsymbol{\rho}$ é obtida por meio das covariâncias entre as variáveis padronizadas que apresentam médias iguais a zero e variâncias iguais a um. Isso significa que, independentemente das suas escalas de medidas e do tipo de variação, aleatória ou não, todas elas apresentam a mesma distribuição de valores, entre -3 e 3, com probabilidade igual a 0,9973.

Portanto, a análise dos CPs por meio da matriz $\boldsymbol{\rho}$ resolve o problema da escala, mas por outro lado, não resolve o problema da importância relativa de cada variável, dado que a padronização é feita com base em uma variabilidade total estimada a partir de todos os valores observados de cada variável, separadamente. Desse modo, não importa *a priori* se as variáveis a serem analisadas por meio dos CPs são aleatórias ou não.

Em um trabalho de revisão, Hamada e Balakrishnan (1998) trazem um resumo dos principais métodos para analisar experimentos fatoriais sem repetição, bem como comparam estes métodos usando um extensivo estudo de simulação. Apresentam, de forma histórica 24 métodos propostos de 1959 a 1996 que estimam, então, a variabilidade aleatória em experimentos com observações individuais. Nesses experimentos, não é possível estimar por métodos tradicionais, como por exemplo, a análise de variância (ANOVA), a variância aleatória (residual) de cada variável original Y, caso ela não seja aleatória.

Consequentemente, a padronização por meio da variância aleatória, caso ela possa ser aproximadamente estimada, dá a possibilidade de diferenciar as variáveis aleatórias das não aleatórias, quando tal variância aleatória substituir a variância total para a padronização dos seus valores, respectivamente.

Atualmente, os métodos propostos para a análise de experimentos sem repetição, aplicam-se aos experimentos fatoriais com dois níveis por fator, completos ou fracionados. Tais métodos têm como objetivo principal, propor formas mais apropriadas de testar os efeitos principais e as diversas interações presentes nesses fatoriais, que quando se decide testar todos eles, a variância aleatória ou residual apresenta zero grau de liberdade.

Dos diversos métodos já existentes na literatura, o método de Daniel (1959), citado por Hamada e Balakrishnan (1998), foi o pioneiro e usou a seguinte proposta: quando um ponto é marcado fora da reta do gráfico seminormal, isso significa que o efeito que este representa é significativo. Ele desenvolveu também um teste formal para avaliar as significâncias dos efeitos, cuja ideia central foi, inicialmente, comparar o efeito de maior valor absoluto com os valores críticos; se este for significativo, então o efeito com o segundo maior valor absoluto é comparado com o seu respectivo valor crítico, e assim por diante. O procedimento termina quando for encontrado o primeiro efeito não significativo.

A ideia principal que está por traz e, portanto, foi precursora para o surgimento do primeiro e, assim sucessivamente, até o 24º método revisado por Hamada e Balakrishnan (1998), é que, dada a ausência de aleatoriedade de uma determinada variável, torna-se inapropriada a utilização de todos os seus valores observados para a obtenção da estimativa do desvio-padrão.

3.4.1 Método de Lenth

Lenth (1989) apresentou uma metodologia relativamente simples, que consiste em obter uma estimativa para o erro-padrão dos contrastes. No seu trabalho, ele definiu um pseudo erro-padrão (PSE) dos efeitos como sendo:

$$PSE = 1{,}5Md\left|\hat{e}_j\right|, \text{ em que:}$$

$\left|\hat{e}_j\right| = |\hat{e}_i| < 2{,}5s_0$; $s_0 = 1{,}5Md|\hat{e}_i|$; $|\hat{e}_i|$ = vetor que contém todas as estimativas absolutas dos efeitos a serem testados (i = 1, 2, ..., n); e Md = mediana.

Definiu também uma margem de erro (ME):

$$ME = t_{\frac{\alpha}{2}\frac{n}{3}}PSE, \text{ em que:}$$

$t_{\frac{\alpha}{2}\frac{n}{3}}$ = valor tabelado da distribuição t de Student com $\frac{n}{3}$ graus de liberdade, tal que se tenha uma probabilidade $\frac{\alpha}{2}$ de ocorrer valores à sua direita.

Considerando-se que as estimativas são independentes, Lenth (1989) definiu ainda uma margem de erro simultânea (SME), dada por:

$$SME = t_{\gamma,\frac{n}{3}}PSE, \text{ em que } \gamma = \frac{1+0{,}95^{\frac{1}{n}}}{2}.$$

Se $|\hat{e}_i| \geq ME$, então rejeita-se a hipótese $H_{0i}: e_i = 0$. Ou então, adicionando-se como referência, linhas relativas a $\pm ME$ e a $\pm SME$, um efeito será julgado significativo se a sua barra, no gráfico de barras dos efeitos, exceder uma das linhas referentes à margem de erro simultânea (SME) e será considerado não significativo quando a barra não ultrapassar uma das linhas referentes à margem de erro (ME).

Observe que o PSE é consistente para σ quando não há efeitos significativos, mas superestima σ, caso contrário. Os graus de liberdade $\frac{n}{3}$ vêm de uma aproximação de PSE^2 por uma distribuição χ^2 escalonada. Usando-se o PSE para padronizar os efeitos, tem-se:

$$\hat{e}_{pi} = \frac{\hat{e}_i}{PSE}.$$

3.4.2 Método de Juan e Pena

Juan e Pena (1992) propuseram uma diferente estimativa do desvio-padrão aleatório, em relação ao PSE proposto por Lenth (1989). No seu método, calcula-se inicialmente a mediana de todas as n estimativas dos efeitos absolutos, denominada de MAD_0. Depois, calcula-se a mediana daquelas estimativas dos efeitos absolutos menores que $wMAD_0$, de acordo com alguma constante w>2. O procedimento de calcular MAD_0 e de obter efeitos menores que $wMAD_0$ continua até o valor de MAD_0 não mudar.

De acordo com o método de Juan e Pena (1992), o último valor de MAD_0 é definido por $IMAD_0$. Desse modo, a estimativa do desvio-padrão aleatório é obtida por:

$$\hat{\sigma}_{IMAD} = \frac{IMAD_0}{a_w}, \text{ em que:}$$

$a_w =$ fator de correção.

Para melhor estimativa, Juan e Pena (1992) recomendam $w = 3{,}5\ e$ $a_w = 0{,}6578$.

Para a identificação dos efeitos significativos, procede-se:

$$\hat{e}_{pi} = \frac{\hat{e}_i}{\hat{\sigma}_{IMAD}}.$$

Se $|\hat{e}_{pi}| \geq z_{\frac{\alpha}{2}}$, então rejeita-se a hipótese H_{0i}: $e_i = 0$, em que $z_{\frac{\alpha}{2}} =$ valor tabelado da distribuição normal padronizada que deixa uma probabilidade $\frac{\alpha}{2}$ na extremidade da cauda à direita.

3.4.3 Método de Dong

Como concorrência também ao método de Lenth (1989), Dong (1993) propôs a seguinte estimativa:

$$\hat{\sigma}_{DONG} = \sqrt{m^{-1} \sum_{j=1}^{m} \hat{e}_j^2}, \text{ em que:}$$

$\hat{e}_j = |\hat{e}_i| < 2{,}5s_0$; $s_0 = 1{,}5Md|\hat{e}_i|$; $|\hat{e}_i|$ = vetor que contém todas as estimativas absolutas dos efeitos a serem testados (i = 1, 2, ..., n); Md = mediana; m = número de $\hat{e}_j$; n = número de $\hat{e}_i$, e m ≤ n.

Dong (1993) usa a seguinte fórmula para padronizar os efeitos:

$$\hat{e}_{pi} = \frac{\hat{e}_i}{\hat{\sigma}_{DONG}}.$$

Dong (1993) também propôs iterativamente calcular $\hat{\sigma}_{DONG}$ até que o valor não mude, quando há um grande número de efeitos significativos.

4 MATERIAL E MÉTODOS

O banco de dados do presente estudo foi obtido por meio de simulação. Foram simulados 60 conjuntos de dados compostos por quatro variáveis aleatórias com 10.000 observações cada, com três diferentes graus de correlação, dois tipos de médias, dois tipos de variâncias e cinco percentuais de outliers. A estimação do desvio-padrão foi feita de acordo com quatro métodos: desvio-padrão total, Lenth, Juan e Pena e Dong. Primeiro avaliaram-se os efeitos desses quatro fatores sob as estimativas dos desvios-padrão obtidas pelos quatro métodos por meio de uma análise de superfície de resposta. Em um segundo estudo, foram estudados em um experimento fatorial sem repetição e sob o delineamento inteiramente casualizado (DIC), quatro matrizes de covariâncias, dois tipos de variâncias e dois percentuais de outliers por meio de análise de variância e teste de Tukey.

4.1 Simulação de Dados

Para a realização do presente trabalho, foram simulados no software R 12 conjuntos de dados. Em cada conjunto, foram consideradas quatro variáveis aleatórias (Y_1, Y_2, Y_3 *e* Y_4) com 10.000 observações cada, classificadas de acordo com três diferentes graus de relações e quatro diferentes combinações de médias e variâncias.

Para isso, foram propostas três matrizes de correlações lineares para as quatro variáveis estudadas:

$$\boldsymbol{\rho_1} = \begin{bmatrix} 1 & 0{,}5 & 0 & 0 \\ 0{,}5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0{,}5 \\ 0 & 0 & 0{,}5 & 1 \end{bmatrix};$$

$$\boldsymbol{\rho_2} = \begin{bmatrix} 1 & 0{,}75 & 0 & 0 \\ 0{,}75 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0{,}75 \\ 0 & 0 & 0{,}75 & 1 \end{bmatrix};$$

$$\boldsymbol{\rho_3} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

A matriz $\boldsymbol{\rho_1}$ foi, de maneira simplificada, denominada de correlação $\rho = 0{,}5$. Do mesmo modo, $\boldsymbol{\rho_2}$ e $\boldsymbol{\rho_3}$, denominadas de correlação $\rho = 0{,}75$ e $\rho = 1$, respectivamente.

De forma geral, a matriz de correlações foi definida por:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix}.$$

Nas diferentes situações, as simulações dos 10.000 valores por variável aleatória foram baseadas na distribuição normal multivariada de probabilidades, isto é, $\mathrm{Y}_1, \mathrm{Y}_2, \mathrm{Y}_3 \; e \; \mathrm{Y}_4 \sim N_4\,(\boldsymbol{\mu}_Y;\, \boldsymbol{\Sigma}_Y)$, sendo $\boldsymbol{\mu}_Y$ o vetor de médias e $\boldsymbol{\Sigma}_Y$ a matriz de covariâncias entre as variáveis aleatórias Y_s dados por:

$$\boldsymbol{\mu}_Y = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} e\; \boldsymbol{\Sigma}_Y = \begin{bmatrix} \sigma_{11} & \rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \rho_{13}\sqrt{\sigma_{11}\sigma_{33}} & \rho_{14}\sqrt{\sigma_{11}\sigma_{44}} \\ \rho_{21}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} & \rho_{23}\sqrt{\sigma_{22}\sigma_{33}} & \rho_{24}\sqrt{\sigma_{22}\sigma_{44}} \\ \rho_{31}\sqrt{\sigma_{11}\sigma_{33}} & \rho_{32}\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} & \rho_{34}\sqrt{\sigma_{33}\sigma_{44}} \\ \rho_{41}\sqrt{\sigma_{11}\sigma_{44}} & \rho_{42}\sqrt{\sigma_{22}\sigma_{44}} & \rho_{43}\sqrt{\sigma_{33}\sigma_{44}} & \sigma_{44} \end{bmatrix},$$

em que: $\mu_w =$ média do $\mathrm{Y_w}$; $\sigma_{ww} =$ variância de $\mathrm{Y_w}$; $\rho_{ww'} =$ coeficiente de correlação entre $\mathrm{Y_w}$ e $\mathrm{Y_{w'}}$, sendo w ≠ w' (w, w' = 1, 2, 3 e 4).

No primeiro cenário as variáveis aleatórias Y_1, Y_2, Y_3 e Y_4 foram simuladas com médias e variâncias iguais de acordo com as três matrizes de correlações, como segue:

$$\mu_Y = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \end{bmatrix} e\; \boldsymbol{\Sigma}_Y = \begin{bmatrix} 10 & 10\rho_{12} & 0 & 0 \\ 10\rho_{12} & 10 & 0 & 0 \\ 0 & 0 & 10 & 10\rho_{34} \\ 0 & 0 & 10\rho_{34} & 10 \end{bmatrix}.$$

No segundo cenário as variáveis aleatórias Y_1, Y_2, Y_3 e Y_4 foram simuladas com médias iguais e variâncias diferentes de acordo com as três matrizes de correlações, como segue:

$$\mu_Y = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \end{bmatrix} e\; \boldsymbol{\Sigma}_Y = \begin{bmatrix} 10 & 5\rho_{12} & 0 & 0 \\ 5\rho_{12} & 2{,}5 & 0 & 0 \\ 0 & 0 & 10 & 5\rho_{34} \\ 0 & 0 & 5\rho_{34} & 2{,}5 \end{bmatrix}.$$

No cenário 3 as variáveis aleatórias Y_1, Y_2, Y_3 e Y_4 foram simuladas com médias diferentes e variâncias iguais de acordo com as três matrizes de correlações, como segue:

$$\mu_Y = \begin{bmatrix} 100 \\ 50 \\ 100 \\ 50 \end{bmatrix} e\ \boldsymbol{\Sigma}_Y = \begin{bmatrix} 10 & 10\rho_{12} & 0 & 0 \\ 10\rho_{12} & 10 & 0 & 0 \\ 0 & 0 & 10 & 10\rho_{34} \\ 0 & 0 & 10\rho_{34} & 10 \end{bmatrix}.$$

Por fim, no quarto cenário as variáveis aleatórias Y_1, Y_2, Y_3 e Y_4 foram simuladas com médias e variâncias diferentes de acordo com as três matrizes de correlações, como segue:

$$\mu_Y = \begin{bmatrix} 100 \\ 50 \\ 100 \\ 50 \end{bmatrix} e\ \boldsymbol{\Sigma}_Y = \begin{bmatrix} 10 & 5\rho_{12} & 0 & 0 \\ 5\rho_{12} & 2{,}5 & 0 & 0 \\ 0 & 0 & 10 & 5\rho_{34} \\ 0 & 0 & 5\rho_{34} & 2{,}5 \end{bmatrix}.$$

4.2 Adição de O*utliers*

Após a simulação dos 10.000 valores para cada um dos 12 conjuntos de dados, foram substituídos 0 (0%), 100 (1%), 200 (2%), 300 (3%) e 400 (4%) valores aleatórios pelas mesmas quantidades respectivas de *outliers*, com o objetivo de quebrar a aleatoriedade das variáveis Y_1, Y_2, Y_3 e Y_4.

Segundo Triola (1999), são considerados *outliers* todos os valores inferiores e superiores a:

$$q_1 - 1{,}5a_{iq};$$

$$q_3 + 1{,}5a_{iq},\ \text{em que:}$$

$a_{iq} = q_3 - q_1 =$ estimativa da amplitude interquartílica; q_1 = primeiro quartil; q_3 = terceiro quartil.

Os *outliers* foram adicionados de forma unilateral à direita, tal que:

$$y_w > q_3 + 1{,}5a_{iq}, para\ w = 1, 2, 3\ e\ 4.$$

Portanto, foram gerados, no total, 60 conjuntos de dados classificados de acordo com os cruzamentos entre três graus de relações (0,5; 0,75 e 1), dois tipos de médias (iguais e diferentes), dois tipos de variâncias (iguais e diferentes) e cinco percentuais de *outliers* (0, 1, 2, 3 e 4%).

Após as constituições dos 60 arquivos de dados, foi realizado, no software R, o teste de Kolmogorov-Smirnov para a verificação da normalidade, a 5% de significância, a fim de confirmar se cada variável ($Y_1, Y_2, Y_3\ e\ Y_4$) é aleatória normal ou não.

4.3 Métodos de Estimação do Desvio-padrão

Os CP's, estimados nos 60 arquivos de dados, foram obtidos com base na matriz (4x4) amostral de covariâncias obtida a partir dos valores padronizados (S_z), como segue:

$$z_w = \frac{y_w - \bar{y}_w}{\sqrt{s_{ww}}} = \frac{y_w - \bar{y}_w}{s_w}, \text{ em que:}$$

y_w = valor simulado da variável Y_w; $\bar{y}_w =$ estimativa da média da variável Y_w; s_{ww} = estimativa da variância da variável Y_w; s_w = estimativa do desvio-padrão da variável Y_w; w = 1, 2, 3 e 4.

Para a obtenção do desvio-padrão s_w, e com auxílio do software Microsoft Excel 2013, foram utilizados os quatro métodos de estimação, sendo:

$$\boldsymbol{S_z} = \begin{bmatrix} s_{z_1z_1} & s_{z_1z_2} & s_{z_1z_3} & s_{z_1z_4} \\ s_{z_2z_1} & s_{z_2z_2} & s_{z_2z_3} & s_{z_2z_4} \\ s_{z_3z_1} & s_{z_3z_2} & s_{z_3z_3} & s_{z_3z_4} \\ s_{z_4z_1} & s_{z_4z_2} & s_{z_4z_3} & s_{z_4z_4} \end{bmatrix}, \text{ em que:}$$

$\boldsymbol{S_z}$ = matriz das estimativas das variâncias e covariâncias; $s_{z_wz_w}$ = estimativa da variância da variável padronizada Z_w; $s_{z_wz_{w'}}$= estimativa da covariância entre as variáveis padronizadas Z_w e $Z_{w'}$.

De acordo com cada método de estimação do desvio-padrão, pode implicar em diferentes padronizações de determinada variável original Y_w. Portanto, pode ocorrer $s_{z_wz_{w'}} \neq r_{y_wy_{w'}}$ de diferentes graus e, não necessariamente, $s_{z_wz_{w'}} = r_{y_wy_{w'}}$.

4.3.1 Método do Desvio-padrão Total

O primeiro método, denominado de desvio-padrão total, forneceu a seguinte estimativa do desvio-padrão da variável Y_w:

$$s_{w(1)} = \sqrt{\frac{\sum_{i=1}^{n}(y_{wi} - \bar{y}_w)^2}{n-1}}, \text{ em que:}$$

i = 1, 2, ...,n; n = 10.000 e w = 1, 2, 3 e 4.

4.3.2 Método de Lenth

Já o segundo método, proposto por Lenth (1989), forneceu a estimativa do desvio-padrão de Y_w da seguinte forma:

$$s_{w(2)} = PSE_w, \text{ em que:}$$

$PSE_w = 1{,}5Md|y_{wj} - \bar{y}_w|$; $|y_{wj} - \bar{y}_w| = |y_{wi} - \bar{y}_w| < 2{,}5s_0$; s_0 = $1{,}5Md\ |y_{wi} - \bar{y}_w|$ $|y_{wi} - \bar{y}_w|$ = vetor que contém todas as estimativas dos desvios absolutos dos valores em relação à média da variável Y_w, para w = 1, 2, 3 e 4 e i = 1, 2, ..., 10.000; e Md = mediana.

4.3.3 Método de Juan e Pena

O método de Juan e Pena (1992), terceiro método, forneceu a seguinte estimativa do desvio-padrão de Y_w:

$$s_{w(3)} = \hat{\sigma}_{IMAD_w}, \text{ em que:}$$

$\hat{\sigma}_{IMAD_w} = \frac{IMAD_{0_w}}{0{,}6578}$; $IMAD_{0_w}$ = último valor de MAD_{0_w}; e MAD_{0_w} = mediana de todas as n estimativas absolutas dos desvios dos valores em relação à média da variável Y_w, para w = 1, 2, 3 e 4 e n = 10.000.

Neste método, o primeiro passo foi encontrar as n estimativas dos desvios absolutos dos valores em relação à média da variável e calcular a mediana delas, o que foi chamado de MAD_{0_w}. Logo após encontrou-se a mediana das estimativas dos efeitos menores que $wMAD_{0_w}$, em que o valor utilizado para w foi igual a 3,5. O procedimento foi feito até que o valor de MAD_{0_w} não mudasse.

4.3.4 Método de Dong

No quarto método, denominado método de Dong (1993), a estimativa foi dada por:

$$s_{w(4)} = \hat{\sigma}_{DONG_w}, \text{ em que:}$$

$\hat{\sigma}_{DONG_w} = \sqrt{m^{-1}\sum_{j=1}^{m}(y_{wj} - \bar{y}_w)^2}$;

$|y_{wj} - \bar{y}_w| = |y_{wi} - \bar{y}_w| < 2{,}5s_0$;

$s_0 = 1{,}5Md|y_{wi} - \bar{y}_w|$;

$|y_{wi} - \bar{y}_w|$ = vetor que contém todas as estimativas dos desvios absolutos dos valores em relação à média da variável Y_w para w = 1, 2, 3 e 4 e i = 1, 2, ..., 10.000; Md = mediana; m = número de $|y_{wj} - \bar{y}_w|$; e n = número de $|y_{wi} - \bar{y}_w|$ (m ≤ n).

4.4 Medidas de Avaliação dos Métodos

Para avaliar a qualidade de estimação pelos quatro métodos estudados, com base nas cinco alterações de adições de *outliers* propostas, foram avaliadas duas variáveis:

$$Dvar = \Delta\sigma_{ww} = \frac{1}{4}\sum_{w=1}^{4}\frac{|s_{ww} - \sigma_{ww}|}{\sigma_{ww}};$$

$$Dcov = \Delta\sigma_{ww'} = \frac{1}{6}\sum_{w,w'=1}^{6}\frac{|s_{ww'} - \sigma_{ww'}|}{\sigma_{ww'}}, \text{ em que:}$$

s_{ww} = estimativa da variância da variável Y_w; σ_{ww} = variância paramétrica da variável Y_w; $s_{ww'}$ = estimativa da covariância entre as variáveis Y_w e $Y_{w'}$; $\sigma_{ww'}$ = covariância paramétrica entre as variáveis Y_w e $Y_{w'}$; w, w' = 1, 2, 3 e 4.

As estimativas dessas variáveis mostrarão as diferenças absolutas entre as variâncias estimadas e paramétricas e entre as covariâncias estimadas e paramétricas, e espera-se que todas sejam iguais a zero.

Para cada variável avaliada, em cada um dos quatro métodos separadamente, foi realizada análise de superfície de resposta, no software R, técnica estatística utilizada para modelagem e análise de problemas nos quais a variável resposta é influenciada por vários fatores, cujo objetivo é a otimização dessa resposta. A variável dependente foi estudada em função dos tipos de médias (iguais = 0 e diferentes = 1), dos tipos de variância (iguais = 0 e diferentes = 1), dos graus de relações (0,5; 0,75 e 1) e dos percentuais de *outliers* (0, 1, 2, 3 e 4%). Para a seleção do melhor modelo, foram retirados os termos não significativos pelo teste t de Student, a 5% de probabilidade, começando pelas interações duplas, cujo maior modelo definido foi dado por:

$$y = \beta_0 + \beta_1 m + \beta_2 v + \beta_3 \rho + \beta_4 o + \beta_5 mv + \beta_6 m\rho + \beta_7 mo + \beta_8 v\rho + \beta_9 vo + \beta_{10}\rho o + \varepsilon, \text{ em que:}$$

β_0 = constante da regressão;

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$ = coeficientes da regressão;

m = tipos de média;

v = tipos de variância;

ρ = graus de relações;

o = percentuais de *outliers*;

ε = erro da regressão, para $\varepsilon \sim N\ (0;\ \sigma^2)$.

Posteriormente, e de acordo com a melhor superfície de resposta ajustada, foram construídos os seus gráficos. E para separar os quatro métodos de estimação do desvio-padrão, dentro de cada combinação entre os níveis dos três fatores estudados, mas apenas dos detectados como significativos pelo teste t de Student a 5% de probabilidade, foi realizado uma análise de agrupamento com base nas variáveis avaliadas Dvar e Dcov com valores padronizados, e por meio do método da ligação média. Plotou-se os dendogramas e a separação dos grupos foi feita com base no mínimo de 70% de similaridade. Tais análises foram feitas no software Minitab 16®. Foram obtidas ainda, médias das variáveis Dvar e Dcov dentro de cada grupo selecionado pelo agrupamento.

4.5 Matrizes de Covariâncias

A análise de CPs foi realizada com base nos autovalores e autovetores de uma matriz de covariâncias, e, a partir dos quatro métodos de estimação do desvio-padrão, foram formadas diferentes matrizes de covariâncias.

Foram estabelecidas oito matrizes de covariâncias. Desse total, foram utilizadas para a análise dos CPs, as matrizes baseadas no método do desvio-padrão total ($\mathbf{S_{ZT}}$), com dados transformados de acordo com seu respectivo coeficiente de variação ($\mathbf{S_{ZCV}}$), e as duas melhores relacionadas aos métodos de Lenth (1989), Juan e Pena (1992) e Dong (1993).

4.5.1 Matriz S_{ZT}

A matriz $\mathbf{S_{ZT}}$ é a matriz de covariâncias entre as variáveis $Z_1, Z_2, Z_3\ e\ Z_4$, cuja estimativa do desvio-padrão utilizada para as obtenções dos seus respectivos valores padronizados, foi obtida por:

$$s_w = \sqrt{\frac{\sum_{i=1}^{10.000}(y_{wi} - \bar{y}_w)^2}{10.000-1}}$$, em que:

y_{wi} = valor simulado da variável Y_w de ordem i (i = 1, 2, ..., 10.000); $\bar{y}_w =$ estimativa da média da variável Y_w; w = 1, 2, 3 e 4.

4.5.2 Matriz S_{ZCV}

A matriz $\mathbf{S_{ZCV}}$ foi obtida com base na transformação dos dados originais, levando-se em consideração o coeficiente de variação (CV) das respectivas variáveis originais. A transformação utilizada foi definida como segue:

$$z_w = \frac{y_w - \bar{y}_w}{\bar{y}_w}$$, em que:

y_w = valor simulado da variável Y_w de ordem i (i = 1, 2, ..., 10.000); $\bar{y}_w =$ estimativa da média da variável Y_w; w = 1, 2, 3 e 4.

4.5.3 Matriz S_{ZL}

A matriz $\mathbf{S_{ZL}}$ é a matriz de covariâncias entre as variáveis Z_1, Z_2, Z_3 *e* Z_4, cuja estimativa do desvio-padrão foi obtida com base no método de Lenth (1989), como segue:

$$s_w = PSE_w.$$

Nesse caso, foram obtidas duas matrizes de covariâncias. A matriz $\mathbf{S_{ZL1}}$ foi encontrada utilizando-se os cálculos das covariâncias para todos os 10.000 valores simulados. E para a matriz $\mathbf{S_{ZL2}}$, os cálculos das covariâncias foram realizados somente sobre os valores considerados como aleatórios por este método, isto é, para aqueles cujos desvios absolutos ($|y_{wi} - \bar{y}_w|$) foram menores do que $2{,}5s_0$.

4.5.4 Matriz S_{ZJP}

A matriz $\mathbf{S_{ZJP}}$ é a matriz de covariâncias entre as variáveis Z_1, Z_2, Z_3 *e* Z_4, cuja estimativa do desvio-padrão foi obtida com base no método de Juan e Pena (1992), definida a seguir:

$$s_w = \hat{\sigma}_{IMAD_w}.$$

Nesta situação, foram obtidas duas matrizes de covariâncias. A matriz $\mathbf{S_{ZJP1}}$ foi encontrada através dos cálculos das covariâncias a partir dos 10.000 valores simulados. Enquanto isso, para matriz $\mathbf{S_{ZJP2}}$, os cálculos das covariâncias foram realizados somente sobre os valores considerados como aleatórios por este método, isto é, para aqueles cujo valor não mudou ao realizar o procedimento necessário.

4.5.5 Matriz S_{ZD}

A matriz $\mathbf{S_{ZD}}$ é, também, uma matriz de covariâncias entre as variáveis $Z_1, Z_2, Z_3 \; e \; Z_4$, em que a estimativa do desvio-padrão foi obtida com base no método de Dong (1993), como segue:

$$s_w = \hat{\sigma}_{DONG_w}.$$

De maneira semelhante à matriz $\mathbf{S_{ZL}}$, foram obtidas duas matrizes de covariâncias, em que a matriz $\mathbf{S_{ZD1}}$ foi encontrada com base nos cálculos das covariâncias utilizando todos os 10.000 valores simulados. Já para a matriz $\mathbf{S_{ZD2}}$, os cálculos das covariâncias foram feitos somente sobre os valores considerados como aleatórios por este método, isto é, para aqueles cujos desvios absolutos ($|y_{wi} - \bar{y}_w|$) foram menores do que $2{,}5s_0$.

4.6 Análise dos Componentes Principais

Após estimadas as quatro matrizes de covariâncias, foram realizadas as análises dos CPs. Foram estimados apenas os dois primeiros componentes, porque as variáveis simuladas Y_1 e Y_2 correlacionam-se entre si com coeficientes iguais a 0,5; 0,75 e 1, mas não se correlacionam com as variáveis Y_3 e Y_4, que estão relacionadas entre si com os mesmos três coeficientes.

As variáveis $Z_1, Z_2, Z_3 \; e \; Z_4$ com vetor (4x1) de médias e matriz amostral (4x4) $\mathbf{S_Z}$ de covariâncias, foram utilizadas para as estimativas dos CPs:

$$CP_1 = \hat{e}_{11}Z_1 + \hat{e}_{12}Z_2 + \hat{e}_{13}Z_3 + \hat{e}_{14}Z_4;$$
$$CP_2 = \hat{e}_{21}Z_1 + \hat{e}_{22}Z_2 + \hat{e}_{23}Z_3 + \hat{e}_{24}Z_4.$$

A estimativa do vetor $\hat{\boldsymbol{e}}_{\boldsymbol{j}} = [\hat{e}_{j1} \quad \hat{e}_{j2} \quad \hat{e}_{j3} \quad \hat{e}_{j4}]$ de autovetores normalizados de cada CP, foi obtido por:

$$\hat{\boldsymbol{e}}_{\boldsymbol{j}} = \frac{1}{||\hat{\boldsymbol{a}}_{\boldsymbol{j}}||}\hat{\boldsymbol{a}}_{\boldsymbol{j}}, \text{ para j = 1 e 2, em que:}$$

$\left[\boldsymbol{S}_{\boldsymbol{Z}} - \hat{\lambda}_j \boldsymbol{I}\right]\hat{\boldsymbol{a}}_{\boldsymbol{j}} = \mathbf{0};$

$\hat{\lambda}_j$ = estimativa do autovalor do CP_j , j = 1 e 2.

No total, foram realizadas 240 análises dos CPs no software R. Para avaliar a análise dos CPs, foram utilizadas as seguintes variáveis associadas aos dois primeiros componentes principais:

$$V_1 = \frac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sum_{j=1}^{4}\hat{\lambda}_j};$$

$$V_2 = |r_{CP_1Z_1} - r_{CP_1Z_2}|;$$

$$V_3 = |r_{CP_1Z_3} - r_{CP_1Z_4}|;$$

$$V_4 = |r_{CP_2Z_1} - r_{CP_2Z_2}|;$$

$$V_5 = |r_{CP_2Z_3} - r_{CP_2Z_4}|, \text{ em que:}$$

$\hat{\lambda}_1$ = estimativa do primeiro autovalor do CP_1;

$\hat{\lambda}_2$ = estimativa do segundo autovalor do CP_2;

$r_{CP_1Z_1}$ = estimativa do coeficiente de correlação do CP_1 com a variável Z_1;

$r_{CP_1Z_2}$ = estimativa do coeficiente de correlação do CP_1 com a variável Z_2;

$r_{CP_1Z_3}$ = estimativa do coeficiente de correlação do CP_1 com a variável Z_3;

$r_{CP_1Z_4}$ = estimativa do coeficiente de correlação do CP_1 com a variável Z_4;

$r_{CP_2Z_1}$ = estimativa do coeficiente de correlação do CP_2 com a variável Z_1;

$r_{CP_2Z_2}$ = estimativa do coeficiente de correlação do CP_2 com a variável Z_2;

$r_{CP_2Z_3}$ = estimativa do coeficiente de correlação do CP_2 com a variável Z_3;

$r_{CP_2Z_4}$ = estimativa do coeficiente de correlação do CP_2 com a variável Z_4;

$r_{CP_jZ_w} = \frac{\hat{e}_{jw}\sqrt{\hat{\lambda}_j}}{\sqrt{s_{ww}}}$, para j = 1 e 2 e w = 1, 2, 3 e 4.

A variável analisada V_1 expressa a diferença entre as estimativas dos autovalores 1 e 2. Esta diferença é esperada aumentar com o aumento da correlação, porque o aumento da correlação aumenta a importância do primeiro CP. Já para V_2, V_3, V_4 e V_5 esperam-se estimativas iguais a zero, porque as variáveis 1 e 2 são correlacionadas entre si e as variáveis 3 e 4 também, sendo que os pares correlacionados são independentes entre si. Isto significa, teoricamente, que os coeficientes e_{11} e e_{12} do CP_1 e e_{21} e e_{22} do CP_2

devem ser iguais. Por outro lado, os coeficientes e_{13} e e_{14} do CP_1 e e_{23} e e_{24} do CP_2 devem ser iguais a zero.

Para cada uma das cinco variáveis analisadas (V_1 até V_5), foi realizada análise de variância de um experimento fatorial segundo o delineamento inteiramente casualizado. Nele, foram estudados, no máximo, cinco fatores definidos por: matrizes de covariâncias (a_1 = $\mathbf{S}_{ZT}$; a_2 = $\mathbf{S}_{Zcv}$; a_3 e a_4 = duas das matrizes escolhidas com base em um dos métodos de estimação do desvio-padrão), tipos de médias (b_1 = iguais e b_2 = diferentes), tipos de variâncias (c_1 = iguais e c_2 = diferentes), graus de relações (d_1 = 0,5, d_2 = 0,75 e d_3 = 1) e percentuais dos *outliers* (e_1 = 0% e e_2 = máximo percentual avaliado adequadamente pelo melhor método de estimação). Desses cinco fatores, foram estudados então, as matrizes e aqueles que apresentaram efeitos significativos para as variáveis Dvar e Dcov de acordo com a análise da superfície de resposta.

Foi, portanto, analisado o efeito principal de cada fator, inclusive as diversas interações entre eles. Para compor o resíduo da análise de variância, foi utilizada a interação de maior ordem e, posteriormente, aplicado o teste de Tukey de acordo com as significâncias das interações. Os testes F e de Tukey foram realizados no software R a 10% de probabilidade.

O objetivo principal dessa análise foi de verificar, principalmente, o melhor método de estimação do desvio-padrão para as análises de componentes principais realizadas em diferentes condições.

5 RESULTADOS E DISCUSSÃO

O fator mais importante em afetar a qualidade da estimativa do desvio-padrão foi o percentual de *outliers*, e o método de estimação que apresentou resultados mais satisfatórios foi o de Lenth. Por outro lado, a matriz que forneceu os melhores resultados para a análise de componentes principais, foi a que utilizou a estimativa do desvio-padrão obtida pelo método do desvio-padrão total, na ausência de *outliers*, com variâncias iguais e diferentes. A matriz com base no método de Lenth, em que os cálculos das covariâncias foram realizados sobre os valores considerados como aleatórios por este método, foi adequada nos casos em que as variâncias são iguais ou diferentes e na presença de *outliers*.

5.1 Métodos de Estimação do Desvio-padrão

Após a constituição dos 60 conjuntos de dados, que contêm cada um deles, quatro variáveis ($Y_1, Y_2, Y_3 \; e \; Y_4$) com 10.000 valores, e obtidos por meio das combinações entre três matrizes de correlações, dois tipos de médias, dois tipos de variâncias e cinco percentuais de *outliers*, foram obtidas 240 estimativas dos desvios-padrão, realizou-se o teste de Kolmogorov-Sminorv para a verificação da normalidade, a 5% de significância, a fim de confirmar quando for variável aleatória ou não. No caso em que o percentual de *outliers* foi zero, observaram-se as quatro variáveis atendendo à pressuposição de normalidade. Todos os demais conjuntos de dados não apresentaram as quatro variáveis como sendo aleatórias normais. Isso significa que as variâncias estimadas com base no método tradicional não serão aleatórias e, consequentemente, se forem estimadas de uma forma total não haverá a possibilidade de captar a parte não aleatória das mesmas.

Para os 60 conjuntos de dados, estão apresentados nas Tabelas 5.1, 5.2, 5.3, 5.4 e 5.5, as diferenças em relação aos respectivos parâmetros, com base no método do desvio-padrão total. Tais diferenças variaram até, no máximo, em 61,7% da variabilidade aleatória.

Tabela 5.1 - Diferenças absolutas entre as estimativas e os parâmetros dos desvios-padrão, com base no método do desvio-padrão total, na ausência de *outliers*

		Variâncias iguais				Variâncias diferentes			
Correlação	Médias	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
0,5	Iguais	0,01	0,00	0,04	0,05	0,00	0,01	0,00	0,00
	Diferentes	0,02	0,00	0,03	0,00	0,01	0,01	0,03	0,01
0,75	Iguais	0,02	0,02	0,02	0,02	0,01	0,00	0,04	0,02
	Diferentes	0,02	0,06	0,01	0,02	0,02	0,02	0,01	0,00
1	Iguais	0,02	0,02	0,01	0,01	0,01	0,01	0,00	0,00
	Diferentes	0,01	0,01	0,02	0,02	0,01	0,01	0,01	0,01

Tabela 5.2 - Diferenças absolutas entre as estimativas e os parâmetros dos desvios-padrão, com base no método do desvio-padrão total, com 1% de *outliers*

		Variâncias Iguais				Variâncias diferentes			
Correlação	Médias	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
0,5	Iguais	0,59	0,57	0,50	0,52	0,31	0,58	0,32	0,56
	Diferentes	0,52	0,55	0,60	0,55	0,55	0,29	0,54	0,31
0,75	Iguais	0,57	0,56	0,57	0,56	0,29	0,54	0,35	0,59
	Diferentes	0,55	0,62	0,58	0,55	0,57	0,31	0,53	0,28
1	Iguais	0,56	0,56	0,57	0,57	0,30	0,56	0,30	0,55
	Diferentes	0,54	0,54	0,55	0,55	0,58	0,29	0,55	0,28

Tabela 5.3 - Diferenças absolutas entre as estimativas e os parâmetros dos desvios-padrão, com base no método do desvio-padrão total, com 2% de *outliers*

		Variâncias Iguais				Variâncias diferentes			
Correlação	Médias	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
0,5	Iguais	1,02	1,01	0,98	0,98	0,62	1,03	0,61	1,01
	Diferentes	0,99	1,03	1,09	1,03	1,03	0,53	0,98	0,53
0,75	Iguais	1,04	1,03	1,02	1,02	0,64	1,03	0,63	1,01
	Diferentes	1,08	1,12	1,08	1,01	1,04	0,54	1,01	0,52
1	Iguais	1,05	1,05	1,03	1,03	0,61	1,01	0,61	1,01
	Diferentes	1,01	1,01	1,05	1,05	1,03	0,51	1,08	0,54

Tabela 5.4 - Diferenças absolutas entre as estimativas e os parâmetros dos desvios-padrão, com base no método do desvio-padrão total, com 3% de *outliers*

		Variâncias Iguais				Variâncias diferentes			
Correlação	Médias	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
0,5	Iguais	1,44	1,42	1,38	1,39	0,88	1,38	0,86	1,39
	Diferentes	1,43	1,47	1,51	1,45	1,40	0,72	1,37	0,73
0,75	Iguais	1,47	1,47	1,40	1,40	0,87	1,38	0,93	1,40
	Diferentes	1,52	1,57	1,48	1,46	1,44	0,75	1,43	0,72
1	Iguais	1,46	1,46	1,46	1,46	0,81	1,34	0,91	1,40
	Diferentes	1,43	1,43	1,46	1,46	1,40	0,70	1,47	0,73

Tabela 5.5 - Diferenças absolutas entre as estimativas e os parâmetros dos desvios-padrão, com base no método do desvio-padrão total, com 4% de *outliers*

		Variâncias Iguais				Variâncias diferentes			
Correlação	Médias	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
0,5	Iguais	1,82	1,80	1,78	1,77	1,10	1,70	1,14	1,71
	Diferentes	1,82	1,85	1,90	1,85	1,77	0,91	1,75	0,94
0,75	Iguais	1,85	1,86	1,81	1,81	1,08	1,71	1,16	1,73
	Diferentes	1,89	1,95	1,79	1,79	1,82	0,94	1,81	0,90
1	Iguais	1,87	1,87	1,86	1,86	1,10	1,70	1,09	1,68
	Diferentes	1,81	1,81	1,85	1,85	1,82	0,91	1,85	0,92

Notou-se que o aumento do *outlier* proporcionou aumento no percentual do erro da estimativa, independentemente das diferentes correlações (0,5; 0,75 e 1), tipos de médias (iguais e diferentes) e tipos de variâncias (iguais e diferentes). De acordo com os resultados obtidos, as diferenças absolutas entre as variâncias estimadas e paramétricas e entre as covariâncias estimadas e paramétricas aumentaram ($P < 0,05$) em função somente do aumento do percentual dos *outliers*.

Conforme Devlin (1981), há grande instabilidade nos métodos tradicionais quando existem *outliers* no conjunto de dados. De acordo com Lawson (2008), os valores discrepantes podem ser o impedimento principal para validar a interpretação de dados de experimentos. Desse modo, a utilização de um método que minimize o efeito da não aleatoriedade do conjunto de dados irá proporcionar uma estimativa da matriz de covariâncias

mais corretamente. Isso significa que na presença de *outliers*, o método do desvio-padrão total deve ser evitado.

Dos quatro métodos estudados, primeiramente quer se avaliar a qualidade de estimação. Um método bom e robusto é aquele não é influenciado por nenhum fator. Para se ter ideia de quais variáveis afetaram significativamente a variação dos Ys, realizou-se uma análise de superfície de resposta que auxilia no processo de seleção dos fatores, eliminando-se aqueles cuja contribuição não seja importante (Tabela 5.6).

Tabela 5.6 - Equações de regressão e coeficientes de determinação para cada variável avaliada por cada um dos quatro métodos

Variável	Método	Eq. Regressão Ajustada	R^2
Dvar	Desvio-padrão total	0,0149 + 0,4072*o	0,93
	Lenth	0,0256 - 0,0086*o + 0,0057*o^2	0,82
	Juan e Pena	0,0706	-
	Dong	0,0283 - 0,0089*o + 0,0055*o^2	0,77
Dcov	Desvio-padrão total	0,0620 + 2,1160*o	0,83
	Lenth	0,0741	-
	Juan e Pena	0,0741	-
	Dong	0,3857 - 0,1356*ρ	0,93

Na Figura 5.1, estão plotados os gráficos dos ajustes das equações de regressão de Dvar em função dos percentuais de *outliers* para os métodos do desvio-padrão total, Lenth, Juan e Pena e Dong. De acordo com ela, os gráficos com base nos métodos de Lenth e Dong se comportaram de maneira semelhante.

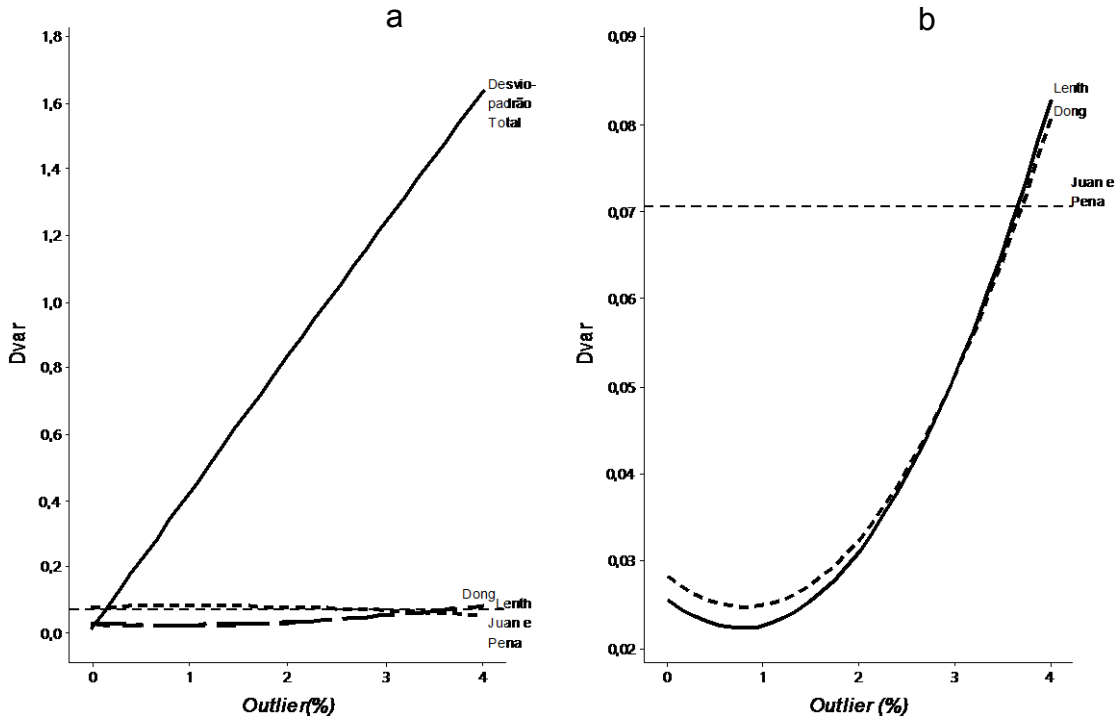


Figura 5.1 - Estimativas de Dvar em função dos percentuais dos *outliers* entre 0 a 4%, para os métodos de estimação do desvio-padrão total, Lenth, Juan e Pena e Dong (a). Diminui-se a escala para os métodos de Lenth, Juan e Pena e Dong (b).

De acordo com a Figura 5.1a, o único método que não conseguiu evitar o efeito da adição dos *outliers* foi o do desvio-padrão total. Isto indicou ser ele um método que sofre o efeito da não aleatoriedade. A fim de ilustrar melhor os métodos de Lenth, Juan e Pena e Dong, diminuiu-se a escala utilizada para eles na Figura 5.1b.

Para Dvar, percebeu-se que os métodos de Lenth e de Dong mostraram sensibilidades em detectar as presenças dos poucos *outliers* e conseguiram estimar adequadamente a variância aleatória dos dados, enquanto o método de Juan e Pena não.

Em geral, o Dvar aumenta ($P < 0,05$) em função do aumento da frequência de *outliers* para os métodos do desvio-padrão total, Lenth e Dong. Porém, ele não sofre efeito ($P > 0,05$) de nenhum dos três fatores estudados (tipos de média, tipos de variâncias e graus de correlação).

Em relação aos métodos de Lenth e de Dong, a interpretação com base na equação de regressão ajustada de 2º grau, se deu somente a partir da fase crescente em relação ao ponto de mínimo.

Já Dcov, conforme Figura 5.2, aumenta (P < 0,05) em função dos percentuais dos *outliers* somente para o método do Desvio-padrão Total. Para os métodos de Lenth e de Juan e Pena não ocorreram efeitos (P > 0,05). E para o método de Dong, o aumento da correlação diminuiu (P < 0,05) a sua média, fato este não esperado. De todo modo, não se recomenda então o método de Dong para situações em que as correlações são mais baixas.

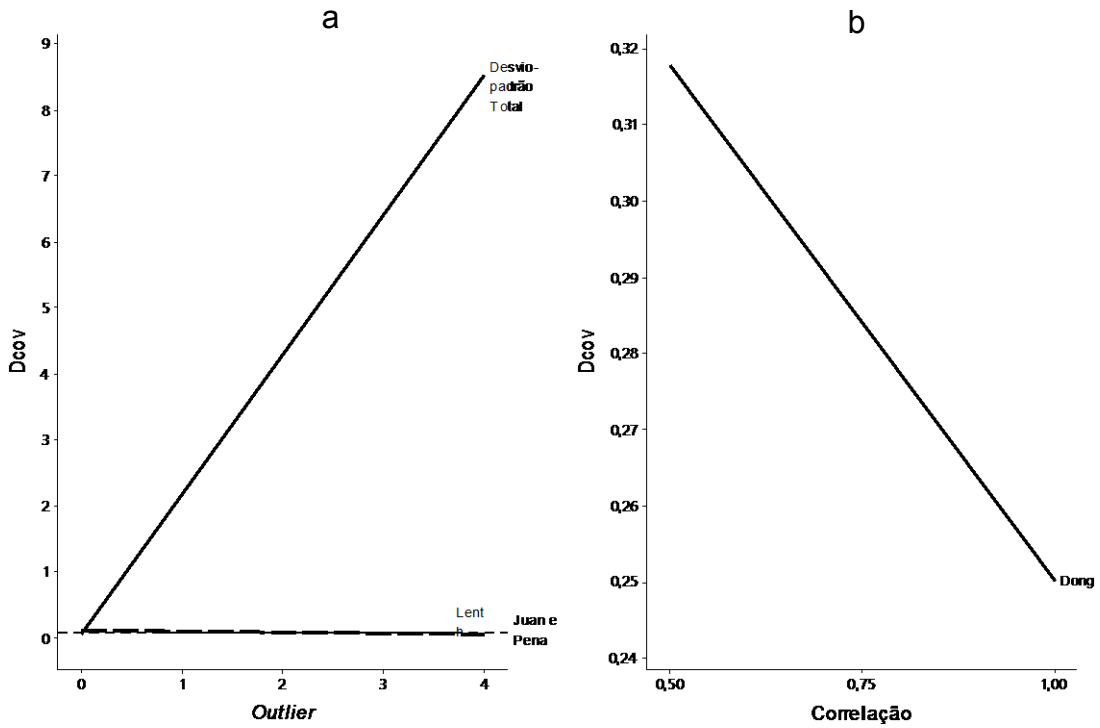


Figura 5.2 - Estimativas de Dcov em função dos percentuais dos *outliers* para os métodos de estimação do desvio-padrão total, Lenth e Juan e Pena (a). Estimativas de Dcov em função dos graus de relação para o método de Dong (b).

Se a variável é totalmente aleatória, a estimativa do seu desvio-padrão está correta. Porém, se não ocorre essa aleatoriedade total, isto é, se há a presença de *outliers*, a estimativa do desvio-padrão é errada e superestimada. Portanto, nessas condições de não aleatoriedade das variáveis estudadas, torna-se necessário estimar o desvio-padrão por outro método.

Dadas as significâncias das frequências de *outliers* e das correlações nas equações de regressão para pelo menos uma variável (Dvar e Dcov) e pelo menos um método de estimação, os agrupamentos foram construídos

para as 15 combinações entre os *outliers* e as correlações (Figuras 5.3, 5.4, 5.5, 5.6 e 5.7).

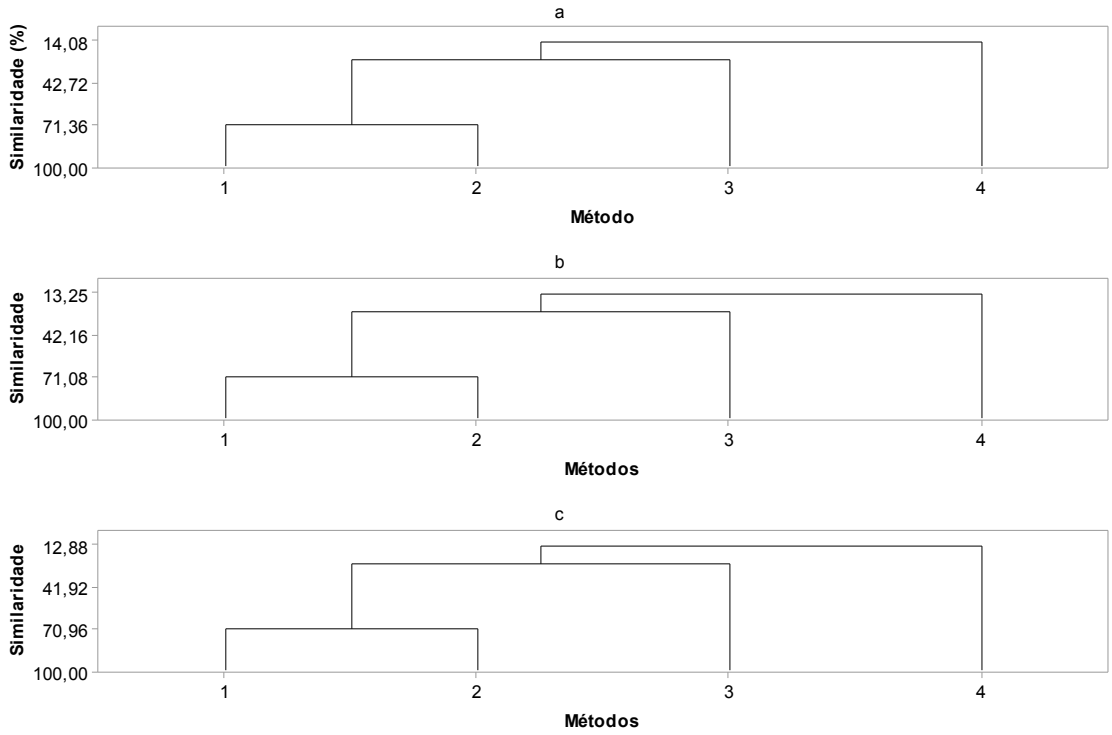


Figura 5.3 - Agrupamento dos quatro métodos (desvio-padrão total = 1, de Lenth = 2, de Juan e Pena = 3 e de Dong = 4) em função dos diferentes graus de correlação: 0,5 (a); 0,75 (b); 1 (c) e na ausência de *outliers*.

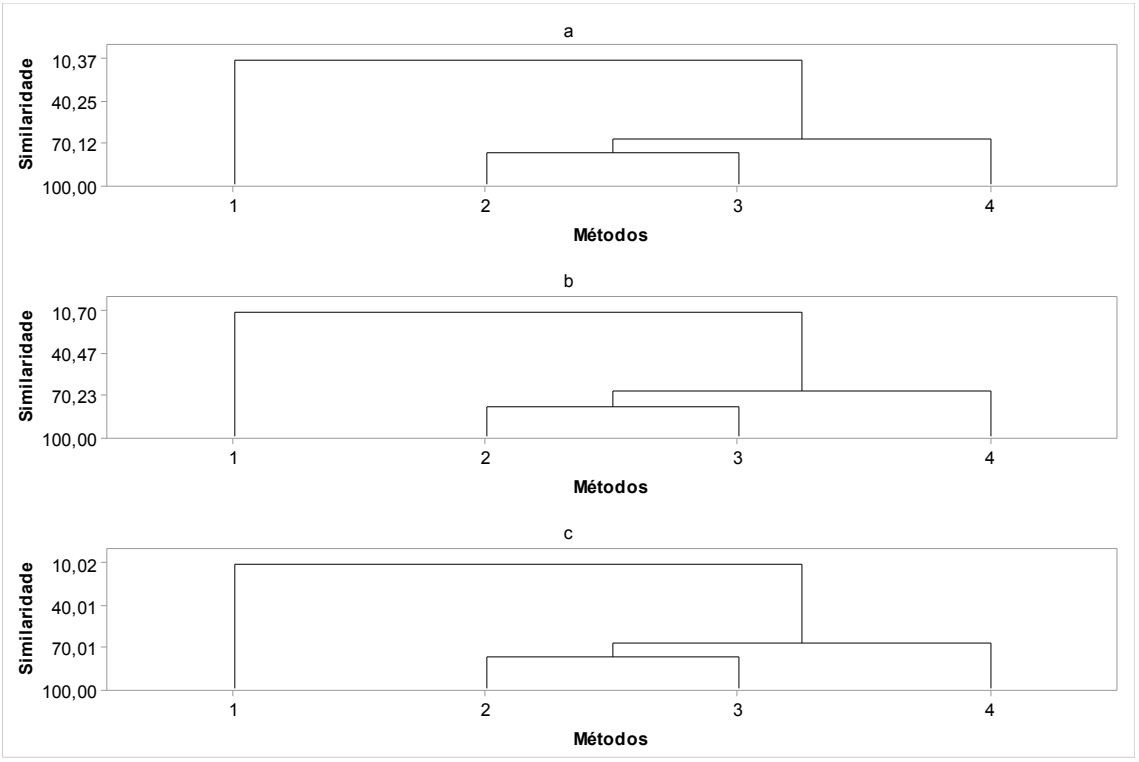


Figura 5.4 - Agrupamento dos quatro métodos (desvio-padrão total = 1, de Lenth = 2, de Juan e Pena = 3 e de Dong = 4) em função dos diferentes graus de correlação: 0,5 (a); 0,75 (b); 1 (c) e *outlier* igual a 1%.

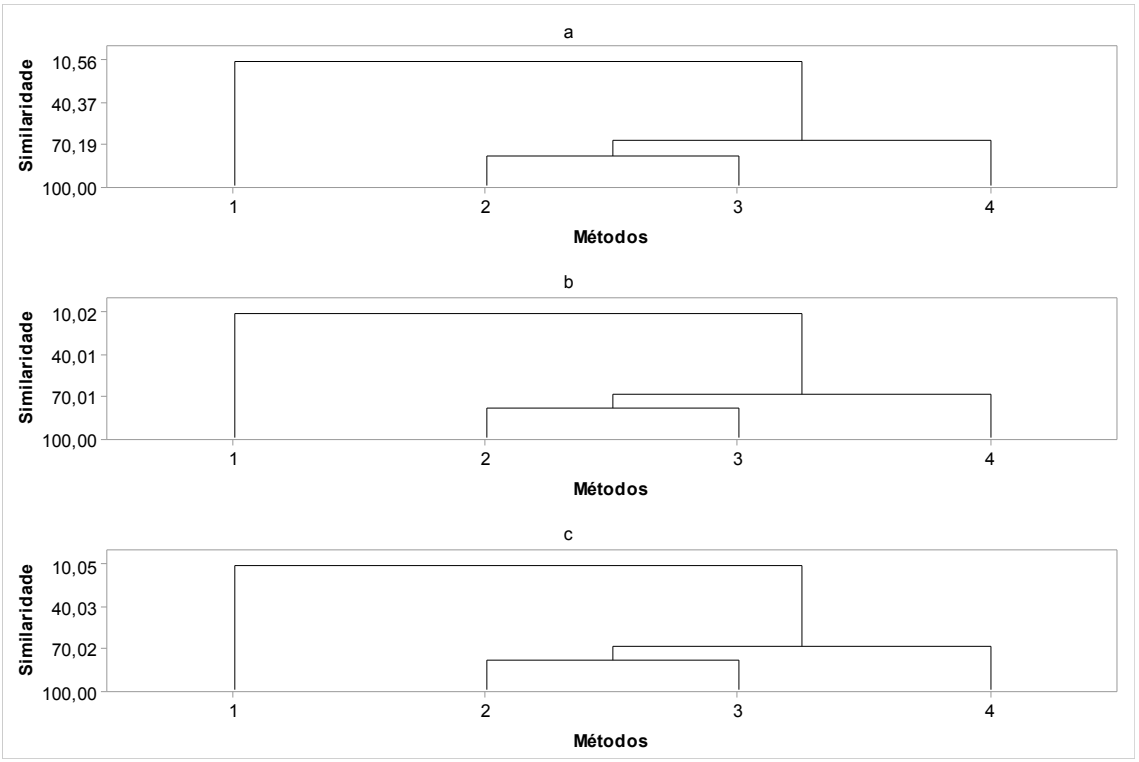


Figura 5.5 - Agrupamento dos quatro métodos (desvio-padrão total = 1, de Lenth = 2, de Juan e Pena = 3 e de Dong = 4) em função dos diferentes graus de correlação: 0,5 (a); 0,75 (b); 1 (c) e *outlier* igual a 2%.

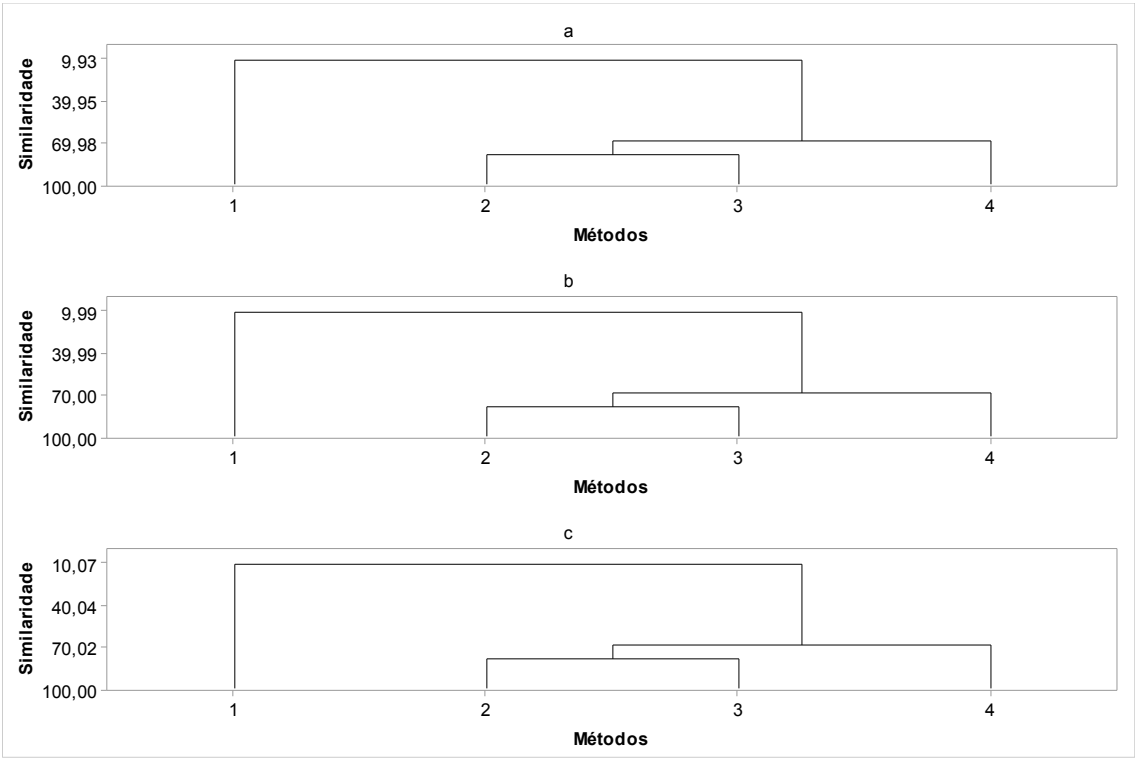


Figura 5.6 - Agrupamento dos quatro métodos (desvio-padrão total = 1, de Lenth = 2, de Juan e Pena = 3 e de Dong = 4) em função dos diferentes graus de correlação: 0,5 (a); 0,75 (b); 1 (c) e *outlier* igual a 3%.

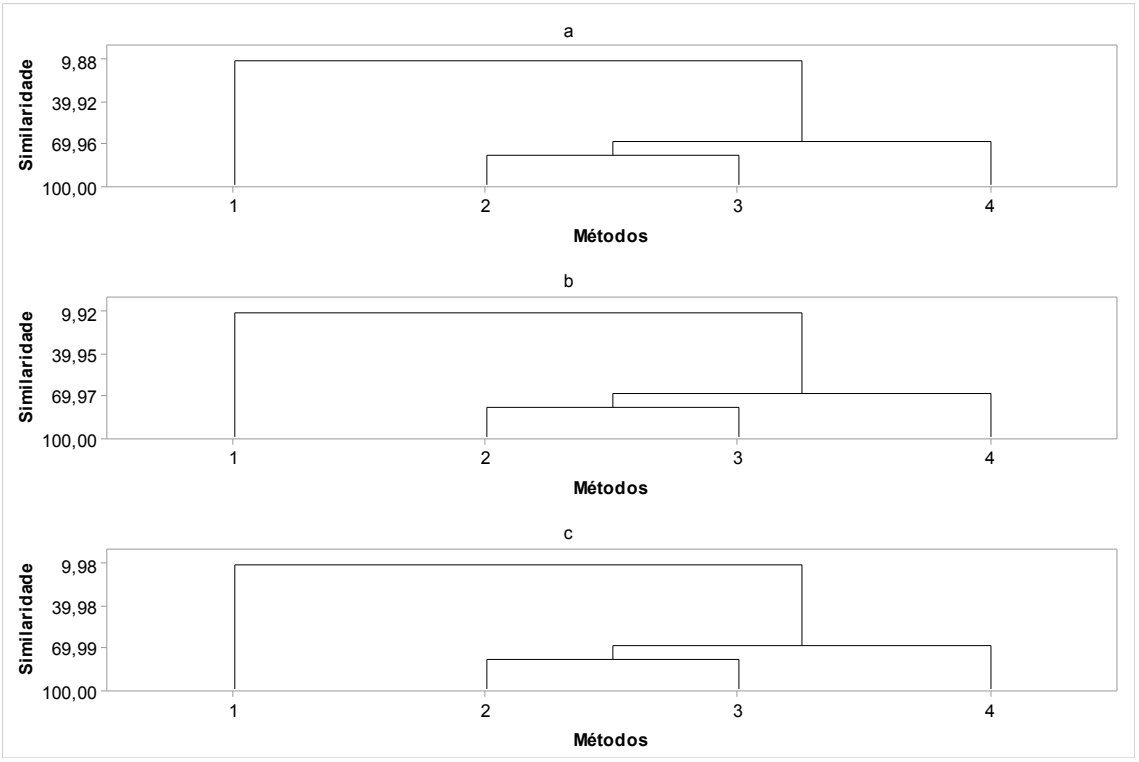


Figura 5.7 - Agrupamento dos quatro métodos (desvio-padrão total = 1, de Lenth = 2, de Juan e Pena = 3 e de Dong = 4) em função dos diferentes graus de correlação: 0,5 (a); 0,75 (b); 1 (c) e *outlier* igual a 4%.

No caso em que não se tem *outliers*, o esperado era que todos os métodos fossem iguais. Dado que o método do desvio-padrão total já é devidamente adequado quando se tem variáveis aleatórias, isso significa, de maneira contrária à expectativa, que os métodos de Juan e Pena e de Dong não apresentaram as mesmas estimativas, nem entre si, e nem em relação aos outros dois métodos, desvio-padrão total e Lenth.

Por outro lado, na presença de *outliers*, os métodos de Lenth, Juan e Pena e Dong forneceram as melhores estimativas de variâncias e covariâncias em comparação com o método do desvio-padrão total, que nessa condição forneceu estimativas inflacionadas. E quanto maior foi a adição dos *outliers*, maior foi a diferença desses grupos, o que implicou na recomendação dos métodos de Lenth, Juan e Pena e Dong e não na utilização do método do desvio-padrão total sempre que houver confirmação de *outliers*.

Com base na visualização da configuração do dendograma, notou-se a ausência do efeito das correlações sobre os agrupamentos. Portanto, foram obtidas as médias de Dvar e Dcov por grupo apenas em função dos percentuais dos *outliers* (Tabela 5.7).

Tabela 5.7 - Médias das variáveis analisadas de acordo com o agrupamento dos métodos em cada percentual de *outlier*

Outlier	Grupo	Método	Dvar	Dcov
0	1	Desvio-padrão total e Lenth	0,0184	0,0492
	2	Juan e Pena	0,0783	0,0627
	3	Dong	0,0285	0,2864
1	1	Desvio-padrão total	0,4212	2,1844
	2	Lenth, Juan e Pena e Dong	0,0405	0,1877
2	1	Desvio-padrão total	0,8383	4,3361
	2	Lenth, Juan e Pena e Dong	0,0503	0,1369
3	1	Desvio-padrão total	1,2360	6,4060
	2	Lenth, Juan e Pena e Dong	0,0544	0,1312
4	1	Desvio-padrão total	1,6396	8,503
	2	Lenth, Juan e Pena e Dong	0,0720	0,1273

Portanto, concluiu-se que o melhor método foi o de Lenth por fornecer melhores estimativas na presença e ausência de *outliers*, e se situar nos

melhores grupos nessas respectivas condições (Tabela 5.7). Hamada e Balakrishnan (1998) recomendam o método de Lenth, sobretudo pela facilidade nos cálculos, bem como por apresentar um bom desempenho.

5.2 Análise de Componentes Principais

Para dar início às análises dos CPs, foram escolhidos dois percentuais de *outliers*, 0 e 2%. Isto porque observou-se no gráfico dos ajustes das equações de regressão de Dvar que o método de Lenth foi eficiente em retirar os efeitos dos *outliers* sobre as estimativas das variâncias aleatórias somente até 2%, aproximadamente. Os *outliers*, dependendo da quantidade que apresentar no conjunto de dados, inflaciona a variância aleatória, ou, numa danificação maior no conjunto de dados por *outliers*, há uma deterioração no método do desvio-padrão total. Apesar dos tipos de variâncias não terem provocado efeito sobre a estimativa do desvio-padrão, ela foi incorporada na análise dos CPs para verificar o efeito das diferenças na variabilidade dos dados por diversos motivos, como por exemplo, o das diferentes escalas.

Como a correlação influenciou somente no método de Dong, fixou-se uma correlação de 0,75 e médias diferentes para prosseguir com as análises.

Os resultados relacionados às matrizes $\boldsymbol{S_{ZT}}$, $\boldsymbol{S_{ZL1}}$ e $\boldsymbol{S_{ZL2}}$ foram bons, pois proporcionaram estimativas dos coeficientes de correlação de CP_1 com a variável Z_1 e de CP_1 com a variável Z_2 e dos coeficientes de correlação de CP_2 com a variável Z_3 e de CP_2 com a variável Z_4 muito próximos entre si. Isso significa que as variáveis Z_1 e Z_2 apresentam-se correlacionadas igualmente com o CP_1 e as variáveis Z_3 e Z_4, também do mesmo modo, com o CP_2. A diferença ocorreu com as estimativas proporcionadas pela matriz $\boldsymbol{S_{ZCV}}$, que apesar de estimar componentes com autovalores semelhantes, ponderou a variável Z_2 com maior peso no CP_1 do que Z_1 e a variável Z_4 com maior peso no CP_2 do que Z_3. Vale a pena ressaltar que o coeficiente de variação de Z_2 foi maior que o de Z_1, e o de Z_4 maior do que o de Z_3 (Tabela 5.8). Entre as três primeiras matrizes, o melhor resultado foi o proporcionado pela $\boldsymbol{S_{ZL2}}$, dadas as maiores diferenças absolutas entre $r_{CP_1Z_1}$ e $r_{CP_2Z_1}$, entre $r_{CP_1Z_2}$ e $r_{CP_2Z_2}$, entre $r_{CP_1Z_3}$ e $r_{CP_2Z_3}$ e entre $r_{CP_1Z_4}$ e $r_{CP_2Z_4}$. As estimativas dos autovalores e autovetores normalizados estão apresentadas na Tabela 5.9.

Tabela 5.8 - Estimativas dos coeficientes de correlação entre os CPs e os Zs na ausência de *outliers* e variâncias iguais

	S_{ZT}		S_{ZCV}		S_{ZL1}		S_{ZL2}	
	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2
Z_1	-0,70	0,62	-0,79	0,24	-0,74	-0,57	0,87	0,26
Z_2	-0,71	0,62	-0,95	0,28	-0,75	-0,56	0,88	0,25
Z_3	-0,63	-0,69	-0,24	-0,79	-0,58	0,74	0,26	-0,87
Z_4	-0,63	-0,69	-0,29	-0,95	-0,57	0,74	0,26	-0,87

Tabela 5.9 - Estimativas dos autovetores e autovalores com ausência de *outliers* e variâncias iguais

	S_{ZT}		S_{ZCV}		S_{ZL1}		S_{ZL2}	
	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$
Z_1	-0,52	0,47	-0,36	0,11	-0,56	-0,43	0,66	0,20
Z_2	-0,53	0,47	-0,88	0,26	-0,55	-0,42	0,68	0,20
Z_3	-0,47	-0,53	-0,11	-0,36	-0,43	0,56	0,20	-0,67
Z_4	-0,47	-0,53	-0,26	-0,88	-0,42	0,55	0,20	-0,67
Auto-valores	1,77	1,72	0,01	0,01	1,72	1,68	17,46	16,82

Quando as variâncias foram diferentes, mas todas as quatro variáveis se comportarem como aleatórias, ou seja, sem a presença de *outliers*, os resultados deveriam ser equivalentes aos encontrados na Tabela 5.8, dado que as variâncias diferem apenas por questão de escala. No entanto, esse fato não ocorreu, e o melhor resultado foi obtido por meio da matriz S_{ZCV} (Tabela 5.10). As estimativas dos autovalores e autovetores normalizados estão apresentadas na Tabela 5.11.

Tabela 5.10 - Estimativas dos coeficientes de correlação entre os CPs e os Zs na ausência de *outliers* e variâncias diferentes

	S_{ZT}		S_{ZCV}		S_{ZL1}		S_{ZL2}	
	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2
Z_1	0,83	0,43	0,93	-0,09	0,66	0,67	0,99	-0,02
Z_2	0,83	0,43	0,93	-0,09	0,66	0,67	0,76	-0,02
Z_3	-0,43	0,83	-0,09	-0,93	-0,67	0,66	-0,02	-0,99
Z_4	-0,43	0,83	-0,09	-0,93	-0,67	0,66	-0,02	-0,74

Tabela 5.11 - Estimativas dos autovetores e autovalores com ausência de *outliers* e variâncias diferentes

	$\boldsymbol{S_{ZT}}$		$\boldsymbol{S_{ZCV}}$		$\boldsymbol{S_{ZL1}}$		$\boldsymbol{S_{ZL2}}$	
	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$
Z_1	0,62	0,32	0,70	-0,06	0,49	0,50	0,93	-0,02
Z_2	0,62	0,32	0,70	-0,06	0,49	0,50	0,36	-0,01
Z_3	-0,32	0,62	-0,06	-0,70	-0,50	0,49	-0,02	-0,94
Z_4	-0,32	0,62	-0,06	-0,70	-0,50	0,49	-0,01	-0,35
Auto-valores	1,75	1,74	0,01	0,01	1,78	1,77	11,22	11,05

Os resultados observados com os conjuntos de dados sem *outliers*, indicaram que as matrizes $\boldsymbol{S_{ZCV}}$ e $\boldsymbol{S_{ZL2}}$ têm comportamentos similares. A primeira sofre o efeito do coeficiente de variação, pois quando foram analisados os dados sem *outliers* e com variâncias iguais (Tabela 5.8), o coeficiente de variação da variável Y_2 foi maior do que o da variável Y_1. Por isso, o $r_{CP_1Z_2}$ foi maior do que o $r_{CP_1Z_1}$. O mesmo aconteceu quando foram comparadas as variáveis Z_3 e Z_4. Já a segunda matriz sofreu o efeito da escala, pois quando foram analisados os dados sem *outliers* e com variâncias diferentes (Tabela 5.10), a variância da variável Y_1 foi maior que a da Y_2. Por isso, o $r_{CP_1Z_1}$ foi maior do que o $r_{CP_1Z_2}$. O mesmo aconteceu com as variáveis Z_3 e Z_4. Nesse caso, isto é, sem a presença de *outliers*, as matrizes $\boldsymbol{S_{ZCV}}$ e $\boldsymbol{S_{ZL2}}$ não respondem adequadamente aos objetivos de ponderarem igualmente as variáveis nos respectivos CPs quando as variâncias ou coeficientes de variação medem variabilidade aleatória.

Na prática, como na grande maioria das vezes, não se tem variâncias iguais para todas as variáveis utilizadas na análise dos CPs, recomenda-se na ausência de *outliers*, a utilização da matriz $\boldsymbol{S_{ZT}}$.

Por outro lado, $\boldsymbol{S_{ZL1}}$ não conseguiu discriminar as variáveis Z_1 e Z_2 no CP_1 e Z_3 e Z_4 no CP_2, o que implicaria em interpretações erradas (Tabelas 5.8 e 5.10).

Portanto, se um determinado teste de aderência à normalidade, acusar que todas as variáveis são normais, recomenda-se utilizar o método do desvio-padrão total para a padronização das variáveis.

Observou-se na Tabela 5.12, que as matrizes $\boldsymbol{S_{ZT}}, \boldsymbol{S_{ZCV}}$ e $\boldsymbol{S_{ZL1}}$ proporcionaram resultados ruins, pois, apesar de terem possibilitado discriminar as variáveis Z_1 e Z_2 no CP_1 e Z_3 e Z_4 no CP_2, tal discriminação não

ocorreu com o mesmo peso. Já a matriz $\boldsymbol{S_{ZL2}}$ foi a que apresentou o melhor resultado, por ter discriminado com a mesma importância. As estimativas dos autovalores e autovetores normalizados estão apresentadas na Tabela 5.13.

Tabela 5.12 - Estimativas dos coeficientes de correlação entre os CPs e os Zs para 2% de *outliers* e variâncias iguais

	$\boldsymbol{S_{ZT}}$		$\boldsymbol{S_{ZCV}}$		$\boldsymbol{S_{ZL1}}$		$\boldsymbol{S_{ZL2}}$	
	$\boldsymbol{CP_1}$	$\boldsymbol{CP_2}$	$\boldsymbol{CP_1}$	$\boldsymbol{CP_2}$	$\boldsymbol{CP_1}$	$\boldsymbol{CP_2}$	$\boldsymbol{CP_1}$	$\boldsymbol{CP_2}$
Z_1	-0,83	0,50	-0,80	0,42	0,83	0,49	0,85	0,31
Z_2	-0,83	0,49	-0,86	0,51	0,83	0,49	0,86	0,30
Z_3	-0,83	-0,49	-0,79	-0,45	0,83	-0,50	0,31	-0,85
Z_4	-0,83	-0,50	-0,83	-0,54	0,82	-0,50	0,31	-0,85

Tabela 5.13 - Autovetores e autovalores para 2% de *outliers* e variâncias iguais

	$\boldsymbol{S_{ZT}}$		$\boldsymbol{S_{ZCV}}$		$\boldsymbol{S_{ZL1}}$		$\boldsymbol{S_{ZL2}}$	
	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$	$\boldsymbol{\hat{e}_1}$	$\boldsymbol{\hat{e}_2}$
Z_1	-0,49	0,50	-0,31	0,26	0,50	0,50	0,66	0,24
Z_2	-0,50	0,49	-0,66	0,64	0,50	0,49	0,67	0,24
Z_3	-0,50	-0,49	-0,30	-0,28	0,50	-0,50	0,24	-0,66
Z_4	-0,49	-0,50	-0,62	-0,67	0,49	-0,49	0,24	-0,67
Auto-valores	1,76	1,75	0,01	0,01	4,71	1,68	17,56	16,9

Já no caso em que se tem 2% de *outliers* e variâncias diferentes, a matriz que forneceu os resultados mais satisfatórios foi a $\boldsymbol{S_{ZL2}}$ (Tabela 5.14), pois foi a única que conseguiu agrupar as variáveis Z_1 e Z_2 no CP_1 e Z_3 e Z_4 no CP_2. O problema foi que esse agrupamento não ocorreu com o mesmo peso para as duas variáveis, respectivamente. Isso significa que nessa condição, ou seja, com a presença de *outliers* e com variâncias diferentes, recomenda-se a utilização dessa matriz. As estimativas dos autovalores e autovetores normalizados estão apresentadas na Tabela 5.15.

Tabela 5.14 - Estimativas dos coeficientes de correlação entre os CPs e os Zs para 2% de *outliers* e variâncias diferentes

	S_{ZT}		S_{ZCV}		S_{ZL1}		S_{ZL2}	
	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2	CP_1	CP_2
Z_1	0,82	0,50	-0,83	0,49	-0,82	0,51	0,99	-0,02
Z_2	0,83	0,50	-0,83	0,49	-0,82	0,51	0,76	-0,02
Z_3	0,82	-0,50	-0,82	-0,51	-0,83	-0,49	-0,02	-0,99
Z_4	0,82	-0,50	-0,82	-0,51	-0,83	-0,49	-0,02	-0,75

Tabela 5.15 - Autovetores e autovalores para 2% de *outliers* e variâncias diferentes

	S_{ZT}		S_{ZCV}		S_{ZL1}		S_{ZL2}	
	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_1$	$\hat{e}_2$
Z_1	0,49	0,50	-0,50	0,49	-0,49	0,51	0,93	-0,02
Z_2	0,50	0,49	-0,50	0,49	-0,49	0,50	0,36	-0,01
Z_3	0,49	-0,50	-0,49	-0,50	-0,50	-0,49	-0,02	-0,94
Z_4	0,49	-0,49	-0,49	-0,50	-0,51	-0,48	-0,01	-0,35
Auto-valores	2,72	1,00	0,01	0,01	4,76	1,76	11,53	11,06

Se houver *outliers*, as variáveis não devem ser igualmente ponderadas, como no caso da matriz S_{ZT}, S_{ZCV} e S_{ZL1} para as estimativas dos coeficientes de correlação. Por outro lado, independentemente se as variâncias forem iguais ou diferentes, mas com *outliers*, recomenda-se S_{ZL2}.

Posteriormente plotou-se os diagramas de dispersão dos escores de CP_1 e CP_2 obtidos com base no conjunto de dados na ausência e com 2% de *outliers*, variâncias iguais e diferentes, e desvio padrão estimados pelos métodos do desvio-padrão total e de Lenth.

Pelos diagramas de dispersão dos escores do CP_1 pelo de CP_2, expostos na Figura 5.16, apesar das matrizes S_{ZT} e S_{ZL2} terem fornecido os melhores resultados na ausência de *outliers*, verificou-se em termos de configurações, resultados semelhantes. Nesses gráficos, observou-se apenas um grupo de elementos, dado que todos os valores diferem-se aleatoriamente entre si. Nos diagramas com a presença de *outliers*, percebeu-se também as mesmas configurações para as matrizes S_{ZT}, S_{ZCV}, S_{ZL1}, S_{ZL2}. Nesses gráficos observou-se a formação de dois grupos de elementos, um para os valores aleatórios e o outro para os *outliers* (Figura 5.17).

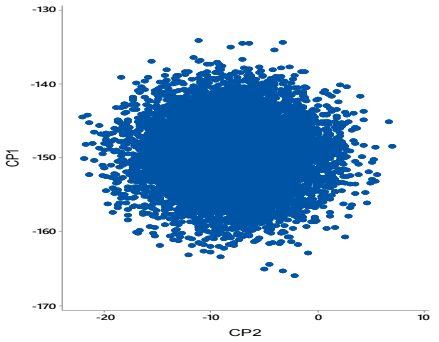
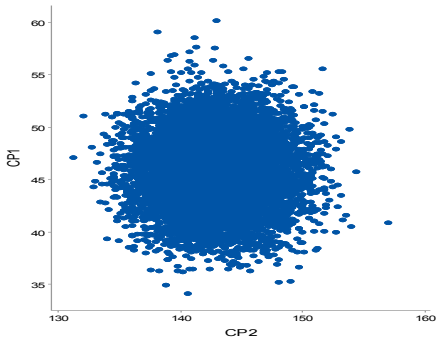
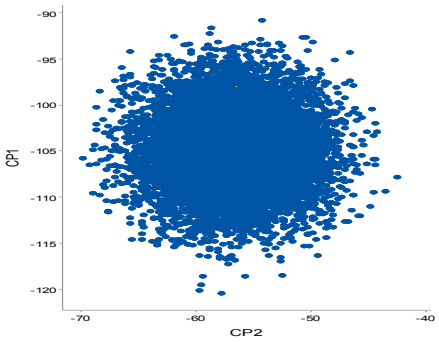
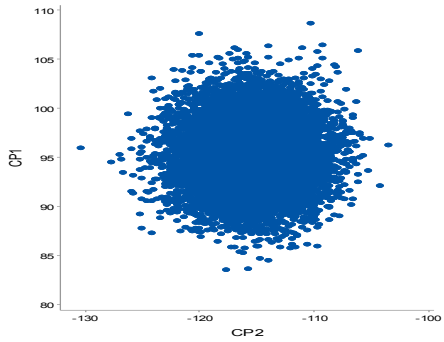
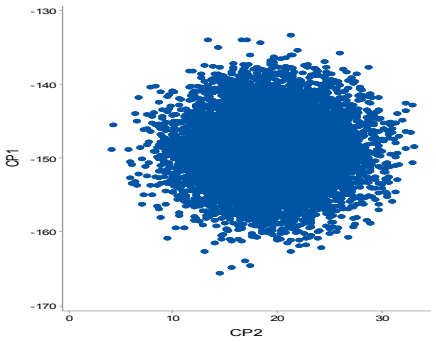
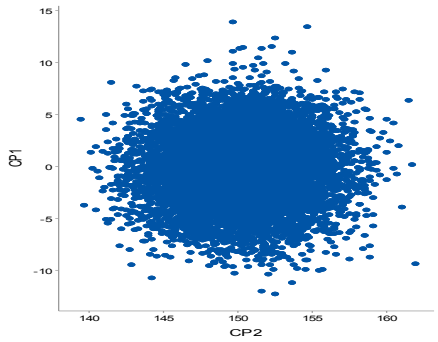
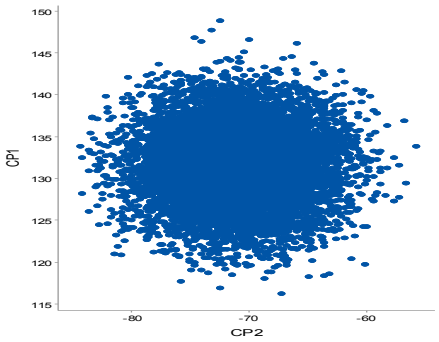
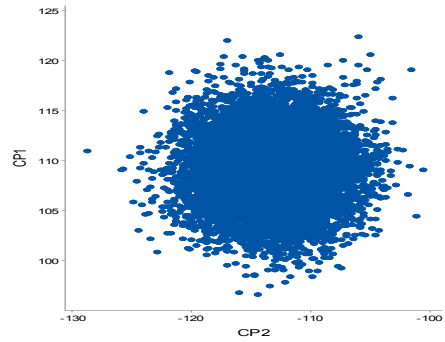
Matriz	*Outlier* = 0	
	Variâncias iguais	Variâncias diferentes
S_{ZT}		
S_{ZCV}		
S_{ZL1}		
S_{ZL2}		

Figura 5.8 - Diagrama de dispersão dos escores do CP_1 e CP_2 obtidos de acordo com o conjunto de dados com 2% de *outliers*, variâncias iguais e diferentes, e desvio padrão estimados pelos métodos desvio-padrão total e Lenth.

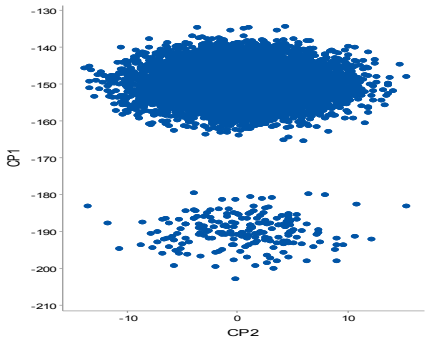
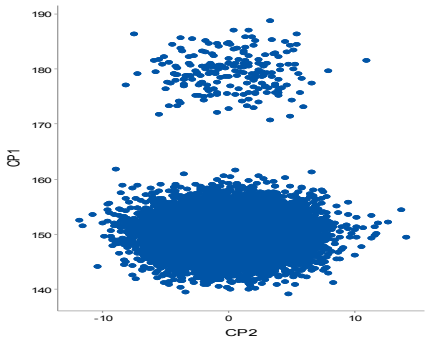
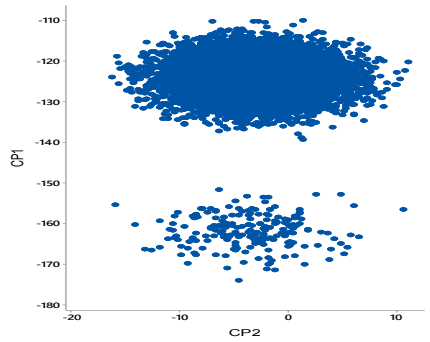
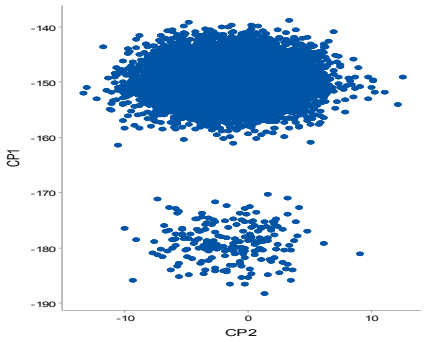
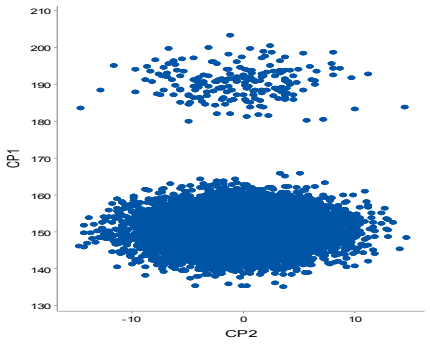
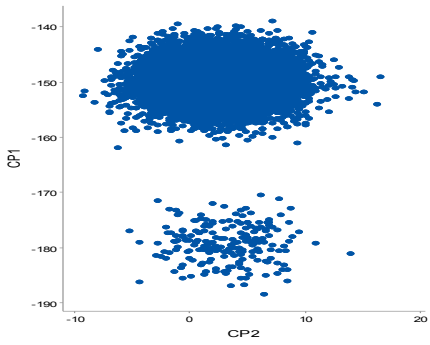
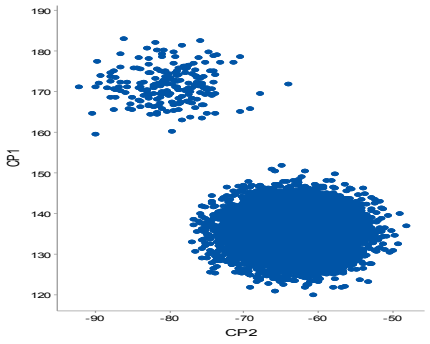
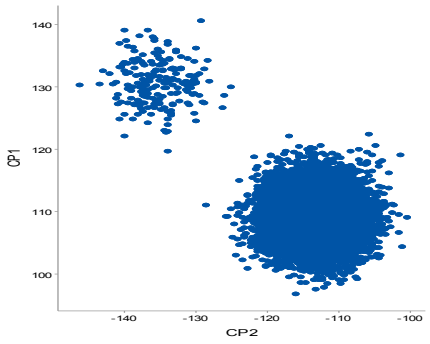
Matriz	*Outlier* = 2%	
	Variâncias iguais	Variâncias diferentes
S_{ZT}		
S_{ZCV}		
S_{ZL1}		
S_{ZL2}		

Figura 5.9 - Diagrama de dispersão dos escores do CP_1 e CP_2 obtidos de acordo com o conjunto de dados na ausência de *outliers*, variâncias iguais e diferentes, e desvio padrão estimados pelos métodos desvio-padrão total e Lenth.

A fim de comparar as diferentes matrizes de covariâncias, nas condições de dois tipos de variâncias e dois tipos de *outliers*, considerando-se médias iguais e correlação de 0,75, foram realizadas as análises de variâncias (Tabela 5.16), e para isso escolheu-se considerar as interações de alta ordem como sendo não ativas, e daí utilizá-las para estimar o erro (HUNTER, 2005). Realizaram-se, também, os testes de Tukey para as comparações de médias (Tabelas 5.17, 5.18, 5.19, 5.20 e 5.21).

Tabela 5.16 - Quadrados médios da análise de variância

		QM				
FV	GL	V_1	V_2	V_3	V_4	V_5
M	3	0,0433*	0,0128*	0,0005*	0,0007	0,0140*
O	1	0,4087*	0,0008	0,0000	0,0001	0,0003
V	1	0,0003	0,0029	0,0006*	0,0014*	0,0033*
M x O	3	0,0435*	0,0007	0,0000	0,0001	0,0003
M x V	3	0,0001	0,0191*	0,0005*	0,0008*	0,0249*
O x V	1	0,0001	0,0009	0,0002	0,0001	0,0002
Resíduo	3	0,0001	0,0007	0,0001	0,0001	0,0003

* significativo pelo teste F (P < 0,10); M = tipos de matrizes; O = percentuais de *outliers*; V = tipos de variâncias.

Tabela 5.17 - Estimativas das médias da variável V_1 estratificada por matriz e *outliers*

Matriz	*Outliers*	
	0	2%
$\boldsymbol{S_{ZT}}$	0,0066 a B	0,4345 a A
$\boldsymbol{S_{ZCV}}$	0,0169 a B	0,4342 a A
$\boldsymbol{S_{ZL1}}$	0,0066 a B	0,4347 a A
$\boldsymbol{S_{ZL2}}$	0,0114 a A	0,0167 b A

Médias seguidas pela mesma letra minúscula, na coluna, e pela mesma letra maiúscula, na linha, não diferem entre si pelo teste de Tukey (P > 0,10).

Tabela 5.18 - Estimativas das médias da variável V_2 estratificada por matriz e variâncias

Matriz	Variâncias	
	Iguais	Diferentes
S_{ZT}	0,0044 b A	0,0015 b A
S_{ZCV}	0,1104 a A	0,0024 b B
S_{ZL1}	0,0028 b A	0,0014 b A
S_{ZL2}	0,0117 b B	0,2318 a A

Médias seguidas pela mesma letra minúscula, na coluna, e pela mesma letra maiúscula, na linha, não diferem entre si pelo teste de Tukey (P > 0,10).

Tabela 5.19 - Estimativas das médias da variável V_3 estratificada por matriz e variâncias

Matriz	Variâncias	
	Iguais	Diferentes
S_{ZT}	0,0027 b A	0,0013 a A
S_{ZCV}	0,0469 a A	0,0017 a B
S_{ZL1}	0,0045 b A	0,0019 a B
S_{ZL2}	0,0008 b A	0,0027 a A

Médias seguidas pela mesma letra minúscula, na coluna, e pela mesma letra maiúscula, na linha, não diferem entre si pelo teste de Tukey (P > 0,10).

Tabela 5.20 - Estimativas das médias da variável V_4 estratificada por matriz e variâncias

Matriz	Variâncias	
	Iguais	Diferentes
S_{ZT}	0,0058 a A	0,0029 a A
S_{ZCV}	0,0623 b A	0,0017 a B
S_{ZL1}	0,0072 b A	0,0038 a A
S_{ZL2}	0,0086 b A	0,0017 a A

Médias seguidas pela mesma letra minúscula, na coluna, e pela mesma letra maiúscula, na linha, não diferem entre si pelo teste de Tukey (P > 0,10).

Tabela 5.21 - Estimativas das médias da variável V_5 estratificada por matriz e variâncias

Matriz	Variâncias	
	Iguais	Diferentes
$\boldsymbol{S_{ZT}}$	0,0041 b A	0,0013 b A
$\boldsymbol{S_{ZCV}}$	0,1297 a A	0,0009 b B
$\boldsymbol{S_{ZL1}}$	0,0023 b A	0,0005 b A
$\boldsymbol{S_{ZL2}}$	0,0011 b B	0,2489 a A

Médias seguidas pela mesma letra minúscula, na coluna, e pela mesma letra maiúscula, na linha, não diferem entre si pelo teste de Tukey (P > 0,10).

Para V_1, os tipos de variâncias não influenciaram (P > 0,10), havendo interação (P < 0,10) entre matrizes e *outliers*. Todas as matrizes apresentaram bons resultados na ausência de *outliers* e a matriz que obteve melhor resultado para 2 % de *outliers* foi a $\boldsymbol{S_{ZL2}}$.

Para as variáveis V_2, V_3, V_4 e V_5, ocorreu interação (P < 0,10) entre as matrizes e os tipos de variância. Isso significou que os quatro métodos, mas de modos diferentes, foram influenciados pelas diferentes variâncias. Desse modo, conclui-se que o ponto mais importante na tomada de decisão em uma análise de CP's, ainda continua sem a recomendação de um método adequado.

No entanto, e de acordo com os resultados obtidos, as melhores matrizes foram: $\boldsymbol{S_{ZL1}}$ e $\boldsymbol{S_{ZL2}}$ para variâncias iguais e $\boldsymbol{S_{ZT}}$, $\boldsymbol{S_{ZCV}}$ e $\boldsymbol{S_{ZL1}}$ para variâncias diferentes.

Por outro lado, a proximidade de valores das variáveis V_2, V_3, V_4 e V_5 próximos de zero não significa, necessariamente, que a matriz foi adequada, mas simplesmente que as estimativas foram próximas entre si. Esse resultado pode ocorrer nos casos em que as estimativas são próximas mas erradas.

6 CONCLUSÕES

O método do desvio-padrão total sofre mais efeito da adição de *outliers* do que os métodos de Lenth, Juan e Pena e Dong. Portanto, ele é indicado somente quando todas as variáveis estudadas são aleatórias, isto é, sem a presença de *outliers*.

Na ausência de *outliers*, o método de estimação que mais se aproximou do método do desvio-padrão total foi o método de Lenth.

Os métodos de estimação de Lenth, Juan e Pena e Dong apresentaram o mesmo desempenho, de maneira geral, quando há presença de até 4% de *outliers*.

O método de Juan e Pena não reconhece diferentes percentuais de *outliers* no intervalo de até 4%.

O método de Dong tem o desempenho afetado pelo grau de relação.

O método de Lenth apresentou o melhor desempenho no que diz respeito às estimativas das variâncias e covariâncias, especificadamente.

Os diferentes tipos de médias e variâncias não afetaram as estimativas dos desvios-padrão pelos métodos do desvio padrão total, Lenth, Juan e Pena e Dong.

A matriz de covariâncias estimada pelo método do desvio-padrão total é adequada para a análise de componentes principais, quando todas as variáveis são aleatórias (ausência de *outliers*) e com variâncias diferentes.

Na ausência de *outliers* e com variâncias iguais, a matriz de covariâncias estimada com base no método de Lenth, em que os cálculos das covariâncias foram realizados sobre os valores considerados como aleatórios por este método, é a recomendada. Da mesma forma, nos casos em que há presença de outliers e com variâncias iguais ou diferentes.

REFERÊNCIAS BIBLIOGRÁFICAS

CAMPANA, A. C. M.; RIBEIRO JÚNIOR, J. I.; NASCIMENTO, M. Uma proposta de transformação de dados para a análise de componentes principais. **Revista Brasileira de Biometria**. v. 28, p. 1-15, 2010.

DEVLIN, S. J.; GNANADESIKAN, R.; KETTENRING, J. R. Robust evaluation of dispersion matrices and principal components. **Journal of the American Statistical Association**, 76, p. 354–362,1981.

DONG, F. On the identification of active contrasts in unreplicated fractional factorials. **Statistica Sinica**, v. 3, p. 209-217, 1993.

FERREIRA, D. F. **Estatística multivariada.** 2 ed. Lavras: Editora UFLA, 2009. 676 p.

HAMADA, M.; BALAKRISHNAN, N. Analyzing unreplicated factorial experiments: a review with some new proposals. **University of Michigan and McMaster University**, 1998.

HONGYU, K. **Comparação do GGEbiplot ponderado e AMMI-ponderado com outros modelos de interação genótipo × ambiente.** 155p. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2015.

HOTELLING, H. Review of the triumph of mediocrity in business. **Journal of the American Statistical Association.** New York, v. 28, n. 184, p. 463 – 465, Dec. 1933.

HUNTER, J. S. **Statistics for experimenters: design, innovation, and discovery**. 2. ed. Hoboken, N.J.: Wiley-Interscience, 2005.

JOHNSON, R. A; WICHERN, D. W. Applied multivariate statistical analysis. 5. Ed. **New Jersey: Prentice Hall**, 2002.

JOLLIFFE, I. T. Principal component analysis. 2nd. edition. **New York: Springer-Verlag**, 2002. 520p.

JUAN, J.; PENA, D. A simple method to identify significant effects in unreplicated two-level factorial designs. Comm. in Statist. **Theory and Methods**, 1992.

KAYSER, J. TENKE, C. E. Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation. **Clinical Neurophysiology**, USA, 2003.

KUBRUSLY, L. S. Um procedimento para calcular índices a partir de uma base de dados multivariados. **Revista Pesquisa Operacional**, v. 21, n.1. 2001.

LAWSON, J. SAS macros for analysis of unreplicated 2k and 2k-p designs with a possible outlier. **Journal of Statistical Software**, v. 25, 2008.

LENTH, R. V. Quick and easy analysis of unreplicated factorials. **Technometrics**, v. 31, n. 4, p. 469-473, 1989.

MANLY, B. J. F. **Métodos estatísticos multivariados: uma introdução.** Porto Alegre: Bookman, 2008.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of cluster in a data set. **Psychometrika**, v.50, p.159-179, 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada.** Belo Horizonte: Editora UFMG, 2007. 297 p.

R DEVELOPMENT CORE TEAM (2007). **R: a language and environment for statistical computing. R fundation for statistical computing**. Vienna, Austria. ISBN 3-900051-07-0. Disponível em: http://r-project .org

REIS, E. **Estatística multivariada aplicada.** Lisboa: Edições Sílabo, 1997.

SANDANIELO, V. L. M. **Emprego de técnicas estatísticas na construção de índices de desenvolvimento sustentável aplicados a assentamentos rurais.** 175p. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agronômicas de Botucatu, 2008.

TRIOLA, M. F. **Introdução à estatística**. 7. ed. Rio de Janeiro: LTC Editora, 1999. 410 p.