IDENTIFICAÇÃO DE *OUTLIERS* VIA COMPONENTES PRINCIPAIS COM AMOSTRAS CORRIGIDAS POR DISTÂNCIAS DO TIPO QUI-QUADRADO

MANOEL VÍTOR DE SOUZA VELOSO

2010

MANOEL VÍTOR DE SOUZA VELOSO

IDENTIFICAÇÃO DE *OUTLIERS* VIA COMPONENTES PRINCIPAIS COM AMOSTRAS CORRIGIDAS POR DISTÂNCIAS DO TIPO QUI-QUADRADO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

Orientador
Prof. Dr. Marcelo Angelo Cirillo

LAVRAS
MINAS GERAIS -BRASIL
2010

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Veloso, Manoel Vítor de Souza.

Identificação de *outliers* via componentes principais com amostras corrigidas por distâncias do tipo qui-quadrado / Manoel Vítor de Souza Veloso. – Lavras : UFLA, 2010.

58 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2010.

Orientador: Marcelo Angelo Cirillo.

Bibliografia.

1. Curtose. 2. MAD. 3. Normal contaminada. 4. Monte Carlo. 5. Bootstrap. I. Universidade Federal de Lavras. II. Título.

CDD-519.56

MANOEL VÍTOR DE SOUZA VELOSO

IDENTIFICAÇÃO DE *OUTLIERS* VIA COMPONENTES PRINCIPAIS COM AMOSTRAS CORRIGIDAS POR DISTÂNCIAS DO TIPO QUI-QUADRADO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 19 de fevereiro de 2010

Prof. Dr. João Domingos Scalon	UFLA
Profa. Dra. Thelma Sáfadi	UFLA
Prof. Dr. Marcelo Tavares	UFU

Prof. Dr. Marcelo Angelo Cirillo
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

À minha amada esposa Luciana,
pelo amor, amizade, carinho, companheirismo, paciência e apoio diário.

DEDICO

Aos meus pais, Celso Veloso e
Maria Aparecida Veloso (in memorian),
aos meus irmãos Celso, Eliane, Cláudia, Flávio e Marcos
por todo amor, cuidado, apoio, confiança, amizade e compreensão
e à Abigail Emília, pelo amor, apoio e carinho especial.

OFEREÇO

AGRADECIMENTOS

Aos amigos que, mesmo distantes, acreditaram e me apoiaram em toda minha trajetória.

Aos amigos Fábio Mathias Corrêa, Ivan Allaman e Walmes Marques Zeviani pela amizade, companheirismo e apoio.

Aos professores do DEX-UFLA, Júlio Silvio de Sousa Bueno Filho, Daniel Furtado Ferreira e Mario Ferrua Vivanco e da UFU, Rogério de Melo Costa Pinto e Ednaldo Guimarães pela amizade, apoio e ensinamentos.

Aos membros da banca, professores do DEX-UFLA João Domingos Scalon, Thelma Sáfadi e, em especial, ao professor Marcelo Tavares da UFU pela amizade, recomendações e confiança.

Ao meu orientador Marcelo Ângelo Cirillo pelas orientações, ensinamentos, confiança, paciência e amizade.

À Universidade Federal de Lavras e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, bem como a todos os funcionários do DEX-UFLA.

Ao CNPq pela bolsa de estudos, essencial para a realização deste trabalho.

SUMÁRIO

Página

LISTA DE TABELAS

Página

LISTA DE FIGURAS

Página

RESUMO

VELOSO, Manoel Vítor de Souza. **Identificação de *outliers* via componentes principais com amostras corrigidas por distâncias do tipo qui-quadrado.** 2010. 58 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras. *

Dentre as inúmeras técnicas utilizadas para identificar *outliers* no âmbito do contexto p-dimensional, a técnica de Componentes Principais tem sido amplamente utilizada. Diante disso, este trabalho teve por objetivo propor um teste de significância baseado nos coeficientes de curtose robustos, com a finalidade de evidenciar, estatisticamente, qual componente é mais apropriado para a identificação dos *outliers* multivariados. Com este propósito, procedeu-se a um estudo Monte Carlo, considerando diferentes números de variáveis, tamanhos de amostras, porcentagem de contaminação da mistura de distribuições e diferentes correções por distâncias do tipo qui-quadrado aplicadas nas amostras. Por fim, diante das conclusões do estudo realizado, recomenda-se tal teste de significância para amostras corrigidas por distâncias do tipo qui-quadrado de Pearson.

Palavras-chaves: Curtose. MAD. Normal contaminada. Monte Carlo. Bootstrap.

* **Orientador:** Marcelo Angelo Cirillo - UFLA.

ABSTRACT

VELOSO, Manoel Vitor de Souza. **Identification of *outliers* by principal components with samples corrected for distances of type chi-square.** 2010. 58 p. Dissertation (Master of Statistics and Agricultural Experimentation) - Federal University of Lavras, Lavras. *

Among the many techniques used to identify *outliers* within the context of p-dimensional, the technique of principal components has been widely used. Thus, this study aimed to propose a test of significance based on robust kurtosis coefficients, in order to show statistically which component is most appropriate for identifying multivariate *outliers*. For this purpose, we proceeded to a Monte Carlo study, considering different numbers of variables, sample size, percentage of contamination of the mixture of different distributions and corrections to distances of type chi-square applied to the samples. Finally, given the findings of the study, it is recommended that test of significance for samples of type distances corrected by chi-square test.

Keywords: Kurtosis. MAD. Contaminated normal. Monte Carlo. Bootstrap.

* **Major Professor:** Marcelo Angelo Cirillo - UFLA.

1 INTRODUÇÃO

Em se tratando da identificação de *outliers* com o uso da técnica de Análise de Componentes Principais (ACP), a utilização de métodos robustos é de extrema importância para que a classificação de observações discrepantes possa ser feita com melhor segurança e confiabilidade. Nesse contexto, cita-se Critchley (1985), que discutiu a influência dos *outliers* pela Curva de Influência, baseada na Influência Global, que envolve a deleção de algumas observações, usando-se a técnica de componentes principais. Um método alternativo para avaliar o efeito local de pequenas perturbações nos dados foi proposto por Cook (1986), tendo por base a curvatura normal, estruturada na verossimilhança.

Dessa forma, segundo Cook (1986), a robustez das estimativas das componentes fornecidas pelo modelo, mediante pequenas perturbações sofridas no modelo ou nos dados, poderá ser contornada não exigindo a deleção de observações e permitindo avaliar a influência conjunta de todos os pontos.

Segundo Jolliffe (2002), a criação de testes mais formais para a identificação de *outliers* com base em Componentes Principais (CP) é possível, assumindo que os CP são normalmente distribuídos. A rigor, isso pressupõe que um conjunto de dados X tem uma distribuição normal multivariada e que, sendo os CP funções lineares das p variáveis aleatórias originais, e usando o recurso do Teorema do Limite Central, pode-se justificar a normalidade aproximada para os CPs mesmo quando as variáveis originais não possuem distribuições normais.

Filzmoser et al. (2008) buscam superar as limitações dos procedimentos clássicos na identificação de *outliers*, propondo um método de fácil implementação computacional, capaz de identificar *outliers* em altas dimensões. Contudo, vale ressaltar que o método proposto por esses autores consiste apenas na descrição de

um procedimento que consistiu em aplicar uma reescalonagem dos dados por meio da mediana (med) e do Devio Absoluto da Mediana (Median Absolute Deviation - MAD). Assim sendo, tendo por base estas informações e com a finalidade de se chegar a melhorias metodológicas no que foi proposto por Filzmoser et al. (2008), este trabalho teve por objetivos:

- propor e auxiliar, via o método Monte Carlo, as correções por distâncias do tipo qui-quadrado para as unidades amostrais a serem utilizadas na identificação de *outliers* por meio da técnica de componentes principais;
- propor um teste de significância com a finalidade de inferir que o primeiro coeficiente de curtose padronizado seja nulo e identifique as observações *outliers* mediante as correções qui-quadrados sugeridas;
- aplicar a metodologia a um conjunto de dados reais referente ao índice de vendas no comércio varejista do estado de Minas Gerais, coletados e divididos em nove segmentos de mercado.

2 REFERENCIAL TEÓRICO

2.1 Algumas distribuições multivariadas específicas

As distribuições multivariadas podem ser construídas por meio de uma distribuição univariada qualquer.

No contexto do presente estudo, necessário se torna recorrer ao conhecimento elaborado e sistematizado no que diz respeito à sensibilidade de um teste ou de um processo de estimação por regiões, aos devios de normalidade dos dados.

Apresenta-se, a seguir, sucintamente, um resumo das distribuições multivariadas normal, normal contaminada e *t-Student*, todas da família de distribuições elípticas, relevantes para se avaliar tal sensibilidade.

2.1.1 Distribuição normal multivariada

Dado o vetor $\boldsymbol{X} = [X_1, \ldots, X_p]'$ em que cada componente X_i, $i = 1, \ldots, p$, é uma variável aleatória com distribuição normal de probabilidade com média μ e variância σ^2; a distribuição conjunta destes componentes gera a distribuição normal multivariada. A distribuição de $\boldsymbol{X}$ é denotada por $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\mu}$ é um vetor de médias de dimensão $p \times 1$ sendo $\mu_i = \mathrm{E}(X_i)$ e a matriz de covariância por $\boldsymbol{\Sigma}$ de dimensão $p \times p$, sendo o (i,j)-ésimo elemento a $Cov(X_i, X_j)$, $i = 1,...,p$ e $j = 1,...,p$ (Johnson, 1987). Ainda segundo este autor, se a matriz $\boldsymbol{\Sigma}$ for singular, então a distribuição de probabilidade de $\boldsymbol{X}$ estará confinada no subespaço em $\mathbb{R}^p$. Se a matriz $\boldsymbol{\Sigma}$ tem posto completo p, então a função densidade de probabilidade de $\boldsymbol{X}$ será definida como

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right], \tag{2.1}$$

com suporte em $\mathbb{R}^p$. A distribuição de $\boldsymbol{X}$ pode ser representada com uma transformação linear de p variáveis normais independente em $\boldsymbol{Y} = [Y_1,Y_2,...,Y_p]'$,

$$\boldsymbol{X} = A\boldsymbol{Y} + \boldsymbol{\mu} \tag{2.2}$$

sendo A qualquer matriz de dimensão $(p \times p)$ para o qual $A'A = \boldsymbol{\Sigma}$. Para muitas aplicações com simulação Monte Carlo, o caso em que $\boldsymbol{\Sigma}$ tem posto completo, a densidade citada na equação 2.1 será suficiente (Johnson, 1987).

2.1.2 Distribuição normal multivariada contaminada

A distribuição normal multivariada contaminada é uma outra distribuição muito importante para o estudo proposto.

Dado o vetor aleatório $\boldsymbol{X} = [X_1, \ldots, X_p]' \in \mathbb{R}^p$ com distribuição normal multivariada contaminada, sua função densidade de probabilidade será

$$\begin{aligned} f(\boldsymbol{x}) = \delta(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right] + \\ +(1-\delta)(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] \end{aligned} \tag{2.3}$$

em que δ é a probabilidade que o processo tem de ser realizado por $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, $(1-\delta)$ é a probabilidade que o processo tem de ser realizado por $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, $\boldsymbol{\Sigma}_i$ é uma matriz positiva definida, $\boldsymbol{\mu}_i \in \mathbb{R}^p$ é o vetor de médias, $i = 1,2$ e $0 \leq \delta \leq 1$.

Segundo Johnson (1987), a geração de variáveis estatísticas a partir da equação 2.3 é fácil e pode ser realizada como a seguir:

I. Gerar um valor u de uma distribuição uniforme contínua, com valores entre 0 e 1. Se $u \leq \delta$, avance para o passo II. Caso contrário, execute o passo III.

II. Gerar $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$.

III. Gerar $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

O problema em usar a equação 2.3, segundo Johnson (1987), não está na geração da variável mas, sim, na seleção dos parâmetros. O número de parâmetros na equação 2.3 é $p^2 + 3p + 1$, os quais correspondem a $2p$ médias, $2p$ variâncias, $p^2 - p$ correlações e a probabilidade γ de mistura.

Recorre-se à Figura 1, com gráficos de contorno, para ilustrar a distribuição normal bivariada, de modo a favorecer a compreensão das distribuições normais multivariadas enquanto abordagens mais complexas e análogas.

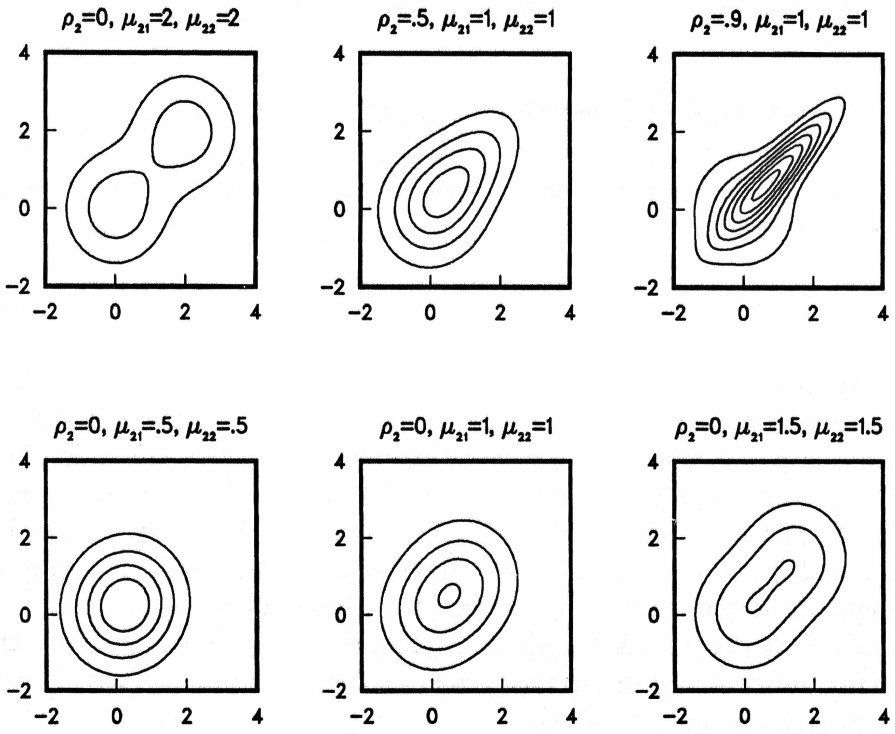

FIGURA 1 Misturas de distribuições normais bivariadas, sendo $\gamma = 0{,}5$, $\rho_1 = 0$, $\mu_{11} = 0$, $\mu_{12} = 0$, $\sigma_{11} = 1$, $\sigma_{12} = 1$, $\sigma_{21} = 1$ e $\sigma_{22} = 1$.

No caso bivariado, a equação 2.3 pode ser escrita como

$$f(x,y) = \gamma f_1(x,y) + (1-\gamma) f_2(x,y), \tag{2.4}$$

em que f_i é definida como na equação 2.1, com médias μ_{i1} e μ_{i2}, desvios padrão σ_{i1} e σ_{i2}, e correlação ρ_i, $i = 1{,}2$.

Na Figura 1, a equação 2.4 foi utilizada, assumindo $\gamma = 0{,}5$ e, como referência, os parâmetros de uma distribuição normal bivariada com médias $\mu_{11} = 0$ e $\mu_{12} = 0$, desvios padrão $\sigma_{11} = 1$, $\sigma_{12} = 1$, $\sigma_{21} = 1$ e $\sigma_{22} = 1$ e coeficiente de correlação $\rho_1 = 0$, enquanto valores fixos. Foram feitas comparações, variando-se apenas o coeficiente de correlação e médias de uma outra distribuição normal bivariada (ρ_2, μ_{21} e μ_{22}). Obteve-se, com a mistura dessas duas distribuições, as seguintes interpretações em torno da Figura 1:

- quando $\rho_2 = 0$, $\mu_{21} = 2$ e $\mu_{22} = 2$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, resultando em duas circunferências centradas, uma em 0 outra em 2;

- quando $\rho_2 = 0{,}5$, $\mu_{21} = 1$ e $\mu_{22} = 1$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, em que uma distribuição está centrada em 0 e outra está centrada em 1, ocorrendo uma interseção constituída, respectivamente, de uma circunferência ($\rho_1 = 0$) e uma elipse ($\rho_2 = 0{,}5$);

- quando $\rho_2 = 0{,}9$, $\mu_{21} = 1$ e $\mu_{22} = 1$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, em que uma distribuição está centrada em 0 e outra está centrada em 1, ocorrendo uma interseção constituída, respectivamente, de uma circunferência ($\rho_1 = 0$) e uma elipse bem alongada ($\rho_2 = 0{,}9$), em que se visualiza um coeficiente de correlação alto;

- quando $\rho_2 = 0$, $\mu_{21} = 0{,}5$ e $\mu_{22} = 0{,}5$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, em que uma distribuição está centrada em 0 e outra está centrada em $0{,}5$; ocorre, então, uma interseção de duas circunferências já que $\rho_1 = \rho_2 = 0$, visualizando-se uma correlação nula nestas distribuições;

- quando $\rho_2 = 0$, $\mu_{21} = 1$ e $\mu_{22} = 1$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, em que uma distribuição está centrada em 0 e outra está centrada em 1; ocorre, então, uma interseção de duas circunferências já que $\rho_1 = \rho_2 = 0$, diferenciando-se da situação anterior na medida em que o valor da média aumentou e permite visualizar, mais uma vez, uma correlação nula nestas distribuições;

- quando $\rho_2 = 0$, $\mu_{21} = 1{,}5$ e $\mu_{22} = 1{,}5$, com a mesma estrutura de covariância, é gerado um gráfico de contorno, em que uma distribuição está centrada em 0 e outra está centrada em $1{,}5$; repete-se a ocorrência de uma interseção de duas circunferências, motivada pelos coeficientes $\rho_1 = \rho_2 = 0$, visualizando-se uma correlação nula nestas distribuições.

2.1.3 Distribuição t-Student multivariada

Além das anteriores, uma distribuição de igual importância é a *t-Student* multivariada. Isto porque, no contexto do presente estudo, necessário se torna aprofundar conhecimentos no que diz respeito à sensibilidade de um teste ou de um processo de estimação por região aos desvios de normalidade dos dados. A t-Student multivariada é um exemplo da família de distribuições elípticas, tornando-se relevante para avaliar tal sensibilidade.

Para definir a função de densidade de probabilidade da distribuição t-Student

multivariada, suponha-se um vetor aleatório $\boldsymbol{X} = [X_1, \ldots, X_p]' \in \mathbb{R}^p$ com

$$f(\boldsymbol{x}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}\Gamma[(\nu+p)/2]}{[\Gamma(1/2)]^p\Gamma(\nu/2)\nu^{p/2}} \left[1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}}, \qquad (2.5)$$

diz-se que $\boldsymbol{X}$ tem distribuição de probabilidade *t* multivariada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ e com ν graus de liberdade, com a notação $\boldsymbol{X} \sim t_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ (Lange et al., 1989).

Nota-se que a distribuição *t* multivariada se aproxima da distribuição normal multivariada com matriz de covariância $\boldsymbol{\Sigma}$ quando $\nu \to \infty$ (Lange et al., 1989).

2.2 Decomposição do Valor Singular

A Decomposição em Valores Singulares (DVS) de uma matriz é a melhor forma de determinar numericamente seu posto, sendo este igual ao número de valores singulares não nulos dessa matriz.

A DVS consiste na determinação de matrizes ortogonais e, também, da matriz dos valores singulares de A de modo a satisfazer o seguinte teorema:

Teorema 2.1 (Decomposição do Valor Singular)**.** *Toda matriz $A_{n\times m}$ de posto r pode ser decomposta em*

$$A = P\Lambda Q' \qquad (2.6)$$

em que $P_{n\times r}$ e $Q_{m\times r}$ são ortonormais por coluna, ou seja, $P'P = Q'Q = I_r$ e $\Lambda = diag(\lambda_i^{\frac{1}{2}})$, $\lambda_i > 0$. As quantidades λ_1,λ_2,...,λ_r são os autovalores não-nulos das matrizes AA' ou $A'A$ e, P e Q correspondem às matrizes formadas pelos r autovetores das matrizes AA' ou $A'A$ dipostos em suas colunas, respectivamente.

É comum definir Λ como sendo

$$\Lambda = \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \tag{2.7}$$

em que σ_i são os valores singulares (e os autovalores da matriz A quando A é simétrica) não nulos da matriz A, com $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq \sigma_{p+1} = \sigma_n = 0$.

2.3 Componentes Principais

A Análise de Componentes Principais (ACP) é uma abordagem estatística que, entre outras alternativas, pode ser usada para analisar correlações entre um grande número de variáveis e explicar essas variáveis em termos de suas dimensões. O objetivo é encontrar um meio de condensar a informação contida em um número de variáveis originais num conjunto menor de variáveis estatísticas com uma perda mínima de informação (Hair Júnior et al., 2005).

Supondo-se $\boldsymbol{X}$ um vetor de p variáveis aleatórias e que a estrutura de covariância entre essas p variáveis seja de interesse; neste caso, segundo Jolliffe (2002), o primeiro passo é procurar uma função linear $\boldsymbol{\alpha_1'X}$ de elementos de $\boldsymbol{X}$ com variância máxima, em que $\boldsymbol{\alpha_1}$ é um vetor de p constantes $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$ tal que

$$\boldsymbol{\alpha_1'X} = \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1p}X_p = \sum_{j=1}^{p} \alpha_{1j}X_j. \tag{2.8}$$

Em seguida, procurar uma função linear $\boldsymbol{\alpha_2'X}$, não correlacionada com $\boldsymbol{\alpha_1'X}$, com variância máxima, e assim por diante, até que seja encontrada uma função linear $\boldsymbol{\alpha_k'X}$ que tenha variância máxima e não seja correlacionada com $\boldsymbol{\alpha_1'X}$, $\boldsymbol{\alpha_2'X}$,..., $\boldsymbol{\alpha_{k-1}'X}$. A função $\boldsymbol{\alpha_k'X}$, também chamada de variável latente, é a k-ésima Componente Principal (CP). Poderiam ser encontradas p CPs, mas espera-

se, em geral, que a maior parte da variação em $\boldsymbol{X}$ seja representada por m CPs, em que $m \ll p$.

Tendo definido Componentes Principais, deve-se saber como encontrá-los. Considera-se o caso em que o vetor de variáveis aleatórias $\boldsymbol{X}$ tenha uma matriz de covariâncias conhecida $\boldsymbol{\Sigma}$. Esta é a matriz em que os (i,j)-ésimos elementos representam a covariância entre o i-ésimo e o j-ésimo elementos de $\boldsymbol{X}$ quando $i \neq j$ e, a variância do j-ésimo elemento de $\boldsymbol{X}$ quando $i = j$. Quando $\boldsymbol{\Sigma}$ for desconhecida, esta será representada pela matriz de covariâncias amostral $\boldsymbol{S}$. Para $k = 1,2,...,p$, o k-ésimo CP será dado por $z_k = \boldsymbol{\alpha_k'X}$, em que $\boldsymbol{\alpha_k}$ será um autovetor de $\boldsymbol{\Sigma}$ correspondente ao k-ésimo autovalor λ_k. Mais ainda, se $\boldsymbol{\alpha_k}$ for escolhido de forma que $\boldsymbol{\alpha_k'\alpha_k} = 1$, então $var(z_k) = \lambda_k$, em que $var(z_k)$ é chamado de variância de z_k (Jolliffe, 2002).

Para se obter as CPs, considera-se, primeiro, $\boldsymbol{\alpha_1'X}$. O vetor $\boldsymbol{\alpha_1}$ maximiza $var[\boldsymbol{\alpha_1'X}] = \boldsymbol{\alpha_1'\Sigma\alpha_1}$. Tal como está, o máximo não será atingido para $\boldsymbol{\alpha_1}$ finito, então uma restrição de normalização deve ser imposta. A restrição utilizada será $\boldsymbol{\alpha_1'\alpha_1} = 1$, isto é, a soma de quadrados dos elementos de $\boldsymbol{\alpha_1}$ é igual a 1 (Jolliffe, 2002). Segundo este autor, o uso de outras restrições, se não $\boldsymbol{\alpha_1'\alpha_1} = constante$, leva a um problema de otimização mais complexo, que é produzir um conjunto de variáveis diferente das CPs.

Para maximizar $\boldsymbol{\alpha_1'\Sigma\alpha_1}$ sujeita a $\boldsymbol{\alpha_1'\alpha_1} = 1$, um procedimento padrão consiste no uso da técnica dos multiplicadores de Lagrange. Ou seja, maximizar

$$\boldsymbol{\alpha_1'\Sigma\alpha_1} - \lambda(\boldsymbol{\alpha_1'\alpha_1} - 1) \tag{2.9}$$

em que λ é um multiplicador de Lagrange. Derivando matricialmente a equação

2.9 em relação a $\boldsymbol{\alpha_1}$ e igualando este resultado a zero, tem-se

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\alpha_1}}[\boldsymbol{\alpha_1'}\boldsymbol{\Sigma}\boldsymbol{\alpha_1} - \lambda(\boldsymbol{\alpha_1'}\boldsymbol{\alpha_1} - 1)] &= 0 \Longrightarrow \\ \Longrightarrow 2\boldsymbol{\Sigma}\boldsymbol{\alpha_1} - 2\lambda\boldsymbol{\alpha_1} &= 0 \end{aligned} \tag{2.10}$$

resultando em

$$\boldsymbol{\Sigma}\boldsymbol{\alpha_1} - \lambda\boldsymbol{\alpha_1} = \mathbf{0} \quad \text{ou} \quad (\boldsymbol{\Sigma} - \lambda\boldsymbol{I}_p)\boldsymbol{\alpha_1} = \mathbf{0}, \tag{2.11}$$

em que $\boldsymbol{I}_p$ é a matriz identidade de ordem p e $\mathbf{0}$ é o vetor nulo. Dessa forma, λ será um autovalor de $\boldsymbol{\Sigma}$ e $\boldsymbol{\alpha_1}$ será o autovetor correspondente. Para decidir qual dos p autovetores resulta em $\boldsymbol{\alpha_1'}\boldsymbol{X}$ com variância máxima, note que a quantidade a ser maximizada será

$$\boldsymbol{\alpha_1'}\boldsymbol{\Sigma}\boldsymbol{\alpha_1} = \boldsymbol{\alpha_1'}\lambda\boldsymbol{\alpha_1} = \lambda\boldsymbol{\alpha_1'}\boldsymbol{\alpha_1} = \lambda. \tag{2.12}$$

Assim, o valor de λ será tão grande quanto possível. Consequentemente, $\boldsymbol{\alpha_1}$ será o autovetor correspondente ao maior autovalor de $\boldsymbol{\Sigma}$ e $Var(\boldsymbol{\alpha_1'}\boldsymbol{X}) = \boldsymbol{\alpha_1'}\boldsymbol{\Sigma}\boldsymbol{\alpha_1} = \lambda_1$ o maior autovalor (Jolliffe, 2002). Os demais autovalores são determinados de forma análoga, porém, não será discutido no presente trabalho.

Cada autovalor λ_k representa a variância de uma componente z_k, $k = 1,...,p$. Como os autovalores estão ordenados de forma decrescente, a primeira componente será a de maior variabilidade e a p-ésima será a de menor.

Dessa forma, segundo Jolliffe (2002), se as q primeiras componentes, $q < p$ explicam uma grande parte da variância total de $\boldsymbol{X}$, pode-se restringir o foco de atenção apenas para o vetor $\boldsymbol{z_k} = [z_1,z_2,...,z_q]'$, $k = 1,2,...,q$ sem perder muita informação sobre a estrutura de covariâncias original de $\boldsymbol{X}$. As evidências

empíricas e científicas conduzem à utilização do critério de reter um número $q < p$ de CPs que contemplem pelo menos 70% da variação total.

2.3.1 Interpretação geométrica das componentes principais

A análise de componentes principais é um tipo de análise multivariada, mas aqui será tratada como uma técnica de desenvolvimento do Escalonamento Multidimensional (*Multidimensional Scaling* - MDS), que se caracteriza por estudar a distribuição espacial dos objetos de forma a reconhecer agrupamentos e relações entre eles. Para melhor compreendê-la, verifique a localização de alguns pontos no plano XY, conforme figura 2 (Pereira, 2004).

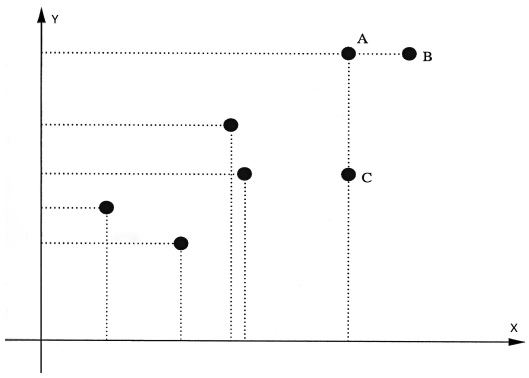

FIGURA 2 Pontos dispersos no plano XY.

Observa-se que nenhum dos eixos discrimina, perfeitamente, os objetos; **A** confunde-se com **B** no eixo dos **Y** e com **C** no eixo dos **X**. Podendo escolher outro sistema de coordenadas em vez de **X** e **Y**, os pontos seriam melhor discriminados, como na figura 3.

Nota-se agora, na figura 3, que as projeções de todos os pontos correspondem a coordenadas distintas. Todos os pontos podem ser distinguidos com o novo sistema de coordenadas, em que o maior eixo representa a melhor distinção e o menor,

perpendicular, complementa tal distinção a fim de detectar pequenas diferenças que o eixo maior não conseguiu sozinho (Pereira, 2004).

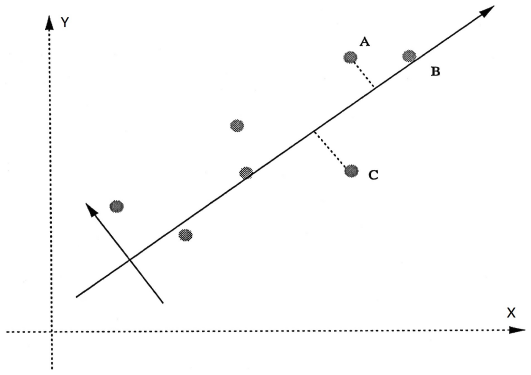

FIGURA 3 Pontos do espaço XY visto por novas coordenadas.

Ainda, uma outra forma de enxergar os novos eixos é fazer uma rotação no sistema de eixos originais que, no caso da figura 2, seria uma rotação no sentido anti-horário de, aproximadamente, $45°$, como se dá na figura 4.

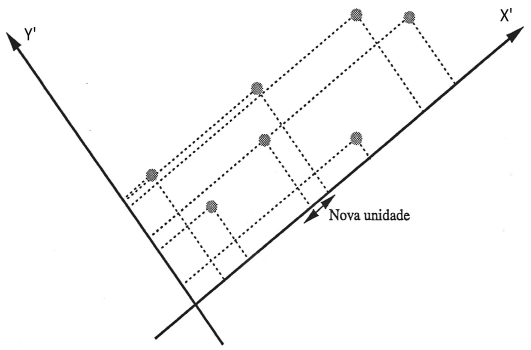

FIGURA 4 Pontos no plano XY com rotação anti-horária dos eixos em $45°$.

Segundo Pereira (2004), a análise de componentes principais identifica, primeiro, a função que melhor discrimina os objetos e, em seguida, deriva a função ortogonal dessa primeira função. Neste novo sistema de coordenadas as unida-

des de medidas serão abstratas, ou seja, não serão dadas pelos eixos **X** ou **Y**. Tais medidas poderiam ser chamadas de escores e a unidade poderia ser a menor medida entre dois pontos, como na figura 4. Pode-se perceber que o maior eixo do novo sistema de coordenadas corresponde à reta de regressão do sistema de eixos original, porém, a reta de regressão é expressa, geralmente, tomando um eixo em função do outro, $y = f(x)$, enquanto a componentes principal uma nova dimensão que será função dos eixos originais, $z = f(x,y)$.

No caso citado nesta seção, a ACP foi trabalhada no espaço bidimensional. O mesmo será feito num espaço multimensional: primeiro identifica-se a dimensão que melhor distingue os objetos, em seguida, identifica-se a função ortogonal a ela; caso os objetos não sejam totalmente discriminados, um novo componente discriminador será buscado, juntamente, com sua função ortogonal derivada e, assim por diante. Dessa forma, o interesse está em saber quanto essas funções, que são uma redução das dimensões originais, estarão, ainda, representando a variabilidade do universo estudado (Pereira, 2004).

2.3.2 Detecção de *outliers* e observações influentes

Um outlier multivariado é aquela observação que apresenta um grande distanciamento das restantes, no espaço p-dimensional definido por todas as variáveis. No entanto, um outlier multivariado não necessita ter valores anormais em qualquer uma das variáveis (Jolliffe, 2002). Muitas das primeiras propostas para a identificação de *outliers* multivariados referem-se a métodos baseados na análise gráfica. Uma das contribuições mais importantes deve-se a Gnanadesikan & Kettenring (1972).

Jolliffe (2002) ainda chama a atenção para um problema importante na detecção de *outliers* multivariados: uma observação que não é outlier em qualquer uma

das variáveis originais, no caso univariado, ainda pode ser um outlier na análise multivariada, por não se conformar com a estrutura de correlação do restante dos dados.

Para grandes valores de p é preciso considerar a possibilidade de que *outliers* irão se manifestar em direções diferentes das que são detectáveis na plotagem simples de pares de variáveis originais. Além do mais, Jolliffe (2002) afirma que, em geral, a identificação de *outliers*, via o método de componentes principais, pode ser contraditória, dependendo de quais componentes são considerados, uma vez que, se forem assumidos os últimos componentes, estes são mais suceptíveis de fornecer informações adicionais que não estão disponíveis na plotagem das variáveis originais. Ainda, segundo este autor, isso acontece porque uma forte estrutura de correlação entre as variáveis originais implica que existam funções lineares dessas variáveis (as componentes principais), com variâncias pequenas, se comparadas com as variâncias das variáveis originais. Examinando os valores das últimas componentes, detectam-se observações que violam a estrutura de correlação imposta pelo conjunto de todos os dados, não sendo necessariamente aberrantes se forem consideradas as variáveis originais. Ainda mais, se um outlier for a causa de um grande aumento de uma ou mais variâncias das variáveis originais, então ele deve ser um extremo dessas variáveis e, portanto, detectável pelo olhar na plotagem das mesmas.

Em se tratando da identificação de *outliers* por meio da técnica de componentes principais, inúmeros testes são propostos na literatura, fundamentados na combinação de informações de vários CPs em vez de examinar individualmente cada um deles. A título de ilustração, Gnanadesikan & Kettenring (1972) apresentam alguns testes que, doravante, serão descritos neste referencial, partindo da definição de quatro estatísticas d_{1i}^2, d_{2i}^2, d_{3i}^2 e d_{4i}^2.

Conforme mencionado por Jolliffe (2002), os últimos CPs tendem a ser mais úteis do que os primeiros na detecção de *outliers* que não são visíveis a partir das variáveis originais. Dessa forma, uma possível estatística de teste, d_{1i}^2, discutida por Gnanadesikan & Kettenring (1972) é a soma dos quadrados dos valores das últimas CPs e é dada por

$$d_{1i}^2 = \sum_{k=p-q-1}^{p} z_{ik}^2, \tag{2.13}$$

em que z_{ik} corresponde ao valor da i-ésima observação na k-ésima CP, $k = 1,2,...,q,q+1,...,p$. A estatística d_{1i}^2, $i = 1,\ldots,n$ terá, aproximadamente, distribuição gama com observações independentes se não houver *outliers*, para que um gráfico de probabilidade gama com parâmetro estimado de forma adequada possa expor *outliers*. Valores exageradamente elevados de d_{1i}^2 indicam que a observação i seja, possivelmente, um outlier ou que tal observação tenha um fraco ajustamento ao espaço de $(p-q)$ dimensões (Gnanadesikan & Kettenring, 1972).

Uma possível crítica à estatística d_{1i}^2 é que ela, ainda, dá peso insuficiente para os últimos CPs, especialmente se q estiver próximo de p. Como os CPs têm variâncias decrescentes, os valores de z_{ik}^2, normalmente, tornam-se menores à medida que k aumenta, e d_{1i}^2, portanto, implicitamente, diminui-se o peso dos CPs à medida que k aumenta. Este efeito pode ser grave se alguns dos CPs resultarem em variâncias muito pequenas. A ocorrência deste fato não é satisfatória, pois são justamente o CPs de baixa variância que podem ser mais eficazes para determinar a presença de certos tipos de *outliers* (Jolliffe, 2002).

Uma alternativa é dar pesos iguais às componentes, padronizando-as. E isso pode ser alcançado através da substituição de z_{ik} por

$$z_{ik}^* = \frac{z_{ik}}{\lambda_k^{\frac{1}{2}}} \tag{2.14}$$

em que λ_k é a variância da k-ésima CP amostral. Neste caso, as variâncias amostrais de z_{ik} serão todas iguais a 1 (um). É importante notar que, quando $q = p$, a estatística d_{1i}^2 transforma-se em

$$d_{2i}^2 = \sum_{k=p-q-1}^{p} \frac{z_{ik}^2}{\lambda_k} \tag{2.15}$$

que é simplesmente o quadrado da Distância de Mahalanobis, D_i^2, entre a i-ésima observação e a média amostral, definida como

$$D_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}). \tag{2.16}$$

Hawkins (1974) preferiu utilizar d_{2i}^2 com $q < p$ em vez de $q = p$ de modo a dar maior importância às componentes de menor variância. No caso de serem consideradas todas as componentes do cálculo de d_{2i}^2, esta estatística terá, aproximadamente, distribuição qui-quadrado com p graus de liberdade (χ_p^2).

Gnanadesikan & Kettenring (1972) consideram, também, a estatística

$$d_{3i}^2 = \sum_{k=1}^{p} \lambda_k z_{ik}^2, \tag{2.17}$$

que enfatiza as observações que têm um grande efeito sobre os primeiros CPs e que é equivalente a $(x_i - \bar{x})' S (x_i - \bar{x})$. Como afirmado anteriormente, os primeiros CPs são úteis na detecção de alguns tipos de outlier e d_{3i}^2 os destaca. Estes autores ainda reafirmam que *outliers* são muitas vezes detectáveis na plotagem das variáveis originais, ao contrário dos *outliers* expostos pelos últimos CPs.

Hawkins (1974), também, mostrou que *outliers* poderão ser detectados de

forma mais expressiva utilizando a estatística

$$d_{4i} = max_{p-q+1 \leq k \leq p} |z^*_{ik}|. \tag{2.18}$$

A estatística de teste para a i-ésima observação é, então, definida pelo valor absoluto máximo dos últimos CPs, renormalizados e rotacionados, para tal observação.

As distribuições exatas para 2.13, 2.15, 2.17 e 2.18 podem ser deduzidas, se for assumido que as observações são de uma distribuição normal multivariada com média μ e matriz de covariância Σ, em que μ e Σ são conhecidos para os resultados de d^2_{2i} e d^2_{4i} (Jolliffe, 2002). Convém ressaltar que Gnanadesikan & Kettenring (1972) afirmam que d^2_{3i} e d^2_{2i}, quando $q = p$, bem como d^2_{1i}, são, aproximadamente, distribuidos por uma gama. Se *outliers* não estiverem presentes e, se a normalidade puder ser, aproximadamente, assumida, o gráfico de probabilidade gama de d^2_{2i} (com $q = p$) e d^2_{3i} poderá ser utilizado, novamente, para detectar *outliers*.

3 METODOLOGIA

A metodologia utilizada para a execução da pesquisa fundamentou-se nos conteúdos e abordagens explicitados nos seguintes tópicos desta seção: 3.1 Correções do tipo Qui-quadrado de Pearson e de Yates nos dados originais; 3.2 Cálculo do coeficiente de curtose para os dados corrigidos pelas distâncias qui-quadrado; 3.3 Construção de um teste de significância para o coeficiente de curtose e 3.4 Misturas de distribuições.

3.1 Correções do tipo Qui-quadrado de Pearson e de Yates nos dados originais

Ressalta-se, aqui, a opção por adaptar as distâncias do tipo qui-quadrado de Pearson e de Yates, pelo fato de que, na literatura corrente, considera-se que as mesmas são altamente influenciadas por *outliers*.

Com base numa amostra aleatória mutivariada, será utilizada a notação de vetor aleatório para cada unidade amostral, em que, cada vetor terá p componentes. Dois índices serão utilizados para os elementos X_{ij} do vetor: i se refere a i-ésima unidade amostral e j se refere a j-ésima variável aleatória ($i = 1,...,n$; $j = 1,...,p$). Tais vetores serão representados em negrito. Tal amostra será representada por $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_i,\ldots,\boldsymbol{X}_n$ e o vetor p-dimensional da i-ésima unidade amostral será representado por $\boldsymbol{X}_i = [\boldsymbol{X}_{i1},\boldsymbol{X}_{i2},\ldots,\boldsymbol{X}_{ij},\ldots,\boldsymbol{X}_{ip}]$, assim como o vetor n-dimensional da j-ésima variável por $\boldsymbol{X}_j = [\boldsymbol{X}_{1j},\boldsymbol{X}_{2j},\ldots,\boldsymbol{X}_{ij},\ldots,\boldsymbol{X}_{nj}]$.

Preliminarmente à identificação de *outliers*, as observações foram corrigidas pelas equações 3.1 e 3.2, distâncias do tipo Qui-quadrado de Pearson e de Yates, respectivamente, conforme expressões abaixo.

$$q_1 = \frac{x_{ij} - \sum_{i=1}^{n} x_{ij} \sum_{j=1}^{p} x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij} \sum_{j=1}^{p} x_{ij}}} \tag{3.1}$$

$$q_2 = \frac{(x_{ij} - |\sum_{i=1}^{n} x_{ij} \sum_{j=1}^{p} x_{ij}| - 1)^2}{\sum_{i=1}^{n} x_{ij} + \sum_{j=1}^{p} x_{ij}}. \tag{3.2}$$

Para cada uma das matrizes, com elementos q_1 e q_2, foi utilizada uma reescalonagem dos dados por meio da mediana e do Desvio Absoluto da Mediana (*Median Absolute Deviation* - MAD). Tal procedimento implicou no cálculo dos coeficientes de curtose, ilustrados na seção 3.2.

3.2 Cálculo do coeficiente de curtose para os dados corrigidos pelas distâncias qui-quadrado

O método proposto por Filzmoser et al. (2008) constitui referência e motivação para definição do objeto de estudo da pesquisa desenvolvida. Seu método consiste na padronização dos dados pela mediana e pelo MAD para se chegar ao cálculo do coeficiente de curtose robusta.

Considerando as matrizes dos elementos q_1 e q_2 obtidas na seção 3.1, a obtenção dos coeficientes de curtose se deu com a execução dos seguintes procedimentos.

Assumindo Q_s, $s = 1,2$, sendo esta, a matriz p-variada dos dados corrigidos, inicialmente foi realizada a tranformação nas observações conforme expressão 3.3:

$$q^*_{s_{ij}} = \frac{q_{s_{ij}} - med(q_{s_{1j}},...,q_{s_{nj}})}{MAD(q_{s_{1j}},...,q_{s_{nj}})};\ j = 1,...,p, \tag{3.3}$$

em que o MAD, já mencionado com o desvio absoluto da mediana, foi obtido por

$$MAD(q_{s_1},...,q_{s_n}) = 1{,}4826 med_j \left|q_{s_j} - med_i(q_{s_i})\right|, \tag{3.4}$$

sendo $1{,}4826$ o valor correspondente ao quantil 75% de uma distribuição normal. Segundo recomendações de Rousseeuw (1984), assumindo este quantil, a estimativa da mediana torna-se consistente.

Desenvolvido este procedimento, baseando-se na observação $q^*_{s_{ij}}$, calculou-se a matriz de covariância, na qual os autovalores e os autovetores foram obtidos pela decomposição dos valores singulares. Representando a matriz dos autovetores por $V_{s_{(p)}}$ foi possível obter a matriz das componentes, conforme a expressão $Z_s = Q^*_s \cdot V_s$, em que Q^*_s é a matriz dos elementos $q^*_{s_{ij}}$. Novamente foram reescaladas as componentes de Z_s pela mediana e pelo MAD de forma similar à equação 3.3, resultando em uma nova matriz Z^*_s:

$$z^*_{s_{ij}} = \frac{z_{s_{ij}} - med(z_{s_{1j}},...,z_{s_{nj}})}{MAD(z_{s_{1j}},...,z_{s_{nj}})};\ j = 1,...,p. \tag{3.5}$$

Após a realização deste procedimento, considerou-se Z^*_s para o cálculo do valor absoluto da curtose robusta (ω_{s_j}) para cada variável, conforme expressão 3.6.

$$\omega_{s_j} = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{[z^*_{s_{ij}} - med(z^*_{s_{1j}},...,z^*_{s_{nj}})]^4}{[MAD(z^*_{s_{1j}},...,z^*_{s_{nj}})]^4} - 3 \right|;\ j = 1,...,p. \tag{3.6}$$

Por uma questão de interpretação, os coeficientes ω_{s_j} foram padronizados pela equação 3.7.

$$\xi_{s_j} = \frac{\omega_{s_j}}{\sum_j \omega_{s_j}} \tag{3.7}$$

Segundo Peña & Prieto (2001), a plotagem das componentes com pequenos e grandes valores do valor absoluto de curtose robusta são indicadores da presença de *outliers*.

3.3 Construção de um teste de significância para o coeficiente de curtose

Os elementos referenciais teóricos, até aqui explicitados, estão na base da construção de um teste de significânica para coeficiente de curtose robusta ora proposto. Acredita-se que tal proposição acrescenta ao tema em estudo uma dimensão inovadora, uma vez que não se encontrou durante a revisão de literatura, a construção de um teste desta natureza.

A construção do teste de significância, assumindo a hipótese nula descrita por $\boldsymbol{H_0}$ **:** O primeiro coeficiente de curtose padronizado é nulo, partiu das seguintes premissas:

(a) Pequenos e grandes valores de ξ_{s_j}, associados aos componentes reescalonados obtidos em Z_s^*, são indicadores de que os componentes a serem utilizados na plotagem dos escores identificam os *outliers* de forma mais eficiente (Peña & Prieto, 2001);

(b) Segundo mencionam Filzmoser et al. (2008), a relação da quantidade de observações *outliers* está associada à magnitude de ξ_{s_j}, de tal forma que valores do coeficiente de curtose indicam maiores quantidades de *outliers*;

(c) O fato de se comprovar, estatisticamente, que o primeiro coeficiente de curtose padronizado é nulo, implica, em termos práticos, uma situação mais elementar na identificação de *outliers*, ou seja, a identificação de *outliers* numa única dimensão.

Tendo por base essas afirmações, o teste de significância foi formulado, conforme a especificação da hipótese nula dada por $\boldsymbol{H_0}$: O primeiro coeficiente de curtose padronizado é nulo, e que, a estatística do teste é dada por

$$T_s = max[\xi_{s_j}];\ j = 1,...,p \tag{3.8}$$

e a regra de decisão é dada por

$$T_s > R_s = \frac{\lambda_{s_1}}{\sum_{j=1}^{p} \lambda_{s_j}};\ j = 1,...,p;\ s = 1,2. \tag{3.9}$$

Mantendo essas especificações, a probabilidade de significância, interpretada como a menor probabilidade que leva à rejeição da hipótese nula, foi computada por

$$\text{valor-p} = \sum_{m=1}^{n} \frac{I(T_s \geq R_s)}{m}, \tag{3.10}$$

em que I é uma função indicadora e m é o número de simulações Monte Carlo, previamente fixado em duas mil simulações. Rejeita-se H_0 caso o valor-p for inferior ao nível de significância especificado pelo pesquisador.

3.4 Misturas de distribuições

Sentiu-se necessidade de criar condições para identificar *outliers* na amostra e, por consequência, para testar o método que, a partir da pesquisa realizada, está sendo introduzido nos estudos da área.

A abordagem metodológica foi concebida em termos computacionais para que as correções por distâncias qui-quadrado nas amostras fossem implementadas, bem como o teste de significância fosse avaliado; os valores paramétricos assumidos na simulação foram definidos nos vetores de médias $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ de dimensões $(p \times 1)$ da seguinte forma:

$$\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\mu_2} = \begin{bmatrix} 100 \\ 100 \\ \vdots \\ 100 \end{bmatrix}. \tag{3.11}$$

Considerou-se, também, as matrizes de covariâncias $\boldsymbol{\Sigma_1}$ e $\boldsymbol{\Sigma_2}$, de ordem p, definidas da seguinte forma:

$$\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & \rho^{1-2} & \cdots & \rho^{1-p} \\ \rho^{2-1} & 1 & \cdots & \rho^{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{3.12}$$

em que $\rho = 0{,}5$ é o coeficiente de correlação assumido.

Na simulação feita recorreu-se a diferentes tamanhos de amostras (n igual a 20, 50, 100 e 150), diferentes quantidades de variáveis (p igual a 5, 10, 20 e 50) e diferentes taxas de misturas (γ igual a 0,1, 0,2 e 0,3).

Alternando os valores paramétricos entre $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ e entre $\boldsymbol{\Sigma_1}$ e $\boldsymbol{\Sigma_2}$, diferentes misturas de distribuições normais e *t-Student* foram geradas, as quais seguem descritas.

(i) Se $u \leq \gamma$, então os dados assumirão valores de uma distribuição normal p-variada com a $A \sim N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$; se $u > \gamma$, os dados assumirão valores de uma normal p-variada com a configuração $A \sim N_p(\boldsymbol{\mu_2},\boldsymbol{\Sigma_2})$.

(ii) Se $u \leq \gamma$, então os dados assumirão valores de uma distribuição normal p-variada com a $B \sim N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$; se $u > \gamma$, os dados assumirão valores de uma normal p-variada com a configuração $B \sim N_p(\boldsymbol{\mu_2},\boldsymbol{\Sigma_1})$.

(iii) Se $u \leq \gamma$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $C \sim N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$; se $u > \gamma$, os dados assumirão valores de uma normal p-variada com a configuração $C \sim N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_2})$.

(iv) Se $u \leq \gamma$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $D \sim N_p(\boldsymbol{\mu_2},\boldsymbol{\Sigma_2})$; se $u > \gamma$, os dados assumirão

valores de uma distribuição t-Student p-variada com a configuração $D \sim t_p(\mathbf{\Sigma_2}, \nu = 5)$, em que ν são os graus de liberdade.

Todos os cálculos e simulações foram feitos no software R (R Development Core Team, 2009) por meio do desenvolvimento de um programa que se encontra descrito no Anexo.

Convém salientar que, em termos práticos, o uso dos métodos, com ou sem correção qui-quadrado, evidenciados pelo teste de significância idealizado neste trabalho, deparou-se com o problema de identificabilidade, ou seja, no sentido de estimar a proporção de amostras contaminadas. Assim sendo, por se tratar de um estudo empírico, norteado pelos objetivos enunciados no início do presente relatório de pesquisa, a estimação de γ não será abordada. Entretanto, para maiores detalhes desta inferência em γ indica-se a referência Chen & Tan (2009).

4 RESULTADOS E DISCUSSÃO

4.1 Probabilidades empíricas a favor de H_0 para o teste de significância proposto para o coeficiente de curtose

Mantendo-se as mesmas configurações especificadas entre os tamanhos amostrais (n), número de variáveis (p) e porcentagem de mistura (γ), a qual interpretou-se como quantidade média de observações *outliers* presentes na amostra, obtida da população de referência, procedeu-se à obtenção das probabilidades de significância, conforme teste descrito na subseção 3.3. Contudo, convém ressaltar que tal probabilidade corresponde ao menor valor que ocasiona a rejeição de H_0, isto é, o valor-p, tratado como uma medida de evidência, independente do nível de significância especificado pelo pesquisador.

Um outro fato a ser registrado é que, na literatura pesquisada não foi encontrado teste de significância para o coeficiente de curtose, como proposto neste trabalho, e sim, apenas os resultados provenientes de simulação, do desempenho de diferentes métodos comparados por Filzmoser et al. (2008) na identificação de *outliers* em altas dimensões, apresentando a taxa de Falso Positivo $(\%FP)$ que corresponde à probabilidade de hipóteses nulas, reconhecidas como verdadeiras, rejeitadas erroneamente dentro de um conjunto de hipóteses especificadas.

Nas tabelas, que se seguem, são descritos os valores-p do teste de significância gerados por simulação, para a amostra original (Amostra original) e para a amostra corrigida pelas distâncias do tipo qui-quadrado de Pearson (Pearson) e de Yates (Yates), obtidos com diferentes configurações em que variam-se os tamanhos amostrais, a quantidade de variáveis e a porcentagem de mistura. Tais amostras foram obtidas a partir de misturas de distribuições multivariadas conforme explicitadas na seção 3.4, itens (i), (ii), (iii) e (iv).

TABELA 1 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias e estruturas de covariâncias diferentes, com amostras ($n = 20{,}50$), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
20	5	0	0,7065	0,2810	0,1235
		0,1	0,5225	0,6790	0,1235
		0,2	0,4355	0,7585	0,0860
		0,3	0,2340	0,7740	0,0750
	10	0	0,6785	0,0180	0,0030
		0,1	0,1935	0,4390	0,0015
		0,2	0,0755	0,6260	0,0010
		0,3	0,0225	0,7000	0,0005
	20	0	0,9505	0,0005	0,0000
		0,1	0,1880	0,1315	0,0000
		0,2	0,0715	0,3080	0,0000
		0,3	0,0695	0,2905	0,0000
	50	0	0,4840	0,0000	0,0000
		0,1	0,0485	0,0100	0,0000
		0,2	0,0085	0,0220	0,0000
		0,3	0,0015	0,0365	0,0000
50	5	0	0,5355	0,3185	0,1665
		0,1	0,8860	0,7240	0,1355
		0,2	0,5100	0,7580	0,1030
		0,3	0,2455	0,8245	0,0775
	10	0	0,3650	0,0095	0,0010
		0,1	0,4095	0,6405	0,0100
		0,2	0,1550	0,7310	0,0040
		0,3	0,0300	0,8275	0,0010
	20	0	0,3515	0,0000	0,0000
		0,1	0,0105	0,1760	0,0005
		0,2	0,0025	0,3310	0,0000
		0,3	0,0000	0,4565	0,0000
	50	0	0,9765	0,0000	0,0000
		0,1	0,0405	0,0000	0,0000
		0,2	0,0145	0,0265	0,0000
		0,3	0,0080	0,0445	0,0000

TABELA 2 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias e estruturas de covariâncias diferentes, com amostras ($n = 100$,150), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
100	5	0	0,2995	0,2510	0,1150
		0,1	0,9270	0,8255	0,1515
		0,2	0,6735	0,8910	0,1030
		0,3	0,2705	0,9005	0,0925
	10	0	0,1880	0,0175	0,0060
		0,1	0,5640	0,7235	0,0055
		0,2	0,0620	0,7970	0,0040
		0,3	0,0090	0,8210	0,0025
	20	0	0,0965	0,0000	0,0005
		0,1	0,0015	0,2660	0,0000
		0,2	0,0000	0,4265	0,0000
		0,3	0,0000	0,5910	0,0000
	50	0	0,0880	0,0000	0,0000
		0,1	0,0000	0,0065	0,0000
		0,2	0,0000	0,0030	0,0000
		0,3	0,0000	0,1330	0,0000
150	5	0	0,2360	0,3855	0,1775
		0,1	0,9940	0,7205	0,0875
		0,2	0,6485	0,8685	0,0865
		0,3	0,1715	0,9045	0,1145
	10	0	0,0510	0,0225	0,0040
		0,1	0,5970	0,7035	0,0025
		0,2	0,2680	0,7015	0,0060
		0,3	0,0040	0,8795	0,0010
	20	0	0,0115	0,0005	0,0000
		0,1	0,0005	0,4115	0,0015
		0,2	0,0000	0,4505	0,0000
		0,3	0,0000	0,5245	0,0005
	50	0	0,0100	0,0000	0,0000
		0,1	0,0000	0,0000	0,0000
		0,2	0,0000	0,0185	0,0000
		0,3	0,0000	0,0360	0,0000

Os resultados encontrados nas Tabelas 1 e 2 foram obtidos com a realização da mistura (i), subseção 3.4, na qual foram considerados os valores paramétricos distintos em ambas as populações. Desta forma, constatou-se que os resultados da simulação quando $n = 20$ e $p = 5$, à medida que houve um aumento na quantidade de *outliers*, representado pelo incremento de γ, o teste, considerando as amostras originais, propiciou uma redução na probabilidade a favor de H_0. Contudo, ao assumir maior número de variáveis p e, à medida que γ aumentou, para cada valor de p, ocorreu uma redução na probabilidade a favor de H_0.

Para a configuração $n = 50$, notou-se o mesmo comportamento, com exceção da situação referida por $p = 5$ e $0 \leq \gamma \leq 0{,}1$, na qual, a probabilidade de significância, obtida pelo teste de significância na amostra original, a favor de H_0 aumentou.

Dessa forma, em geral, o aumento da probabilidade de mistura γ resultou em uma redução na probabilidade a favor de H_0, o que, de certa forma, conduziu a um resultado inesperado, pois a ocorrência de *outliers* é mais expressiva. De forma equivalente, notou-se que os resultados relativos à ausência de *outliers*, isto é, $\gamma = 0$, em geral para amostras de tamanho $n = 20$ e $n = 50$, foram não condizentes, pois esperava-se uma baixa probabilidade a favor de H_0 na ausência de *outliers*. Porém, para tal problema, em consonância com os resultados propiciados pela correção qui-quadrado de Pearson, foram promissores, de modo a evidenciar mais expressivamente que, o primeiro coeficiente de curtose padronizado é nulo. Este fato é perceptível para todos os tamanhos amostrais avaliados, incluindo até mesmo situações com o número de variáveis, excessivamente, elevado.

Considerando que a amostra da população de referência foi contaminada por unidades amostrais provenientes de uma população que diferiu apenas no parâmetro de posição, os resultados obtidos via simulação, assumindo a mistura (ii),

TABELA 3 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias diferentes e mesma estrutura de covariância, com amostras ($n = 20{,}50$), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
20	5	0	0,8530	0,3360	0,1505
		0,1	0,7740	0,6435	0,1335
		0,2	0,5730	0,6780	0,0850
		0,3	0,3450	0,7565	0,1005
	10	0	0,9465	0,0175	0,0005
		0,1	0,5610	0,3235	0,0010
		0,2	0,3565	0,4825	0,0020
		0,3	0,2315	0,6075	0,0015
	20	0	1,0000	0,0000	0,0000
		0,1	0,2490	0,1045	0,0000
		0,2	0,0800	0,1600	0,0000
		0,3	0,0505	0,1895	0,0000
	50	0	0,9855	0,0000	0,0000
		0,1	0,1260	0,0000	0,0000
		0,2	0,0225	0,0030	0,0000
		0,3	0,0010	0,0175	0,0000
50	5	0	0,8520	0,4395	0,1920
		0,1	0,9640	0,6240	0,1925
		0,2	0,8425	0,7005	0,1105
		0,3	0,3640	0,8030	0,1105
	10	0	0,9505	0,0155	0,0020
		0,1	0,9725	0,4770	0,0020
		0,2	0,8075	0,5970	0,0015
		0,3	0,3755	0,6465	0,0005
	20	0	0,9955	0,0000	0,0000
		0,1	0,9575	0,1090	0,0000
		0,2	0,7220	0,1745	0,0000
		0,3	0,2785	0,2350	0,0000
	50	0	1,0000	0,0000	0,0000
		0,1	0,0405	0,0000	0,0000
		0,2	0,0145	0,0000	0,0000
		0,3	0,0190	0,0240	0,0000

TABELA 4 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias diferentes e mesma estrutura de covariância, com amostras (n = 100,150), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
100	5	0	0,8415	0,3940	0,1750
		0,1	0,9980	0,6695	0,2050
		0,2	0,9310	0,8245	0,1730
		0,3	0,4455	0,8195	0,1360
	10	0	0,9190	0,0505	0,0035
		0,1	1,0000	0,4685	0,0020
		0,2	0,9580	0,6175	0,0030
		0,3	0,5480	0,6065	0,0015
	20	0	0,9890	0,0000	0,0000
		0,1	1,0000	0,1220	0,0000
		0,2	0,9485	0,2970	0,0000
		0,3	0,6080	0,2495	0,0000
	50	0	1,0000	0,0000	0,0000
		0,1	0,9995	0,0000	0,0000
		0,2	0,8415	0,0000	0,0000
		0,3	0,3190	0,0515	0,0000
150	5	0	0,8570	0,5230	0,2265
		0,1	1,0000	0,7105	0,2145
		0,2	0,9925	0,9055	0,2075
		0,3	0,6015	0,8020	0,1555
	10	0	0,9195	0,0615	0,0080
		0,1	1,0000	0,4565	0,0080
		0,2	1,0000	0,5995	0,0010
		0,3	0,5865	0,8170	0,0005
	20	0	0,9950	0,0000	0,0000
		0,1	1,0000	0,1250	0,0000
		0,2	0,9880	0,2380	0,0000
		0,3	0,6615	0,4070	0,0000
	50	0	1,0000	0,0000	0,0000
		0,1	1,0000	0,0000	0,0000
		0,2	1,0000	0,0190	0,0000
		0,3	0,5755	0,0275	0,0000

subseção 3.4, e ilustrados nas Tabelas 3 e 4, indicaram que, na ausência de *outliers* ($\gamma = 0$) a probabilidade a favor de H_0 foi mais relevante, considerando amostra original. Fato este, também ocorrido nos resultados descritos anteriormente na Tabela 1.

Ainda, na Tabela 3, ao analisar os resultados específicos, a probabilidade de significância a favor de H_0, confrontando os tamanhos amostrais n, verificou-se que, para $n = 20$, as probabilidades provenientes do teste submetido às amostras corrigidas pela distância qui-quadrado de Pearson mantiveram-se promissoras em relação à tomada de decisão a favor da hipótese nula para $\gamma = 0{,}2$ e $\gamma = 0{,}3$. No entanto, aumentando o tamanho amostral para $n = 50$, observou-se que o fato de utilizar a correção qui-quadrado de Pearson nas amostras repercutiu em uma baixa probabilidade a favor de H_0, podendo ocasionar uma decisão não coerente a ser tomada pelo pesquisador, em assumir que o primeiro coeficiente de curtose padronizado seja nulo, dado um nível de significância arbitrariamente escolhido.

Em situação de tamanhos amostrais superiores a $n = 50$, como pode ser visto na Tabela 4, a redução na probabilidade a favor de H_0 também ocorreu quando as amostras foram submetidas à correção qui-quadrado de Pearson. Assim, torna-se possível inferir que, quando é sabido que *outliers* são provenientes de uma população com o mesmo parâmetro de posição, recomenda-se que a identificação destas observações seja feita considerando a amostra original.

Mantendo a mesma linha de discussão, em avaliar a probabilidade de significância do teste do coeficiente de curtose associado à identificação dos componentes que melhor discriminam *outliers*, os resultados encontrados nas Tabelas 5 e 6 foram obtidos considerando a mistura (iii), descrita na subseção 3.4, na qual se partiu do pressuposto que as unidades amostrais foram originadas de uma população que diferiu-se apenas no parâmetro de dispersão em relação à população

TABELA 5 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias iguais e estruturas de covariâncias diferentes, com amostras ($n = 20{,}50$), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
20	5	0	0,6855	0,2885	0,7090
		0,1	0,7060	0,3265	0,6970
		0,2	0,7600	0,3035	0,7135
		0,3	0,7450	0,3125	0,7200
	10	0	0,7180	0,0675	0,6990
		0,1	0,7310	0,0755	0,6755
		0,2	0,7705	0,0560	0,7035
		0,3	0,7845	0,0840	0,7165
	20	0	0,9485	0,0295	0,6565
		0,1	0,9590	0,0365	0,6620
		0,2	0,9840	0,0305	0,6855
		0,3	0,9900	0,0325	0,7030
	50	0	0,5060	0,0130	0,5270
		0,1	0,5340	0,0075	0,5875
		0,2	0,6065	0,0160	0,5590
		0,3	0,6855	0,0180	0,6235
50	5	0	0,5345	0,2605	0,7650
		0,1	0,5815	0,2165	0,7500
		0,2	0,6015	0,2530	0,7895
		0,3	0,6160	0,2460	0,7765
	10	0	0,4455	0,0695	0,7630
		0,1	0,4380	0,0630	0,7660
		0,2	0,4965	0,0630	0,7880
		0,3	0,5840	0,0590	0,7580
	20	0	0,3630	0,0180	0,7620
		0,1	0,4295	0,0260	0,7705
		0,2	0,4670	0,0240	0,7870
		0,3	0,6030	0,0285	0,7820
	50	0	0,9625	0,0100	0,7250
		0,1	0,9905	0,0100	0,7470
		0,2	0,9990	0,0085	0,7530
		0,3	1,0000	0,0130	0,7620

TABELA 6 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de duas distribuições normais multivariadas com vetores de médias iguais e estruturas de covariâncias diferentes, com amostras (n = 100,150), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
100	5	0	0,3265	0,2185	0,7805
		0,1	0,4255	0,1780	0,7860
		0,2	0,4440	0,2030	0,7960
		0,3	0,5090	0,1870	0,7800
	10	0	0,1620	0,0385	0,7835
		0,1	0,1715	0,0380	0,8015
		0,2	0,2360	0,0475	0,7920
		0,3	0,3295	0,0430	0,7855
	20	0	0,0885	0,0135	0,7725
		0,1	0,1365	0,0190	0,7830
		0,2	0,1450	0,0205	0,7855
		0,3	0,1800	0,0190	0,7900
	50	0	0,0915	0,0105	0,7840
		0,1	0,1660	0,0090	0,7915
		0,2	0,2000	0,0140	0,7940
		0,3	0,3165	0,0120	0,7865
150	5	0	0,1180	0,1810	0,7775
		0,1	0,2680	0,2210	0,7940
		0,2	0,3520	0,1840	0,8130
		0,3	0,3035	0,2050	0,7885
	10	0	0,0850	0,0385	0,8090
		0,1	0,1295	0,0325	0,8110
		0,2	0,1115	0,0325	0,7890
		0,3	0,1565	0,0410	0,7825
	20	0	0,0280	0,0195	0,8020
		0,1	0,0400	0,0120	0,8155
		0,2	0,0500	0,0195	0,8090
		0,3	0,1110	0,0155	0,8060
	50	0	0,0110	0,0065	0,7740
		0,1	0,0205	0,0075	0,7900
		0,2	0,0320	0,0060	0,7815
		0,3	0,0685	0,0085	0,8025

de referência. Neste contexto, os resultados encontrados nestas tabelas permitiram observar que, de modo geral, a correção de qui quadrado de Yates utilizada na amostra propiciou resultados mais satisfatórios em relação à correção de qui-quadrado de Pearson. Entretanto, ressalva-se que as probabilidades de significância propostas foram submetidas às amostras originais. Este fato é notório nas situações de baixo tamanho amostral e baixo número de variáveis para todas as concentrações de *outliers*, supostamente previsto na amostra.

Ao analisar o aumento do tamanho amostral, na Tabela 6, em conjunto com o incremento do número de variáveis, é possível observar que, de fato, a probabilidade do teste de significância proposto, quando a amostra é submetida à correção qui-quadrado de Yates, em média, está acima de 70%, ao passo que a probabilidade resultante do teste, quando submetido à amostra original, oscila com valores entre 50% e 99%, incluindo resultados mais aberrantes, no caso de $n = 150$, com valores entre 1% e 30%.

A Tabela 7 é referente aos resultados da mistura de uma distribuição normal multivariada com uma distribuição *t-Student* multivariada com mesmas estruturas de covariâncias, como citada na subseção 3.4, item (iv), com $n = 20$ e $n = 50$. A importância deste estudo, considerando a distribuição *t-Student* multivariada, se deve ao fato de que esta distribuição ter caudas mais pesadas do que a distribuição normal, e esta característica tem chamado a atenção em estudos de robustez. Já a Tabela 8, referente à esta mesma mistura, apresenta os resultados para $n = 100$ e $n = 150$.

Diante do exposto, os resultados que mereceram destaque foram descritos na Tabela 7, referindo-se ao desempenho do teste proposto quando foram consideradas as amostras corrigidas pela distância qui-quadrado de Pearson. Neste contexto, observou-se que, de modo geral, o aumento do número de variáveis ocasionou

TABELA 7 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de uma distribuição normal multivariada com uma distribuição *t-Student* multivariada com mesmas estruturas de covariâncias, amostras ($n = 20{,}50$), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
20	5	0	0,6805	0,2510	0,7095
		0,1	0,4445	0,6450	0,0635
		0,2	0,2760	0,7035	0,0265
		0,3	0,1885	0,7540	0,0320
	10	0	0,6750	0,0460	0,6845
		0,1	0,1930	0,2610	0,1050
		0,2	0,0795	0,4545	0,0165
		0,3	0,0235	0,4825	0,0000
	20	0	0,9380	0,0200	0,6710
		0,1	0,2005	0,0745	0,0850
		0,2	0,0925	0,0490	0,0065
		0,3	0,0635	0,1450	0,0020
	50	0	0,4885	0,0180	0,5465
		0,1	0,0835	0,0025	0,1545
		0,2	0,0075	0,0000	0,0155
		0,3	0,0000	0,0005	0,0010
50	5	0	0,5315	0,2300	0,7605
		0,1	0,6295	0,6625	0,0125
		0,2	0,3240	0,6855	0,0205
		0,3	0,1235	0,8170	0,0225
	10	0	0,3180	0,0330	0,7715
		0,1	0,1925	0,3350	0,0000
		0,2	0,0650	0,5230	0,0000
		0,3	0,0100	0,6210	0,0000
	20	0	0,3230	0,0160	0,7690
		0,1	0,0100	0,0230	0,0115
		0,2	0,0085	0,1250	0,0000
		0,3	0,0000	0,1795	0,0000
	50	0	0,9495	0,0090	0,7545
		0,1	0,0550	0,0025	0,0485
		0,2	0,0190	0,0000	0,0000
		0,3	0,0155	0,0000	0,0000

TABELA 8 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates na mistura de uma distribuição normal multivariada com uma distribuição *t-Student* multivariada com mesmas estruturas de covariâncias, amostras ($n = 100{,}150$), diferentes quantidades de variáveis (p) e porcentagens de mistura (γ).

n	p	γ	Amostra original	Correções Qui-quadrado	
				Pearson	Yates
100	5	0	0,3625	0,1440	0,7960
		0,1	0,8090	0,6975	0,0110
		0,2	0,2790	0,6770	0,0180
		0,3	0,0065	0,7240	0,0260
	10	0	0,2075	0,0455	0,7815
		0,1	0,1270	0,3895	0,0000
		0,2	0,0810	0,5605	0,0000
		0,3	0,0000	0,5685	0,0000
	20	0	0,1865	0,0190	0,8010
		0,1	0,0000	0,0635	0,0000
		0,2	0,0000	0,0950	0,0000
		0,3	0,0000	0,1835	0,0000
	50	0	0,1500	0,0085	0,7810
		0,1	0,0000	0,0000	0,0000
		0,2	0,0000	0,0000	0,0000
		0,3	0,0000	0,0000	0,0000
150	5	0	0,3545	0,1615	0,7965
		0,1	0,8605	0,6750	0,0090
		0,2	0,2110	0,7355	0,0120
		0,3	0,0140	0,8115	0,0205
	10	0	0,1465	0,0205	0,8050
		0,1	0,1275	0,4100	0,0000
		0,2	0,0000	0,5030	0,0000
		0,3	0,0000	0,5300	0,0000
	20	0	0,1490	0,0150	0,8030
		0,1	0,0000	0,0435	0,0000
		0,2	0,0000	0,0455	0,0000
		0,3	0,0000	0,0855	0,0000
	50	0	0,1190	0,0140	0,7840
		0,1	0,0000	0,0000	0,0000
		0,2	0,0000	0,0000	0,0000
		0,3	0,0000	0,0000	0,0000

uma redução nos valores das probabilidades empíricas a favor de H_0. A ocorrência deste resultado foi bastante pronunciada em todas as quantidades de *outliers* presentes na amostra e quantificadas pela taxa de mistura γ e tamanhos amostrais.

4.2 Aplicação: seleção das componentes que melhor discriminam *outliers*

Recorrendo-se ao método apresentado na seção 3.2, a metodologia proposta foi ilustrada com uma aplicação em dados reais. Os dados utilizados neste trabalho tornaram-se públicos através do site do IBGE - Instituto Brasileiro de Geografia e Estatística.

Utilizou-se o algoritmo proposto por Filzmoser et al. (2008), descrito na seção 3.2 deste trabalho, para discriminar *outliers* para os dados do índice de volume de vendas do comércio varejista de Minas Gerais, de janeiro de 2007 a dezembro de 2009, em nove segmentos: 1-Combustíveis e Lubrificantes, 2-Hipermercados, supermercados, produtos alimentícios, bebidas e fumo, 3-Hipermercados e supermercados, 4-Tecidos, vestuário e calçados, 5-Móveis e eletrodomésticos, 6-Artigos farmacêuticos, médicos, ortopédicos, de perfumaria e cosméticos, 7-Livros, jornais, revistas e papelaria, 8-Equipamentos e materiais para escritório, informática e comunicação, 9-Outros artigos de uso pessoal e doméstico; que será a matriz $X_{36\times 9}$.

Os coeficientes de curtose padronizados ξ_j, estimados para cada componente principal, encontram-se descritos na Tabela 9.

Tendo por base estes resultados e seguindo a recomendação de Filzmoser et al. (2008) e Peña & Prieto (2001), ao afirmarem que as componentes associadas aos maiores e menores valores de ξ_j resultam na melhor escolha das componentes que identificam os *outliers*, escolheu-se realizar a Figura 5.

Esta figura contém três gráficos sendo que: o gráfico A se refere à plotagem

TABELA 9 Valores do coeficiente de curtose padronizado (ξ_j), para cada componente principal.

Coeficiente de curtose (ξ)	Componentes principais
$\xi_1 = 0{,}5067$	Componente1
$\xi_2 = 0{,}4066$	Componente2
$\xi_3 = 0{,}0026$	Componente3
$\xi_4 = 0{,}0033$	Componente4
$\xi_5 = 0{,}0098$	Componente5
$\xi_6 = 0{,}0029$	Componente6
$\xi_7 = 0{,}0552$	Componente7
$\xi_8 = 0{,}0040$	Componente8
$\xi_9 = 0{,}0089$	Componente9

dos escores da primeira componente principal com os da segunda componente principal, componentes estas geradas a partir dos dados originais; o gráfico B se refere à plotagem da primeira componente, gerada após padronização pela mediana e MAD proposta por Filzmoser et al. (2008), que obteve o maior valor do coeficiente de curtose padronizado ($\xi_1 = 0{,}5067$); o gráfico C se refere à plotagem das componentes 5 e 6, as quais tiveram valores intermediários do coeficiente de curtose padronizados ($\xi_5 = 0{,}0098$ e $\xi_6 = 0{,}0029$).

Os gráficos da Figura 5 evidenciam, de fato, a ocorrência de alguns resultados esperados, como por exemplo, no gráfico **C**; como se pode observar, com pouca discriminação de *outliers*, ressaltam-se que as componentes utilizadas para a construção deste gráfico estão associadas aos coeficientes de curtose ξ_5 e ξ_6 com valores intermediários.

Tendo por referência a plotagem dos escores das duas primeiras componentes obtidas da matriz dos dados originais X, associadas à maior explicação da variabilidade total, nota-se que a plotagem da componente associada ao maior coeficiente de curtose resultou nas mesmas observações identificadas como *outliers*, indicando que pode-se considerar apenas a plotagem dessa primeira componente,

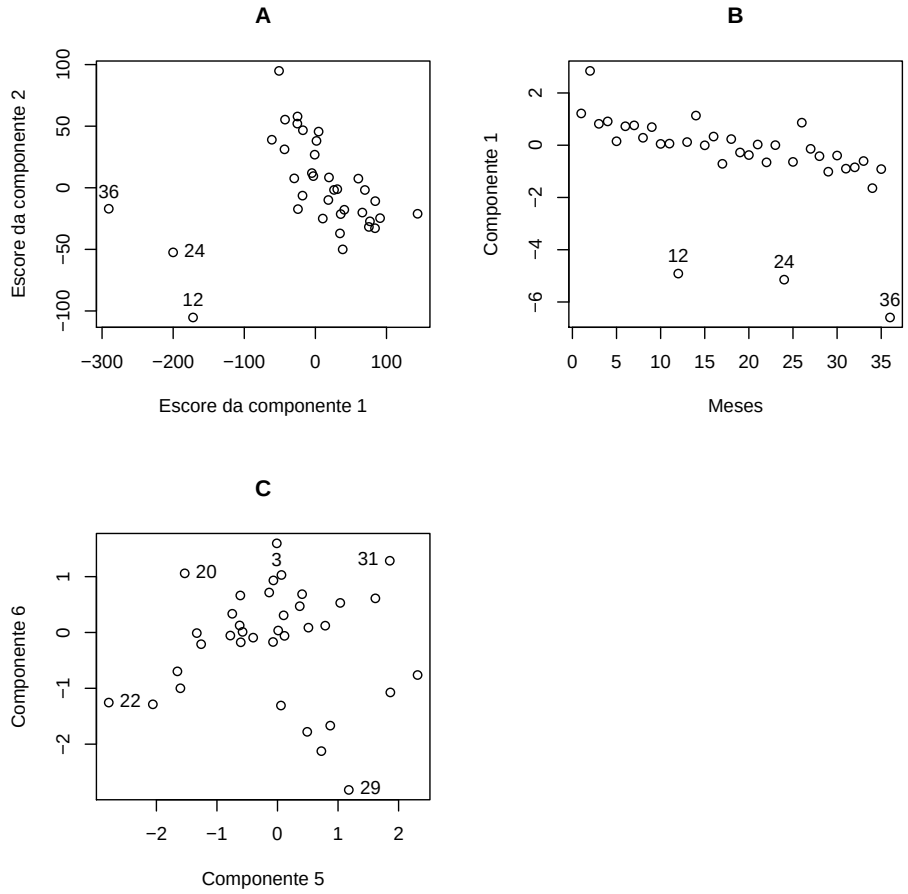

FIGURA 5 Representação gráfica referente aos escores das duas primeiras componentes principais geradas pelos dados originais (A); da componente 1 (B) e das componentes 5 e 6 (C) que, respectivamente, apresentaram o maior e os intermediários valores do coeficiente de curtose robusta.

sendo este o caso mais elementar para que possamos identificar *outliers*. Este fato é perceptível ao comparar os gráficos **A** e **B**, respectivamente, da Figura 5.

A partir dos resultados obtidos, concluiu-se que a componente associada ao maior valor do coeficiente de curtose permitiu discriminar as observações *outli-*

ers, possibilitando uma interpretação adequada ao índice de volume de vendas no comércio varejista de Minas Gerais, nos segmentos analisados.

Pelo fato dos dados serem séries temporais, serão apresentados, a seguir, os gráficos que representam tais séries a fim de identificar sazonalidades. Se os *picos* de sazonalidades acusarem os mesmos pontos considerados como *outliers* na execução do método de Filzmoser et al. (2008), ao invés de usar a plotagem de cada uma das variáveis para identificar possíveis *outliers*, basta que se use tal método.

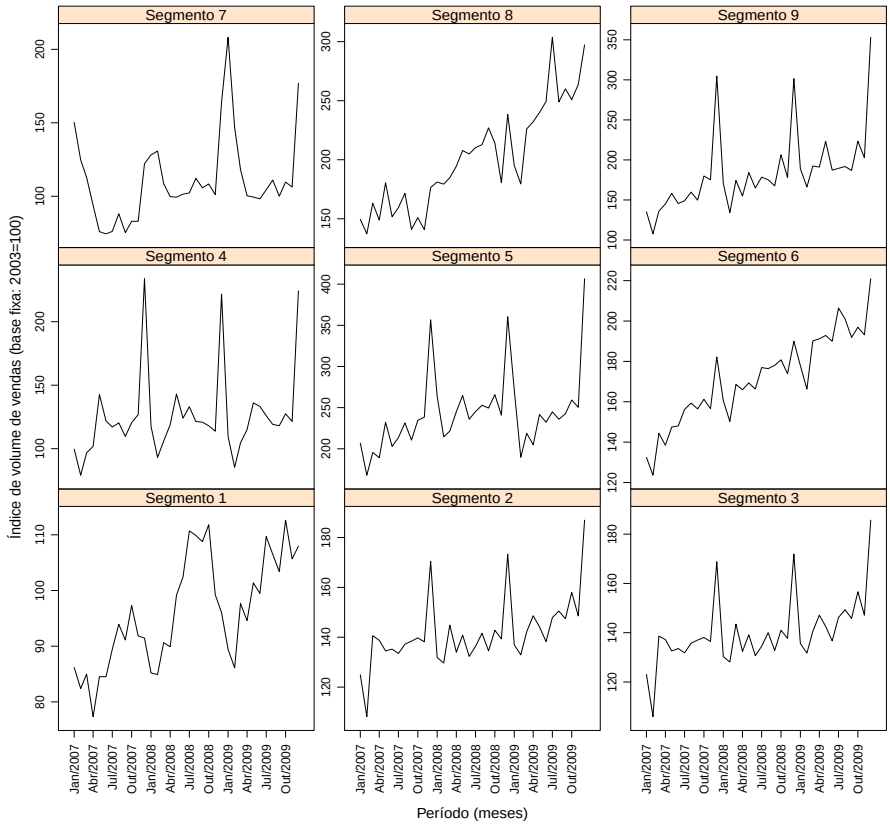

FIGURA 6 Plotagem do índice de volume de vendas, coletados mensalmente, para cada um dos segmentos de mercado.

Desta forma, pode-se afirmar, ainda, que no gráfico B, apresentado na Figura 5, foram identificados os mesmos pontos considerados *picos* de sazonalidade (12, 24 e 36, representando, respectivamente, os meses de Dez/2007, Dez/2008 e Dez/2009), indicando a vantagem trabalhar com o gráfico B ao invés da plotagem de todos os gráfico apresentados na Figura 6.

Com a finalidade de aplicação da metodologia proposta no presente relatório de pesquisa, a matriz X foi corrigida pela distância qui-quadrado de Pearson como descrita na equação 3.1 da seção 3.1 e desenvolveu-se o processo da seção 3.2. Os coeficientes de curtose ξ_{1_j}, estimados para cada segmento de mercado encontram-se na Tabela 10.

TABELA 10 Valores do coeficiente de curtose padronizado (ξ_{1_j}), para cada componente principal, calculados a partir de Q_1, a matriz dos dados corrigida pela distância qui-quadrado de Pearson.

Coeficiente de curtose (ξ)	Componente principal
$\xi_{1_1} = 0{,}1126$	Componente1
$\xi_{1_2} = 0{,}7052$	Componente2
$\xi_{1_3} = 0{,}0101$	Componente3
$\xi_{1_4} = 0{,}0306$	Componente4
$\xi_{1_5} = 0{,}0141$	Componente5
$\xi_{1_6} = 0{,}0112$	Componente6
$\xi_{1_7} = 0{,}0937$	Componente7
$\xi_{1_8} = 0{,}0131$	Componente8
$\xi_{1_9} = 0{,}0095$	Componente9

A Figura 7 contém três gráficos sendo que: o gráfico A se refere à plotagem dos escores da primeira componente principal com os da segunda componente principal, componentes estas geradas a partir dos dados corrigidos pela distância qui-quadrado de Pearson, conforme equação 3.1; o gráfico B se refere à plotagem da segunda com a primeira componente, gerada após padronização, pela mediana e MAD, dos dados corrigidos conforme equação 3.1, que apresentaram, respecti-

vamente, o maior valor do coeficiente de curtose padronizado ($\xi_1 = 0{,}7052$) e a maior explicação da variabilidade total; o gráfico C se refere à plotagem das componentes 5 e 6, as quais tiveram valores intermediários do coeficiente de curtose padronizados ($\xi_5 = 0{,}0141$ e $\xi_6 = 0{,}0112$).

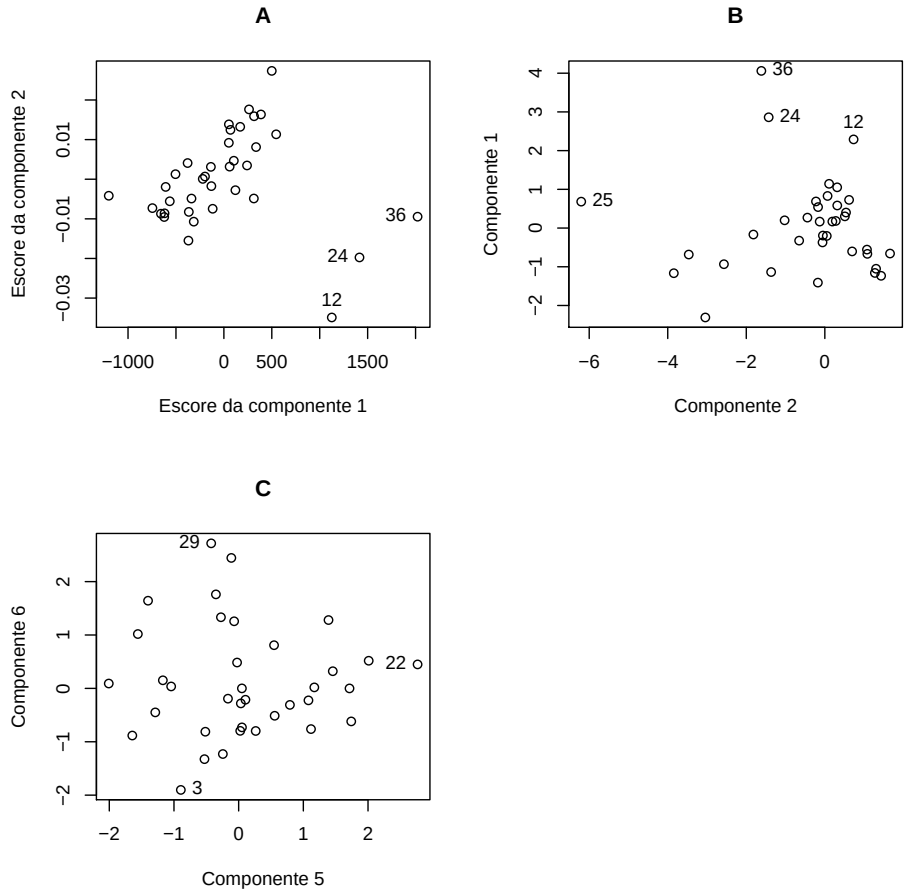

FIGURA 7 Representação gráfica, após os dados serem corrigidos pela distância qui-quadrado de Pearson, referente aos escores das duas primeiras componentes principais geradas pelos dados originais (A); das componentes 1 e 2 (B) e das componentes 5 e 6 (C) que, respectivamente, apresentaram o maior e os intermediários valores do coeficiente de curtose robusta.

Nos gráficos da figura 7 evidencia-se, de fato, a ocorrência de alguns resultados esperados, como por exemplo, no gráfico C; da mesma forma que na Figura 5, observou-se pouca discriminação de *outliers*, estando os pontos menos distantes uns dos outros, ressaltando que as componentes utilizadas para a construção deste gráfico estão associadas aos coeficientes de curtose ξ_{1_5} e ξ_{1_6}, cujos valores são intermediários.

A plotagem dos escores das duas primeiras componentes principais (A) obtidas da matriz dos dados corrigida pela distância Qui-quadrado de Pearson, Q_1, associadas à maior explicação da variabilidade total, e a plotagem das componentes 1 e 2 (B), já padronizadas pela mediana e MAD, associadas, respectivamente, à maior explicação da variabilidade total e ao maior coeficiente de curtose padronizado ($\xi_{1_2} = 0{,}7052$), resultaram nas mesmas observações identificadas como *outliers*, a menos do gráfico B, que indicou que a observação 25 (Jan/2009) pode ser um possível outlier.

O próximo passo foi aplicar o teste de significância. Tem-se que todo teste de significância é baseado em repetição. No caso dessa aplicação, a repetição foi representada por cinco mil reamostragens da amostra original feita pelo método *bootstrap*. Com base nessas reamostras, o teste de significância proposto no presente relatório de pesquisa foi aplicado nessa seção, obtendo as probabilidades (valor-p) apresentadas na Tabela 11:

TABELA 11 Probabilidades (valor-p) do teste de significância, considerando a amostra original e as amostras corrigidas pelas distâncias qui-quadrado de Pearson e de Yates, com base nas reamostras obtidas pelo método *bootstrap*.

Teste Aleatorização	valor-p
Amostra original	0,3814
Pearson	0,4936
Yates	0,3908

As probabilidades obtidas na Tabela 11 indicam que os dados corrigidos pela distância do tipo qui-quadrado de Pearson aumenta a probabilidade a favor de H_0, mostrando-se superior ao valor obtido pela correção qui-quadrado de Yates.

A partir desses resultados obtidos, mais uma vez a correção qui-quadrado foi promissora, mesmo sem se conhecer a estrutura de covariância populacional e média populacional e muito menos a porcentagem de *outliers*. Com esse resultado é confirmada a eficiência de se utilizar a correção qui-quadrado de Pearson na matriz original dos dados.

5 CONCLUSÕES

- Os resultados propiciados pela correção Qui-quadrado de Pearson foram promissores, de modo a evidenciar mais expressivamente que, o primeiro coeficiente de curtose padronizado é nulo. Este fato foi perceptível para ambos os tamanhos amostrais avaliados, incluindo até mesmo situações com o número de variáveis, excessivamente, elevado.

- Para o tamanho amostral $(n = 20)$ é viável aplicar a correção de Qui-quadrado de Pearson apenas quando o número de variáveis é pequeno. Para tamanhos amostrais elevados $(n = 50{,}100{,}150)$ ocorreu uma redução na probabilidade a favor de H_0 quando aplicado a correção Qui-quadrado. Assim, recomenda-se que a identificação destas observações seja feita considerando a amostra original.

- Para uma quantidade pequena de variáveis, a correção de Qui-quadrado de Pearson apresentou melhores resultados. Porém, de modo geral, o aumento do número de variáveis ocasionou uma redução na probabilidade a favor de H_0.

- Em relação à aplicação da metodologia nos dados referentes ao índice de volume de vendas no comércio varejista de Minas Gerais, a plotagem das componentes de maior coeficiente de curtose padronizado com a componente associada à maior porcentagem de explicação da variabiliade total, calculadas para a amostra corrigida pela distância qui-quadrado de Pearson, discriminam melhor *outliers*.

REFERÊNCIAS BIBLIOGRÁFICAS

CHEN, J.; TAN, X. Inference for multivariate normal mixtures. **Journal of Multivariate Analysis**, New York, v.100, n.7, p.1367–1383, Aug. 2009.

COOK, R. D. Assessment of local influence (with discussions). **Journal of Royal Statistical Society:** series B, statistical methodology, Oxford, v.48, n.2, p.133–169, Feb. 1986.

CRITCHLEY, F. Influence in principal component analysis. **Biometrika**, London, v.72, n.3, p.627–636, Dec. 1985.

FILZMOSER, P.; MARONNA, R.; WERNER, M. Outlier identification in high dimensions. **Computational Statistics and Data Analysis**, Amsterdam, v.52, n.3, p.1694–1711, Jan. 2008.

GNANADESIKAN, R.; KETTENRING, J. R. Robust estimates, residuals and outlier detection with multiresponse data. **Biometrics**, Washington, v.28, n.1, p.81–124, Jan. 1972.

HAIR JÚNIOR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005. 593p.

HAWKINS, D. M. The detection of errors in multivariate data using principal components. **Journal of the American Statistical Society**, New York, v.69, n.345, p.340–344, Jun. 1974.

JOHNSON, M. E. **Multivariate statistical simulation**. New York: J. Wiley, 1987.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer Verlag, 2002. 487p.

LANGE, K. L.; RODERICK, J. A. L.; TAYLOR, J. M. G. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, New York, v.84, n.408, p.881–896, Dec. 1989.

PEÑA, D.; PRIETO, F. Multivariate outlier detection and robust covariance matrix estimation. **Technometrics**, Washington, v.43, n.3, p.286–310, Aug. 2001.

PEREIRA, J. C. R. **Análise de dados qualitativos:** estratégias metodológicas para as Ciências da Saúde, Humanas e Sociais. 3. ed. São Paulo: USP, 2004. 115–118p.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2009.

ROUSSEEUW, P. J. Least median of squares regression. **Journal of the American Statistical Association**, New York, v.79, n.388, p.871–880, Dec. 1984.

ANEXOS

ANEXO A Rotina para a simulação de uma configuração na mistura, neste caso, de duas distribuições normais multivariadas com vetores de médias e estruturas de covariâncias diferentes.

```
# Definindo os Parâmetros #
namo <- 20   # tamanho da amostra
p    <- 50   # quantidade de variáveis
gama <- 0.3  # probabilidade de mistura
pho  <- 0.1  # coeficiente de correlação
nsim <- 20   # número de simulações

# Definindo as matrizes de covariâncias #
AR  <- matrix(0,p,p)
eco <- matrix(0,p,p)

# Definindo os vetores de médias #
mi1 <- c(rep(0,p))
mi2 <- c(rep(100,p))

# Definindo o vetor para os coeficientes de curtose
# padronizados #
wp  <- c(rep(0,p))

# Definindo as matrizes dos dados corrigidos pelas
# distâncias qui-quadrado #
Qsy <- matrix(0,namo,p) # Pearson
Qy  <- matrix(0,namo,p) # Yates

# Definindo a matriz dos dados originais #
X   <- matrix(0,1,p)

# Configuração para gerar os valores-p na amostra
# original (usual), na amostra corrigida pela
# distância qui-quadrado de Pearson (qui) e pela
# de Yates (yates) #
conta1 <- 0
conta2 <- 0
conta0 <- 0

est_av <- matrix(0,nsim,6)
mdat   <- matrix(0,3,1, byrow=TRUE,
          dimnames = list(c("usual", "qui", "yates"),
          c("prob"))) # lista os valores-p

# Definindo os tipos de matrizes de covariâncias #
```

```
for (i in 1:p)
{
   for (j in 1:p)
   {
      if (i==j)  AR[i,j]  <- 1
      if (i!=j)  AR[i,j]  <- pho^(abs(i-j))
      if (i==j)  eco[i,i] <- 1
      if (i!=j)  eco[i,j] <- pho
   }
}
covp1 <- AR
covp2 <- eco

# Funções #

# Definindo as matrizes dos dados corrigidos pelas
# distâncias qui-quadrado #
cor_qui <- function(X,p)
{
   sl <- 0
   for ( i in 1:nrow(X))
   {
      sl <- sum(X[i,])
      for (j in 1:p)
      {
         Qsy[i,j] <- (X[i,j] - sl*sum(X[,j]))/
                     (sqrt(abs(sl*sum(X[,j]))))
         # Pearson
         Qy[i,j]  <- (X[i,j] - abs(sl*sum(X[,j])) - 1)^2/
                     (sl + sum(X[,j]))
         # Yates
      }
   }
   return(list(Q1=Qsy,Q2=Qy))
}

# REESCALANDO OS DADOS ROBUSTAMENTE COM MEDIANA E MAD:
res <- function(X)
{
   xmed <- apply(X,2,median)
   xmad <- apply(X,2,mad)
   Xij  <- matrix(0,nrow(X),ncol(X))
   for (j in 1:length(xmed))
   {
      for (i in 1:nrow(X))
      {
         Xij[i,j] <- ((X[i,j] - xmed[j]) / xmad[j])
         # matriz dos dados padronizados pela
         # mediana e MAD
      }
   }
   return(Xij)
```

```
}

curt <- function(Zesc,Zaux)
{
   K    <- matrix(0,nrow(Zesc),ncol(Zesc))
   zmed <- apply(Zaux,2,median)
   zmad <- apply(Zaux,2,mad)
   for (j in 1:length(zmed))
   {
      for(i in 1:nrow(Zesc))
      {
         K[i,j] <- ((Zesc[i,j] - zmed[j])^4/
                    zmad[j]^4) - 3
      }
   }
   W  <- abs(colSums(K) / nrow(Zesc))
   # Coeficiente de curtose robusta
   return(W)
}

pesocurt <- function(coef,p)
{
   for (i in 1:p)
   {
      totp1 <- 0
      waux  <- coef[i]
      for (j in 1:p)
      {
         totp1 <- totp1 + coef[j]
      }
      wp[i] <- waux / totp1
      # Coeficiente de curtose padronizado
   }
   resw1 <- wp
   return(resw1)
}


# INÍCIO DO PROGRAMA #
for (s in 1:nsim)
{
# Normal contaminada #
   for (r in 1:namo)
   {
      u <- runif(1)
      if (u <= gama)   Xaux <- mvrnorm(1, mi1, covp1)
      if (u >  gama)   Xaux <- mvrnorm(1, mi2, covp2)
      X <- rbind(X,Xaux)
   }
   X <- X[2:namo,1:p]

# Algoritmo usual #
```

```
   chama_res  <- res(X)
   sigma      <- cov(chama_res)
   svdsigma   <- svd(sigma)

   V <- svdsigma$v
   Z <- chama_res %*% V

   chama2_res <- res(Z)

# Modificação com correção qui-quadrado #
   chama_cor_qui <- cor_qui(X,p)
   MCQ1   <- t(chama_cor_qui$Q1) %*% chama_cor_qui$Q1
          ### Matriz de covariância associada as variáveis
   MCQ2   <- t(chama_cor_qui$Q2) %*% chama_cor_qui$Q2
          ### Matriz de covariância associada as variáveis

   vq1 <- svd(MCQ1)
   vq2 <- svd(MCQ2)

# Calculo dos coeficientes de curtose #
   chama_curtq1 <- curt(chama_cor_qui$Q1,vq1$v)
   chama_curtq2 <- curt(chama_cor_qui$Q2,vq2$v)
   chama_curt   <- curt(chama2_res,Z)

# Calculo dos coeficientes de curtose padronizados #
   chama_pesocurt   <- pesocurt(chama_curt,p)
   chama_pesocurtq1 <- pesocurt(chama_curtq1,p)
   chama_pesocurtq2 <- pesocurt(chama_curtq2,p)

# Estatística para seleção dos dois primeiros componentes #
   wq0  <- max(chama_pesocurt)
   w1q1 <- max(chama_pesocurtq1)
   w1q2 <- max(chama_pesocurtq2)

   pvarq0 <- (svdsigma$d[1]) / sum(svdsigma$d)
   pvarq1 <- (vq1$d[1]) / sum(vq1$d)
   pvarq2 <- (vq2$d[1]) / sum(vq2$d)

   if (wq0  >= pvarq0)  conta0 <- conta0 + 1
   if (w1q1 >= pvarq1)  conta1 <- conta1 + 1
   if (w1q2 >= pvarq2)  conta2 <- conta2 + 1

   est_av[s,1] <- wq0
   est_av[s,2] <- pvarq0
   est_av[s,3] <- w1q1
   est_av[s,4] <- pvarq1
   est_av[s,5] <- w1q2
   est_av[s,6] <- pvarq2
}
pAM0 <- conta0 / nsim # valor-p para a amostra original
pAM1 <- conta1 / nsim # valor-p para a amostra corrigida
                      # pela distância qui-quadrado de
```

```
                      # Pearson
pAM2 <- conta2 / nsim # valor-p para a amostra corrigida
                      # pela distância qui-quadrado de
                      # Yates
mdat[1,1] <- pAM0
mdat[2,1] <- pAM1
mdat[3,1] <- pAM2
mdat # lista os valores-p
```

ANEXO B Rotina para a aplicação da metodologia proposta num conjunto de dados reais.

```
# Requerendo a biblioteca MASS do software R #
require(MASS)

# Importando o aquivo dos dados amostrais e definindo a
# matriz X #
arq <- read.table("dados.txt",h=T)
attach(arq)

X <- as.matrix(arq)
X

# Informando o número de colunas e tamanho da amostra #
p=ncol(arq) ; namo=nrow(arq)  ; nboot=5000

# Definindo as matrizes de covariância #
Qsy=matrix(0,namo,p)
Qy=matrix(0,namo,p)

# Definindo o vetor dos coeficientes de curtose
# padronizados #
wp=c(rep(0,p))

# Definindo o vetor com os valores-p #
conta1=0 ;   conta2=0 ; conta0=0

mdat <- matrix(0,3,1,byrow=TRUE,
               dimnames = list(c("usual", "qui", "yates"),
               c("prob")))

# FUNÇÕES #
# Definindo as matrizes dos dados corrigidos pelas
# distâncias qui-quadrado #
cor_qui=function(X,p)
{
 sl=0
 for( i in 1:nrow(X))
 {
    sl=sum(X[i,])
    for (j in 1:p)
    {
        Qsy[i,j]=(X[i,j]-sl*sum(X[,j]))/
                 (sqrt(abs(sl*sum(X[,j]))))
        # Pearson
        Qy[i,j]=(X[i,j]-abs(sl*sum(X[,j]))-1)^2/
                 sl+sum(X[,j])
        # Yates
    }
```

```
 }
 return(list(Q1=Qsy,Q2=Qy))
}

# REESCALANDO OS DADOS ROBUSTAMENTE COM MEDIANA E MAD:
res=function(X)
{
   xmed   <- apply(X,2,median)
   xmad   <- apply(X,2,mad)
   Xij     <- matrix(0,nrow(X),ncol(X))
   for(j in 1:length(xmed))
   {
       for(i in 1:nrow(X))
       {
          Xij[i,j] <- ((X[i,j] - xmed[j])/xmad[j])
       }
   }
   return(Xij)
}

curt=function(Zesc,Zaux)
{
    K <- matrix(0,nrow(Zesc),ncol(Zesc))
    zmed   <- apply(Zaux,2,median)
    zmad   <- apply(Zaux,2,mad)
    for(j in 1:length(zmed))
    {
        for(i in 1:nrow(Zesc))
        {
          K[i,j] <- ((Zesc[i,j] - zmed[j])^4/zmad[j]^4)-3
        }
    }
    W  <- abs(colSums(K)/nrow(Zesc))
    return(W) # coeficientes de curtose
}

pesocurt=function(coef,p)
{
   for (i in 1:p)
   {
      totp1=0
      waux=coef[i]
      for (j in 1:p)
      {
         totp1=totp1+coef[j]
      }
      wp[i]=waux/totp1
   }
   resw1=wp # coeficientes de curtose padronizados
   return (resw1)
}
```

```
# INÍCIO DO PROGRAMA #
# Algoritmo usual #
chama_res=res(X)
sigma=cov(chama_res)
svdsigma <- svd(sigma);
V <- svdsigma$v
Z <- chama_res%*%V
chama2_res=res(Z)

# Modificando a amostra com correção qui-quadrado #
chama_cor_qui=cor_qui(X,p)
MCQ1=t(chama_cor_qui$Q1)%*%chama_cor_qui$Q1
### Matriz de covariância associada as variáveis
MCQ2=t(chama_cor_qui$Q2)%*%chama_cor_qui$Q2
### Matriz de covariância associada as variáveis
vq1=svd(MCQ1) ; vq2=svd(MCQ2)

# Cálculo dos coeficientes de curtose #
chama_curtq1=curt(chama_cor_qui$Q1,vq1$v)
chama_curtq2=curt(chama_cor_qui$Q2,vq2$v)
chama_curt=curt(chama2_res,Z)

# Cálculo dos coeficientes de curtose padronizados #
chama_pesocurt=pesocurt(chama_curt,p)
chama_pesocurtq1=pesocurt(chama_curtq1,p)
chama_pesocurtq2=pesocurt(chama_curtq2,p)

# Estatística para seleção dos dois primeiros
# componentes #
wq0=max(chama_pesocurt)
w1q1=max(chama_pesocurtq1)
w1q2=max(chama_pesocurtq2)


# INÍCIO DO BOOTSTRAP #
for (a in 1:nboot)
{
  u=round(runif(nrow(X),1,nrow(X)))
  cmb=matrix(0,1,ncol(X))
  for (j in 1:nrow(X))
  {
    aux=u[j] ;  mb=X[aux,]
    cmb=rbind(cmb,mb);
  }
  cmb=cmb[2:nrow(cmb),]
  chama_resb=res(cmb) ; sigmab=cov(chama_resb)
  svdsigmab <- svd(sigmab) ;
  Vb <- svdsigmab$v  ;  Zb <- chama_resb%*%Vb
  chama2_resb <- res(Zb)

  ### Modificando a amostra com correção qui-quadrado
```

```
  ### versão bootstrap #####
      chama_cor_quib=cor_qui(cmb,p)
      ### Matriz de covariância associada as variáveis
      MCQ1b=t(chama_cor_quib$Q1)%*%chama_cor_quib$Q1
      ### Matriz de covariância associada as variáveis
      MCQ2b=t(chama_cor_quib$Q2)%*%chama_cor_quib$Q2
      vq1b=svd(MCQ1b) ; vq2b=svd(MCQ2b)

 ### Cálculo dos coeficientes de curtose
 ### versão bootstrap ######
     chama_curtq1b=curt(chama_cor_quib$Q1,vq1b$v)
     chama_curtq2b=curt(chama_cor_quib$Q2,vq2b$v)
     chama_curtb=curt(chama2_res,Zb)

 ### Cálculo dos coeficientes de curtose padronizados
 ### versao bootstrap ######
     chama_pesocurtb=pesocurt(chama_curtb,p)
     chama_pesocurtq1b=pesocurt(chama_curtq1b,p)
     chama_pesocurtq2b=pesocurt(chama_curtq2b,p)

 ### Estatística bootstrap para seleção dos dois primeiros
 ### componentes ########
     wq0b=max(chama_pesocurtb)
     w1q1b=max(chama_pesocurtq1b)
     w1q2b=max(chama_pesocurtq2b)
     if (wq0b>=wq0)    conta0=conta0+1
     if (w1q1b>=w1q1)  conta1=conta1+
     if (w1q2b>=w1q2)  conta2=conta2+1
 }
pAM0=conta0/nboot
pAM1=conta1/nboot
pAM2=conta2/nboot
mdat[1,1]=pAM0
mdat[2,1]=pAM1
mdat[3,1]=pAM2
mdat # lista os valores-p
```

ANEXO C Rotina para a construção dos gráficos de sazonalidade num conjunto de dados reais.

```
# Importando o conjuntos de dados
ibge <- read.table("dados_boot.txt", header=TRUE)
attach(ibge)
X <- ibge
X
# Definindo o nome das variáveis:
names(X) <- c("Segmento 1", "Segmento 2", "Segmento 3",
              "Segmento 4", "Segmento 5", "Segmento 6",
              "Segmento 7", "Segmento 8", "Segmento 9")

# INÍCIO DO PROGRAMA PARA CONSTRUÇÃO GRÁFICA:
# Abrindo o pacote para construções gráficas do
# software R, Lattice:
require(lattice)
# Definindo o eixo x:
xtime <- seq(as.Date("2007-01-01"),
             as.Date("2009-12-01"), by="month")
# Definindo a matrix X:
X <- as.data.frame(X)
dados <- stack(X)
dados$x <- xtime
# Salvando o gráfico em pdf com o comando
# pdf(...); dev.off()):
pdf(file="plotsazon.pdf",h=10, w=10)
# Gerando os gráficos numa única figura:
xyplot(values~as.POSIXct(x, format="%b/%y")|ind,
       data=dados, type="l", col="black", layout=c(3,3),
       scales=list(alternating=FALSE, tck=c(1,0),
       y=list(relation="free"),
       x=list(at=as.POSIXct(unique(dados$x)[seq(1, by=3,
              length.out=length(unique(dados$x)))],
              format="%b/%Y"), rot=90,
              labels=format(unique(dados$x)[seq(1, by=3,
              length.out=length(unique(dados$x)))],
              format="%b/%Y"))),
       xlab="Período (meses)",
       ylab="Índice de volume de vendas
             (base fixa: 2003=100)")
dev.off()
```