

**PROPOSTA DE TESTES *BOOTSTRAP* PARA INFERIR  
SOBRE O NÚMERO DE COMPONENTES PRINCIPAIS  
RETIDOS**

**DEYSE MÁRCIA PACHECO GEBERT**

**2010**

**DEYSE MÁRCIA PACHECO GEBERT**

**PROPOSTA DE TESTES *BOOTSTRAP* PARA INFERIR SOBRE O  
NÚMERO DE COMPONENTES PRINCIPAIS RETIDOS**

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

Orientador  
Prof. Dr. Daniel Furtado Ferreira

LAVRAS  
MINAS GERAIS -BRASIL  
2010

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

Gebert, Deyse Márcia Pacheco.

Proposta de testes *bootstrap* para inferir sobre o número de componentes principais retidos / Deyse Márcia Pacheco Gebert. – Lavras : UFLA, 2010.

75 p. : il.

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

Orientador: Daniel Furtado Ferreira.

Bibliografia.

1. Multivariada. 2. Teste *bootstrap*. 3. Reamostragem. I. Universidade Federal de Lavras. II. Título.

CDD-519.535

**DEYSE MÁRCIA PACHECO GEBERT**

**PROPOSTA DE TESTES *BOOTSTRAP* PARA INFERIR SOBRE O  
NÚMERO DE COMPONENTES PRINCIPAIS RETIDOS**

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

APROVADA em 22 de fevereiro de 2010

Prof. Dr. Eric Batista Ferreira

UNIFAL

Prof. Dr. Renato Ribeiro de Lima

UFLA

Prof. Dr. Daniel Furtado Fereira  
UFLA  
(Orientador)

LAVRAS  
MINAS GERAIS-BRASIL

DEDICO

*A minha mãe, Rosa,  
meu filho, Vinícius,  
minha irmã, Débora,  
e meu sobrinho, Gabriel.*

## AGRADECIMENTOS

A Deus, pela luz nos momentos que me pareciam escuros demais.

A minha irmã Débora, que desde os tempos de faculdade me incentiva e apoia. Sem a sua ajuda não teria alcançado meu objetivo.

A minha mãe Rosa, pela força nos momentos de fraqueza e pelo ombro (via telefone) nos momentos de tristeza.

A meu filho Vinícius, pela compreensão das minhas ausências devido à distância.

Ao meu orientador Daniel, pela sua grande ajuda, atenção e confiança depositada em mim.

A CAPES e o programa REUNI, pela oportunidade de trabalho e ajuda financeira.

A todos os amigos do DEX, em especial o “Clube da Luluzinha”: Flávia, Franciella, Rosi e Simone, por compartilharem comigo momentos de tensão e também de alegrias. Uma verdadeira amizade!

A todos os meus amigos, que de alguma forma contribuíram para essa vitória.

Muito Obrigada!

## SUMÁRIO

LISTA DE TABELA . . . . .	i
LISTA DE FIGURAS . . . . .	ii
RESUMO . . . . .	iv
ABSTRACT . . . . .	v
1 INTRODUÇÃO . . . . .	1
2 REFERENCIAL TEÓRICO . . . . .	3
2.1 Análise multivariada . . . . .	3
2.1.1 Análise de componentes principais . . . . .	4
2.1.2 Critérios de retenção de componentes principais . . . . .	8
2.2 Simulação <i>bootstrap</i> . . . . .	12
2.2.1 <i>Bootstrap</i> paramétrico . . . . .	13
2.2.2 <i>Bootstrap</i> não-paramétrico . . . . .	14
2.3 Simulação Monte Carlo . . . . .	14
2.4 Erro tipo I, erro tipo II, poder e robustez . . . . .	16
2.5 Distribuições . . . . .	19
2.5.1 Distribuição normal multivariada . . . . .	20
2.5.2 Distribuição <i>t</i> de Student multivariada . . . . .	21
3 METODOLOGIA E VALIDAÇÃO MONTE CARLO . . . . .	23
4 RESULTADOS E DISCUSSÃO . . . . .	30
4.1 Resultados metodológicos . . . . .	30
4.1.1 Testes de <i>bootstrap</i> paramétrico . . . . .	30
4.1.2 Teste de <i>bootstrap</i> não-paramétrico . . . . .	34
4.2 Validação Monte Carlo . . . . .	35
4.2.1 Validação sob normalidade . . . . .	35

4.2.2 Validação sob a $t$ multivariada . . . . .	44
5 CONCLUSÕES . . . . .	51
REFERÊNCIAS BIBLIOGRÁFICAS . . . . .	52
ANEXOS . . . . .	55



## LISTA DE TABELA

1	Resultados possíveis de um teste de hipóteses e suas probabilidades de ocorrência. . . . .	17
---	--	----

## LISTA DE FIGURAS

- 1 Curva da função poder em relação a um parâmetro  $\theta$  . . . . . 19
- 2 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . . 37
- 3 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . . 39
- 4 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 41
- 5 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 43

- 6 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 45
- 7 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 46
- 8 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 48
- 9 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ . 49

## RESUMO

GEBERT, Deyse Márcia Pacheco. **Proposta de testes *bootstrap* para inferir sobre o número de componentes principais retidos.** 2010. 75 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras. \*

Uma de muitas metas da análise multivariada é a redução da dimensionalidade, ou seja, a busca por modelos mais parcimoniosos. Neste sentido, uma técnica utilizada é a análise de componentes principais. Após a determinação dos componentes principais deve-se definir qual o número ideal a ser usado, de modo a sintetizar a informação contida nas variáveis originais em um modelo mais simplificado. Alguns critérios para essa escolha são propostos na literatura, a maioria deles possui sérias limitações sendo essa uma justificativa para esse estudo. O objetivo do presente trabalho foi propor novos testes para a determinação do número ideal de componentes principais a serem retidos, baseados na proporção de explicação da variação total dos  $k$  primeiros componentes. A avaliação das taxas de erro tipo I, poder e robustez desses métodos foi feita por meio de simulações Monte Carlo. Tais testes foram comparados entre si e com o teste de Fujikoshi. Três novos testes *bootstrap* para a proporção de explicação de componentes principais retidos foram propostos com sucesso. Sob normalidade multivariada os dois testes de *bootstrap* paramétrico são recomendados, pois tiveram os maiores valores de poder e controlaram o erro tipo I. Sob não-normalidade o teste *bootstrap* não-paramétrico é recomendado por ser robusto, uma vez que foi o único a controlar o erro tipo I. Os testes propostos superaram o teste de Fujikoshi na maioria das situações.

**Palavras-chave:** Multivariada, componentes principais e teste *bootstrap*.

---

\*Orientador: Daniel Furtado Ferreira -UFLA .

## ABSTRACT

GEBERT, Deyse Márcia Pacheco. **On Bootstrap tests for the retained number of principal components.** 2010. 75 p. Dissertation (Master in Statistics and Agricultural Experimentation) - Federal University of Lavras, Lavras. \*

One of the many goals of the multivariate analysis is to reduce dimensionality, i.e., search for more parsimonious models. In this sense, the technique of principal component can be used. After the principal components were determined one should choose the ideal number of components to be retained in order to synthesize the information contained in the  $p$  original variates in a simplified model. Some criteria for this choice are proposed in the literature, most of them have serious limitations that were used as justification for this study. The aim of this study was to propose new tests for determining the optimal number of principal components to be retained, based on the proportion of explanation of the total variation of the  $k$  first components ( $k < p$ ). The evaluation of type I error rates, power and robustness of these tests was made by Monte Carlo simulations. These tests were compared among themselves and with Fujikoshi's test. The three new bootstrap tests were proposed with success. Under multivariate normality the two parametric bootstrap tests are recommended, since they showed the highest power and controlled the type I error. Under non-normality the non-parametric bootstrap test was considered robust, since it was the only one to control the type I error and should be recommended. The proposed tests overcome the test of Fujikoshi in most circumstances.

**Key words:** Multivariate, principal components and test bootstrap.

---

\* Adviser: Daniel Furtado Ferreira - UFLA

## 1 INTRODUÇÃO

Na grande maioria das pesquisas realizadas, mensura-se mais do que uma variável resposta em cada indivíduo, objeto ou amostra. Nem sempre esses dados são analisados de forma multivariada. As variáveis mensuradas possuem relações entre si e a estatística multivariada explora essa relação, levando à análises mais robustas e informativas. De modo geral, análises que envolvem muitas variáveis, têm interpretações mais complexas. Uma das metas da estatística multivariada é a busca por modelos mais simplificados, com menos variáveis. Neste sentido, uma técnica utilizada é a análise de componentes principais.

A análise de componentes principais é uma transformação linear das variáveis originais, que são correlacionadas, em novas variáveis não correlacionadas. Essa transformação deve ser feita de forma a maximizar a variância dos componentes e então organizá-los de forma decrescente. A geração de componentes principais (CP) nada mais é que a determinação dos autovalores e autovetores da matriz de covariâncias. Após a determinação dos CP deve-se definir qual o número ideal a ser usado, de modo a sintetizar a informação contida nas variáveis originais em um modelo mais simplificado.

Muitos critérios para a escolha do número ideal de CP são propostos na literatura. Grande parte deles é baseada em argumentos empíricos, como por exemplo o gráfico de autovalores (Jolliffe, 2002). Além desse método, existe o teste para a igualdade dos últimos autovalores (Ferreira, 2008), o método baseado no autovalor médio (Perez-Neto et al., 2005), o teste de Bartlett (Grossman et al., 1991), o critério de Kaiser-Guttman (Perez-Neto et al., 2005), o teste de Fujikoshi (Fujikoshi, 1980), entre outros. A maioria dos testes existentes possuem sérias limitações, sendo essa uma justificativa para esse estudo.

O presente trabalho teve como objetivo os seguintes pontos:

- propor novos testes para a determinação do número ideal de componentes principais a serem retidos no modelo, baseados na proporção de explicação da variação total dos  $k$  primeiros componentes;
- avaliação de taxas de erro tipo I, poder e robustez desses métodos por meio de simulações Monte Carlo;
- comparações entre os testes propostos e com o teste de Fujikoshi (1980).

## 2 REFERENCIAL TEÓRICO

### 2.1 Análise multivariada

A expressão análise multivariada, como o próprio nome já diz, é usada para descrever análises de dados que são coletados em  $p$  variáveis ( $p > 1$ ) para cada indivíduo, unidade experimental ou amostral estudada. De acordo com Rencher (2002), a análise multivariada consiste em uma coleção de métodos que podem ser utilizados quando mensura-se mais do que uma variável resposta em cada indivíduo, objeto ou amostra. O desafio está em desvendar o complicado interrelacionamento entre essas várias medidas, sendo a interpretação destes resultados o que torna a análise multivariada uma atividade gratificante para o pesquisador (Afifi, 1996).

Na prática, conjuntos de dados multivariados são muito comuns, embora nem sempre esses dados sejam analisados simultaneamente. Na atualidade a utilização exclusiva de procedimentos univariados não é justificável, dada a disponibilidade de técnicas computacionais multivariadas para realizar tais análises.

A grande maioria das pesquisas realizadas raramente está interessada em apenas uma variável. Essas variáveis em questão, possuem relações entre si e a estatística multivariada explora essa relação, levando a análises mais robustas e informativas. Os experimentos ou amostragens, de modo geral, envolvem muitas variáveis, o que dificulta a interpretação das análises univariadas realizadas. Uma de muitas metas da análise multivariada é a simplificação, ou seja, a busca por modelos mais parcimoniosos, modelos mais simples porém, com um percentual de explicação da variância total adequado para a situação. Neste sentido, uma técnica utilizada é a análise de componentes principais.



### 2.1.1 Análise de componentes principais

Jolliffe (2002) fala que reduzir a dimensionalidade de um conjunto de dados composto de um grande número de variáveis e manter tanto quanto possível, a variação presente nele é a ideia central da análise de componentes principais. Esse resultado é obtido pela transformação linear de um conjunto de variáveis em um novo conjunto, os componentes principais, que não são correlacionados e que são ordenados de forma que os primeiros componentes detém a maior parte da variação contida nas variáveis originais. Um pequeno grupo de variáveis não correlacionadas é mais fácil de ser interpretado e usado para análises futuras do que um grande grupo de variáveis correlacionadas (Dunteman, 1989). A ideia da análise de componentes principais foi originalmente concebida por Pearson (1901) e independentemente desenvolvida por Hotelling (1933).

A análise de componentes principais não é somente utilizada no sentido de redução da dimensionalidade. Muitos pesquisadores a empregam para resolver problemas de multicolinearidade em regressão múltipla, detecção de *outliers*, análise de modelos *AMMI* (*Additive main effects and multiplicative interaction*), sendo também usada em outras técnicas multivariadas como: análise discriminante, análise fatorial, análise de correlação canônica, entre outras possibilidades.

Harris (2001) especifica componentes principais como sendo combinações lineares das  $p$  variáveis observadas, que são correlacionadas. Estas combinações lineares não possuem correlação e são em número, iguais as variáveis originais. Considere uma situação em que  $p = 3$  variáveis ( $X_1$ ,  $X_2$  e  $X_3$ ), tal que, as combinações lineares,

$$\begin{aligned}
Y_1 &= a_{11}X_1 + a_{12}X_2 + a_{13}X_3, \\
Y_2 &= a_{21}X_1 + a_{22}X_2 + a_{23}X_3, \\
Y_3 &= a_{31}X_1 + a_{32}X_2 + a_{33}X_3
\end{aligned}$$

são os componentes principais, e  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ ,  $\mathbf{a}_3$ , em que  $\mathbf{a}_i = [a_{i1}, a_{i2}, a_{i3}]^\top$  é o vetor de coeficientes correspondente ao  $i$ -ésimo componente principal ( $i = 1, 2$  e  $3$ ). O vetor  $\mathbf{a}_1$  relativo ao primeiro componente principal ( $Y_1$ ), é obtido de tal maneira que maximize sua variância. O vetor do segundo componente principal ( $Y_2$ ), é determinado de forma que detenha a segunda maior variância e o terceiro, e último componente, terá vetor de coeficientes que lhe proporcione a menor variância.

Interpretando esta situação de forma geométrica, Ferreira (2008) diz que, pode-se entender essas combinações lineares como sendo uma rotação no sistema de eixos coordenados, fazendo com que os novos eixos ( $Y_1$ ,  $Y_2$  e  $Y_3$ ) sejam posicionados no sentido de maior variabilidade, ou seja, o primeiro componente principal é responsável pela direção que está associada a maior variabilidade dos dados, e o último componente é responsável pela direção de menor variabilidade. A idéia é maximizar a variabilidade dos componentes principais  $Y_i$ , em relação a  $\mathbf{a}_i$ .

Ainda usando o mesmo exemplo e considerando cada conjunto de coeficientes como um vetor  $\mathbf{a}_i$ , pode-se definir  $Y_i$  como sendo:

$$Y_i = \mathbf{a}_i^\top \mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + a_{i3}X_3,$$

em que a variância de  $Y_i$  é obtida por:

$$\begin{aligned}
\text{Var}(Y_i) &= \text{Var}(\mathbf{a}_i^\top \mathbf{X}) \\
&= \mathbf{a}_i^\top \text{Var}(\mathbf{X}) \mathbf{a}_i \\
&= \mathbf{a}_i^\top \boldsymbol{\Sigma} \mathbf{a}_i.
\end{aligned}$$

Entretanto, o máximo da variância de um componente principal não existe, pois a medida que os componentes do vetor de coeficientes crescem, a variância crescerá para infinito. Desse modo, deve-se impor uma restrição, por exemplo,  $\mathbf{a}_i^\top \mathbf{a}_i = 1$ .

Morrison (1990) usa a técnica de multiplicadores de Lagrange em que deve-se maximizar

$$\mathbf{a}_i^\top \boldsymbol{\Sigma} \mathbf{a}_i - \lambda_i (\mathbf{a}_i^\top \mathbf{a}_i - 1),$$

em relação a  $\mathbf{a}_i$ , sendo  $\lambda_i$  o multiplicador de Lagrange.

A derivada da expressão a ser maximizada em relação a  $\mathbf{a}_i$  igualada a zero, resulta em

$$(\boldsymbol{\Sigma} - \lambda_i \mathbf{I}) \mathbf{a}_i = \mathbf{0}, \quad (2.1)$$

em que  $\mathbf{I}$  é uma matriz identidade de ordem  $p$ ,  $\boldsymbol{\Sigma}$  é a matriz de covariâncias de  $\mathbf{X}$ , também de ordem  $p$  e  $\mathbf{X}$  é a matriz de dados ( $n \times p$ ).

Observando a equação (2.1), nota-se a seguinte relação

$$\boldsymbol{\Sigma} \mathbf{a}_i = \lambda_i \mathbf{a}_i. \quad (2.2)$$

A partir disso pode-se verificar que

$$\text{Var}(Y_i) = \mathbf{a}_i^\top \boldsymbol{\Sigma} \mathbf{a}_i = \mathbf{a}_i^\top \lambda_i \mathbf{a}_i = \lambda_i \mathbf{a}_i^\top \mathbf{a}_i = \lambda_i,$$

e também que a covariância é dada por:

$$\text{Cov}(Y_i, Y_k) = \mathbf{a}_i^\top \boldsymbol{\Sigma} \mathbf{a}_k = \mathbf{a}_i^\top \lambda_k \mathbf{a}_k = \lambda_k \mathbf{a}_i^\top \mathbf{a}_k = 0,$$

para  $i \neq k$ , sendo que  $\mathbf{a}_i$  e  $\mathbf{a}_k$  são ortogonais.

Pela observação da equação (2.2), pode-se concluir que a definição dos componentes principais nada mais é que a obtenção de autovalores e autovetores da matriz de covariância dos dados ( $\boldsymbol{\Sigma}$ ). Sendo que os autovetores  $\mathbf{a}_i$  ( $i = 1, 2, \dots, p$ ) indicam as coordenadas dos novos eixos, ou seja, representam o sentido da rotação dos eixos coordenados definidos pelas variáveis originais. Já os autovalores  $\lambda_i$  ( $i = 1, 2, \dots, p$ ) indicam as variâncias dos novos eixos coordenados (Ferreira, 2008).

Utilizando a decomposição espectral da matriz de covariâncias  $\boldsymbol{\Sigma}$ , dada por

$$\boldsymbol{\Sigma} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^\top,$$

sendo  $\mathbf{P}$  uma matriz composta em suas colunas por autovetores da matriz  $\boldsymbol{\Sigma}$  e  $\boldsymbol{\Lambda}$  a matriz diagonal de autovalores, pode-se verificar que  $\text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^p \lambda_i$ , da seguinte forma:

$$\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^\top) = \text{tr}(\boldsymbol{\Lambda} \mathbf{P}^\top \mathbf{P}) = \text{tr}(\boldsymbol{\Lambda} \mathbf{I}) = \text{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^p \lambda_i.$$

A partir disso é óbvia a conclusão de que  $\sum_{i=1}^p \lambda_i = \sum_{i=1}^p \sigma_{ii}$ , em que  $\sigma_{ii}$

( $i = 1, 2, \dots, p$ ) são as variâncias das  $p$  variáveis, correspondentes aos elementos da diagonal da matriz de covariâncias ( $\Sigma$ ), o que significa que a variabilidade total das variáveis originais é igual a variabilidade dos componentes principais.

Após a determinação dos componentes principais (CP) deve-se definir qual o número ideal a ser usado, de modo a sintetizar a informação contida nas variáveis originais em um modelo mais simplificado. Alguns critérios para essa escolha são propostos na literatura, sendo grande parte deles baseada em argumentos empíricos.

### 2.1.2 Critérios de retenção de componentes principais

Uma das maneiras para determinação do número de componentes principais a serem retidos em um modelo é baseada na porcentagem cumulativa da explicação da variância total. A idéia deste método é determinar um número  $k$  de componentes principais que retenham uma proporção  $\rho_k^2$  da explicação da variância total contida nos dados originais. A dificuldade na aplicação deste método está no fato de que o próprio pesquisador deve determinar quanto o modelo reduzido deve explicar da variação total. Em geral, a proporção ideal para essa explicação, segundo Timm (2002), deve estar entre 70% e 80%, já Jolliffe (2002) afirma estar entre 70% e 90%.

A proporção da variação total explicada pelos  $k$  primeiros componentes principais pode ser estimada por

$$R_k^2 = \frac{\sum_{i=1}^k \hat{\lambda}_i}{\sum_{i=1}^p S_{ii}} \times 100.$$

Além do fato do método da escolha da proporção de explicação utilizando

o estimador  $R_k^2$  ser um dos mais intuitivos, existem situações em que o  $p$  é muito grande, e um ponto de corte de pelo menos 70% não conduz a modelos parcimoniosos devido ao alto valor de  $k$  necessário para alcançar tal ponto de corte.

Uma outra maneira para determinar o número ideal de componentes principais a serem retidos, e que também complementa o método descrito anteriormente, é o uso do gráfico dos autovalores, que é denominado como *scree plot* (Jolliffe, 2002). Neste gráfico são plotados, os autovalores no eixo das ordenadas e a sua ordem  $k$  no eixo das abscissas. Como os primeiros autovalores contém a maior variância e os últimos uma variância menor, a aparência do gráfico é de início uma curva decrescente que depois de um determinado ponto tende a ser paralela ao eixo das abscissas. Sob essas circunstâncias o objetivo é encontrar o ponto, um valor de  $k$ , em que a partir dele a curva torna-se quase paralela ao eixo das abscissas, ou seja, a partir de onde quase não existe variação. Na tentativa de reduzir o caráter pessoal desta técnica, pode-se ajustar modelos de regressão para determinar esse ponto. Timm (2002), indica o modelo de regressão não-paramétrico por *spline*. Modelos de *linear response plateau* podem também atender a esse objetivo (Ferreira, 2008). A desvantagem de determinar um modelo para o gráfico de autovalores é que a precisão do método fica a mercê da qualidade desse ajuste. Fleck & Bourdel (1998), verificaram que esse método retém poucos componentes principais.

Um outro critério para determinar o número de componentes principais a serem retidos é o teste de hipóteses de igualdade dos últimos autovalores (Ferreira, 2008). Neste critério almeja-se testar a hipótese:

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \dots = \lambda_p.$$

Se esta hipótese for rejeitada, deve-se então adicionar um componente

principal ao modelo reduzido escolhido e testar novamente a hipótese nula. Esse processo deve ser repetido até que seja obtido o menor valor de  $k$  para o qual  $H_0$  não seja mais rejeitada. Com isso, almeja-se que os  $k$  primeiros componentes detenham toda a informação sobre a variabilidade e os últimos componentes sejam responsáveis apenas por ruídos.

Para a aplicação deste teste é necessário assumir que a distribuição dos dados seja normal multivariada, porque é um teste baseado na razão de verossimilhanças. A primeira desvantagem deste método é a suposição de normalidade, que é muito forte para o tipo de dado geralmente utilizado nesse tipo de análise. Uma das críticas de Ferreira (2008), em relação a este critério, é o fato de serem realizados vários testes, em número não definido, não independentes e em sequência, impedindo que se faça algum tipo de correção no nível de significância.

No artigo publicado por Perez-Neto et al. (2005), foram comparados alguns métodos, entre eles um baseado em permutação aleatória dos autovalores médios. Essa regra é conduzida da seguinte forma: (1) permutar os valores dentro da matriz de dados das variáveis; (2) conduzir a análise de componentes principais na matriz reformulada; (3) repetir os passos (1) e (2) 999 vezes; (4) para cada análise de componentes da matriz reformulada, calcular o valor médio dos autovalores. Se o valor observado de um dado componente exceder o valor médio calculado, este componente é então considerado diferente de zero. Segundo Perez-Neto et al. (2005), esse é um método recomendável, desde que se tenha aplicado o teste de Bartlett para verificação da independência das variáveis. A estatística desse teste é dada por:

$$\chi^2 = - \left[ n - \frac{1}{6}(2p + 11) \right] \ln|R|, \quad (2.3)$$

sendo  $|R|$  o determinante da matriz de correlação,  $n$  o tamanho da amostra e  $p$  o

número de variáveis. A estatística dada em 2.3 tem distribuição aproximadamente  $\chi^2$  com  $p(p - 1)/2$  graus de liberdade. A hipótese nula desse teste é que todas as variáveis são correlacionadas. Segundo uma comparação feita por Perez-Neto et al. (2005), o teste de Bartlett é uma boa opção.

Outro critério muito utilizado, baseado em matrizes de correlação, é o critério de Kaiser-Guttman (Perez-Neto et al., 2005). Ele propõe considerar apenas os autovalores superiores a um, demonstrando que esses seriam os valores estatisticamente significativos. No entanto, esta condição não é suficiente. Nem todos os autovalores superiores a 1 correspondem a componentes com significado evidente, a ponto de serem retidos. Fleck & Bourdel (1998), em um estudo para comparação de métodos na determinação do número de componentes, salientaram que este método é muito conservador, ou seja, retém componentes demais.

Ainda existem outros métodos disponíveis na literatura para a determinação do número ideal de CP, como por exemplo, o teste de Fujikoshi (1980), que será discutido posteriormente.

Está claro que existe a necessidade de se estudar e construir testes mais adequados para a determinação do número ideal de componentes principais. O primeiro passo para um estudo nesse sentido poderia ser a determinação da distribuição do percentual de explicação dos  $k$  primeiros componentes principais ( $\rho_k^2$ ). Fujikoshi (1980), determinou a distribuição assintótica para o estimador de  $\rho_k^2$  e criou uma estatística para realizar o teste ( $Z_c$ ).

Sobretudo, técnicas de reamostragem e de simulação computacionalmente intensivas são alternativas para construção de testes com esse objetivo.



## 2.2 Simulação *bootstrap*

Dois dos problemas mais importantes em estatística é a determinação do estimador de um parâmetro de interesse e a avaliação de sua precisão. Efron (1979a), ao introduzir sua versão de *bootstrap*, foi particularmente motivado por estes dois problemas. Por causa de sua generalidade, técnicas de *bootstrap* têm sido aplicadas a uma classe de problemas muito maior do que apenas a estimativa de erros padrões e intervalos de confiança. Em várias disciplinas, como psicologia, geologia, econometria, biologia, engenharia, química, etc, empregam-se essas técnicas (Chernick, 2008).

Esse método tem por alicerce o fato de que o pesquisador pode tratar sua amostra como se ela fosse a população que deu origem aos dados e usar amostragem com reposição da amostra original para gerar pseudo-amostras. A partir destas pseudo-amostras, seria possível estimar parâmetros da população. De maneira mais geral, seria determinar uma distribuição empírica (ou de *bootstrap*) para o parâmetro de interesse a partir da amostra inicial de tamanho  $n$ , considerando que cada unidade amostral tem probabilidade igual a  $1/n$  de ser sorteado. A idéia é simplesmente substituir a distribuição da população desconhecida pela distribuição conhecida empírica, possibilitando assim estimar o parâmetro, determinar seu desvio-padrão, intervalo de confiança ou até mesmo testar uma hipótese sobre ele.

Como já é sabido, o método de simulação *bootstrap* é baseado em reamostragens. Segundo Chernick (2008), o número de reamostragens necessárias para se obter boas estimativas, em intervalos de confiança e problemas de testes de hipóteses, seria de pelo menos 1000 repetições *bootstrap*. Para se fazer estas repetições, existem duas maneiras distintas: o paramétrico e o não-paramétrico.

Quando existe um modelo matemático particular, com constantes ajustá-

veis ou parâmetros que determinam totalmente a função de densidade ou de probabilidade, tal modelo é chamado de paramétrico. Métodos estatísticos baseados neste modelo são métodos paramétricos. Quando nenhum modelo matemático é usado, a análise estatística é não-paramétrica, e usa apenas o fato de que as variáveis aleatórias  $Y_j$  são independentes e identicamente distribuídas. Mesmo que exista um modelo paramétrico plausível para a situação, uma análise não paramétrica pode ser usada para avaliar a robustez das conclusões de uma análise paramétrica (Davison & Hinkley, 2008).

### 2.2.1 *Bootstrap* paramétrico

Neste método as reamostragens são feitas a partir de uma função de distribuição conhecida  $F$ . Quando os dados são originários de uma função de distribuição paramétrica, então  $X_1, X_2, \dots, X_n \sim i.i.d. F(x, \theta)$ . Supondo que os dados observados são usados para estimar  $\theta$  por  $\hat{\theta}$ , que frequentemente, mas nem sempre, é o estimador de máxima verossimilhança. Assim, sua substituição no modelo fornece o modelo ajustado,  $X_1^*, X_2^*, \dots, X_n^* \sim i.i.d. F(x, \hat{\theta})$ , o qual pode ser usado para calcular estimativas de parâmetros (Davison & Hinkley, 2008). Quando o modelo é conhecido ou acredita-se ser uma boa representação da realidade, o *bootstrap* paramétrico pode ser uma ferramenta poderosa, permitindo fazer inferências em situações que de outra forma são intratáveis (Givens & Hoeting, 2005).

Em um procedimento de *bootstrap* paramétrico, o primeiro passo seria obter estimativas dos parâmetros que caracterizam a distribuição em questão. A partir desses parâmetros geram-se  $B$  amostras aleatórias, cada uma de tamanho  $n$ , com a distribuição conhecida e para cada amostra calcula-se uma estimativa da estatística de interesse, que tem a mesma forma funcional da estatística original.

Com essa distribuição de *bootstrap* é possível fazer inferências sobre o parâmetro de interesse (Gentle, 2003).

### 2.2.2 *Bootstrap* não-paramétrico

Uma outra maneira de se obter amostras *bootstrap* é o procedimento de reamostragem não-paramétrico. Ao contrário do método paramétrico, nesse caso não são feitas pressuposições sobre a distribuição dos dados, ou seja, o conjunto de dados tem função de distribuição desconhecida.

As  $B$  pseudo-amostras, de tamanho  $n$ , são obtidas através de reamostragem com reposição da amostra original. A probabilidade de selecionar qualquer valor do conjunto de dados é igual a  $1/n$ , pois considera-se que a amostra original é independente e identicamente distribuída com uma função de distribuição desconhecida. O passo seguinte é calcular a estimativa do parâmetro de interesse para cada uma das  $B$  pseudo-amostras, assim tem-se a distribuição de *bootstrap* desse parâmetro, a partir da qual pode-se fazer inferências.

Em casos onde os dados originais são correlacionados, ou seja, não são independentes e identicamente distribuídos, a aproximação da função de distribuição de *bootstrap* ( $\hat{F}^*(x)$ ) em relação a distribuição verdadeira ( $F(x)$ ) não é segura. Métodos de *bootstrap* para dados dependentes são muito bem descritos em Efron & Tibshirani (1993) e Kunsch (1989).

## 2.3 Simulação Monte Carlo

O método Monte Carlo herdou seu nome do mais famoso cassino da Europa, localizado em Monaco. Distribuições de probabilidade envolvendo jogos de azar são frequentemente muito complicadas, mas podem ser avaliadas via simulações. Simulações computacionais referem-se a reprodução de processos, com

um programa computacional adequado, e a observação desses resultados. Métodos Monte Carlo são aqueles baseados em simulações computacionais envolvendo números aleatórios (Baron, 2007).

O método Monte Carlo pode ser definido como a representação da solução de um problema, sendo utilizada sequência de números aleatórios para se fazer essa representação, com o objetivo de criar uma amostra da população. Sendo assim, possível o cálculo de estimativas das estatísticas de parâmetros representativos do problema (Checchetti, 1999).

Segundo Dachs (1988), os métodos Monte Carlo são frequentemente utilizados para produzir amostras de acordo com distribuições teóricas conhecidas, objetivando avaliar novos comportamentos de diferentes técnicas estatísticas que podem ser empregadas em determinado problema. Nesses casos, esse método pode ser a única maneira satisfatória de prover as informações necessárias.

O primeiro passo para aplicação desse método seria a simulação de amostras, de tamanho  $n$ , seguindo distribuições previamente estruturadas, que serão avaliadas pelas novas técnicas estatísticas em estudo, considerando a hipótese nula estabelecida. Com as técnicas computacionais disponíveis, torna-se fácil a simulação de realizações de variáveis ou de amostras baseadas em modelos estatísticos apropriados de parâmetros conhecidos, com a finalidade de se verificar a adequação de determinada metodologia ou na realização de comparações entre métodos (Dachs, 1988).

No estudo do desempenho ou avaliação de testes estatísticos, frequentemente torna-se muito complicado obter analiticamente informações sobre taxas de erro tipo I e poder. Uma maneira de obter as informações desejadas de forma eficiente e simples, é o uso do método de simulação Monte Carlo, evitando assim eventuais dificuldades analíticas.

Vários autores já avaliaram o desempenho de testes estatísticos através de simulação, como por exemplo Checchetti (1999) que estudou quatro critérios multivariados para o teste de igualdade de efeitos de tratamentos, Silva (1998) que estudou o desempenho do teste de Scott-Knott, Cirillo (2003) que analisou o problema de Behrens-Fisher no espaço p-dimensional, Amorim et al. (2010) na avaliação de testes para a unidimensionalidade de painéis sensoriais entre outros.

#### **2.4 Erro tipo I, erro tipo II, poder e robustez**

De forma geral, o pesquisador depara-se muitas vezes com a necessidade de testar hipóteses sobre parâmetros populacionais. Quando uma hipótese nula é testada tem-se apenas duas alternativas de decisão, rejeitar ou não rejeitar  $H_0$ . Observa-se que ao tomar essa decisão, de rejeição ou não da hipótese nula, é possível estar incorrendo em erros, mesmo quando tudo é feito corretamente. Logo, é necessário considerá-los. O primeiro deles chama-se erro tipo I, e consiste em rejeitar a hipótese nula quando de fato ela é verdadeira. O segundo erro é aquele que é cometido quando  $H_0$  não é rejeitada e na verdade ela é falsa, esse é o chamado erro tipo II (Mood et al., 1974).

Em geral, a probabilidade de cometer o erro tipo I é representada pela letra grega  $\alpha$ , sendo habitualmente chamado de nível de significância. A probabilidade de que se cometa o erro tipo II é simbolizada pela letra  $\beta$  e não tem nome especial. No entanto, quando se aceita  $H_0$  e ela é verdadeira, ou quando se rejeita  $H_0$  e ela é falsa isso consiste em uma decisão correta. Esta última decisão tem probabilidade de ocorrência  $1 - \beta$ , pois é o complemento do erro tipo II ( $\beta$ ), e normalmente é chamado de poder do teste. O complemento da probabilidade de ocorrência do erro tipo I ( $\alpha$ ), que consiste em aceitar  $H_0$  sendo ela ser verdadeira, é representada por  $1 - \alpha$  e comumente chamado de coeficiente de confiança (Oliveira et al., 2009).

Esses conceitos estão sintetizados na Tabela 1.

TABELA 1 Resultados possíveis de um teste de hipóteses e suas probabilidades de ocorrência.

DECISÃO TOMADA	REALIDADE	
	<i>H<sub>0</sub> é verdadeira</i>	<i>H<sub>0</sub> é falsa</i>
<i>H<sub>0</sub> é aceita</i>	Decisão correta Probabilidade : $1 - \alpha$ (Coeficiente de confiança)	Decisão errada Probabilidade : $\beta$ (Erro Tipo II)
<i>H<sub>0</sub> é rejeitada</i>	Decisão errada Probabilidade : $\alpha$ (Erro Tipo I)	Decisão correta Probabilidade : $1 - \beta$ (Poder do teste)
Probabilidade total	$(1 - \alpha) + \alpha = 1$	$(1 - \beta) + \beta = 1$

Sabe-se que não é possível evitar esses erros quando se aplica um teste de hipóteses, mas deve-se manter relativamente pequena a probabilidade de cometê-los. O pesquisador pode controlar diretamente apenas o erro tipo I, pois ele deve ser previamente fixado. Como  $\alpha$  e  $\beta$  são inversamente proporcionais, a única maneira de causar redução simultânea dos dois erros é aumentando o tamanho da amostra. Devido a essa relação inversa, a fixação de  $\alpha$  entre 0,10 e 0,01 é, sempre que possível, uma boa medida (Ferreira, 2005). Testes que apresentam um valor real de taxas de erro tipo I menor que o nominal  $\alpha$  são ditos conservativos ou rigorosos, ou seja, sob  $H_0$  a probabilidade de se encontrar um resultado significativo é inferior ao valor  $\alpha$  estabelecido. No entanto, quando testes trazem um valor real de taxa de erro tipo I maior que o nominal, estes são ditos liberais, pois a probabilidade de detectar resultado significativo é maior que o nível de significância estabelecido.

No desenvolvimento de novos testes, a avaliação do poder é um critério largamente empregado para a escolha, ou não, desse teste. Entretanto, o poder de

um teste não é simples de ser calculado em situações reais, pois depende do valor desconhecido do parâmetro em questão.

De acordo com Ramalho et al. (2000), poder de um teste é a habilidade que ele tem em detectar todas as reais diferenças entre tratamentos, ou seja, é a probabilidade  $(1 - \beta)$  de rejeitar  $H_0$  quando ela de fato, é falsa. Costa (1992) define poder como sendo “a sensibilidade da região crítica” para perceber e rejeitar uma hipótese falsa.

Na Figura 1 tem-se um exemplo de curva da função poder em relação a um parâmetro  $\theta$  com valor hipotético igual a 10. No eixo das ordenadas estão os valores da função poder e na abscissa os valores do parâmetro. Observa-se, que a medida em que os valores de  $\theta$  se afastam do valor hipotético, o poder aumenta, pois quanto mais distante o parâmetro estiver do valor hipotético mais fácil é detectar essa diferença.

Um bom teste é aquele que aceita  $H_0$  com alta probabilidade quando realmente ela é verdadeira, e que rejeita a hipótese nula com grande probabilidade quando de fato ela é falsa. Para qualificar um bom teste é também necessário avaliar a sua robustez.

Quando se aplica um teste estatístico, normalmente algumas condições devem ser satisfeitas. Um exemplo, muito comum, é a análise de variância, a qual tem as seguintes exigências (Banzatto & Kronka, 2006):

- Os efeitos dos fatores que se encontram no modelo matemático de um dado experimento devem ser aditivos;
- Os erros experimentais devem ser aleatórios, independentes e normalmente distribuídos com média zero e variância comum.

Um procedimento deve ser considerado robusto, se ao violar uma das pres-

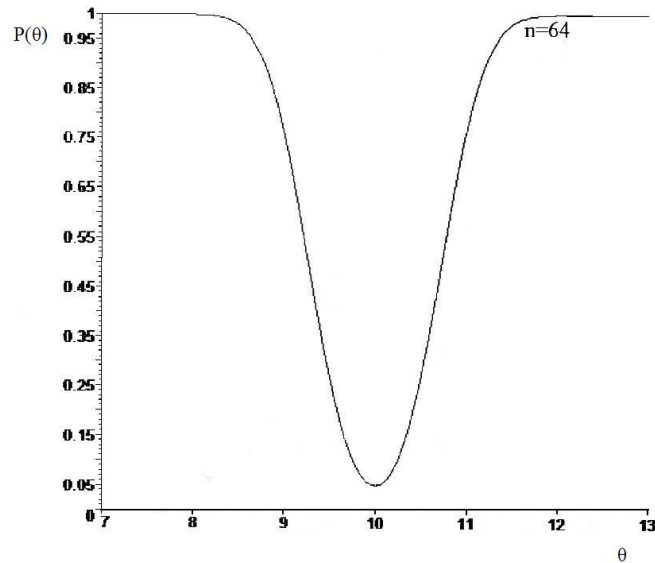


FIGURA 1 Curva da função poder em relação a um parâmetro  $\theta$

suposições estabelecidas, este procedimento, pelo menos de forma aproximada, mantenha as características delineadas originalmente na elaboração de sua teoria.

Um exemplo para a avaliação da robustez seria aplicar um teste baseado na distribuição normal (exige que os dados sigam esta distribuição), em dados com distribuição  $t$  de Student, por exemplo. Deve-se repetir esse processo várias vezes e avaliar se o comportamento é similar ou não quando os dados são normais. Quanto mais o desempenho se aproximar dos resultados obtidos quando as pressuposições são atendidas, mais robusto será o teste.

## 2.5 Distribuições

Nesta subseção estão descritas as distribuições Normal e  $t$  de Student Multivariadas, que estão envolvidas na construção e avaliação do desempenho dos



testes propostos neste trabalho.

### 2.5.1 Distribuição normal multivariada

A distribuição normal é a distribuição de uma variável aleatória contínua mais importante no âmbito univariado. Na estatística multivariada não poderia ser diferente, sendo ela muito utilizada em pesquisas científicas.

Muitas técnicas empregadas na estatística multivariada são baseadas nessa distribuição (Rencher, 2002). Inferências são realizadas assumindo normalidade dos resíduos dos modelos adotados, como por exemplo a análise de variância multivariada (MANOVA).

Um vetor de dados  $p$ -dimensional é dito ter distribuição normal multivariada  $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , com parâmetros  $\boldsymbol{\mu}$  e  $\boldsymbol{\Sigma}$ , se sua função densidade de probabilidade é dada por:

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

sendo  $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_p]^{\top}$  o vetor de médias e  $\boldsymbol{\Sigma}_{(p \times p)}$  a matriz de covariâncias dada por:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix},$$

sendo  $\boldsymbol{\Sigma}$  uma matriz positiva definida (Anderson, 1971).

Em muitos trabalhos científicos o pesquisador se depara com a necessidade de simular dados de alguma distribuição, geralmente com o intuito de validar modelos, comparar testes de hipóteses ou métodos de estimação. Para a geração de uma amostra aleatória de tamanho  $n$  com distribuição normal  $p$ -dimensional

$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , pode-se utilizar o teorema fundamental da transformação de probabilidade (Magalhães, 2006), onde a partir de variáveis uniformes pode-se obter variáveis de qualquer outra distribuição, basta inverter a função de distribuição da variável que se deseja gerar.

Como exemplo, suponha que se deseja gerar uma variável aleatória normal-padrão. Primeiramente deve-se gerar uma realização  $u$  de uma variável aleatória uniforme  $U(0, 1)$ , em seguida obter  $z$  de forma que:

$$u = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy,$$

ou seja,  $z = \Phi^{-1}(u)$ . Deve-se repetir esse processo até se obter  $p$  elementos do vetor  $\mathbf{Z}$ .

A partir disso, para gerar uma variável aleatória com distribuição  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , deve-se aplicar a seguinte transformação linear no vetor  $\mathbf{Z}$ ,  $\mathbf{X} = \boldsymbol{\Sigma}^{1/2} \mathbf{Z} + \boldsymbol{\mu}$ , em que  $\boldsymbol{\Sigma}^{1/2}$  é a matriz raiz quadrada de  $\boldsymbol{\Sigma}$ . Segundo Ferreira (2008)  $\mathbf{X}$  possui distribuição normal multivariada com média  $\boldsymbol{\mu}$  e matriz de covariâncias  $\boldsymbol{\Sigma}$ . Esse processo deve ser repetido até se obter o tamanho de amostra  $n$  desejado.

### 2.5.2 Distribuição $t$ de Student multivariada

Segundo Sorensen & Gianola (2002), um vetor aleatório  $p$ -dimensional  $\mathbf{X} = [X_1, X_2, \dots, X_p]^\top$  é dito ter distribuição  $t$  de Student  $p$ -variada com  $\nu$  graus de liberdade, vetor de médias  $\boldsymbol{\mu}$ , e matrix de covariâncias  $\boldsymbol{\Sigma}$ , se sua função densidade de probabilidade é dada por:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\Gamma(\frac{p+\nu}{2})}{\Gamma(\frac{\nu}{2})|\nu\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left[ 1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\nu} \right]^{-\frac{p+\nu}{2}}.$$

Essa distribuição é dita centrada se o vetor de médias  $\boldsymbol{\mu}$  for igual a zero, caso contrário ela é dita não central.

Note que  $\nu$  pode assumir qualquer valor positivo real e não necessariamente precisa ser um número inteiro. Quando  $p = 1$ , a função resulta na distribuição  $t$ -univariada.

Para a geração de realizações de variáveis aleatórias  $p$ -dimensionais  $t$  de Student com  $\nu$  graus de liberdade e parâmetros  $\boldsymbol{\mu}$  e  $\boldsymbol{\Sigma}$ , considere um vetor aleatório  $\mathbf{Z}$  com distribuição  $N_p(\mathbf{0}, \mathbf{I})$  e uma variável aleatória  $Q$  com distribuição qui-quadrado com  $\nu$  graus de liberdade.

O vetor aleatório  $\mathbf{Y}$  dado pela transformação

$$\mathbf{Y} = \sqrt{\nu} \frac{\mathbf{Z}}{\sqrt{Q}},$$

tem distribuição  $t$  multivariada esférica com  $\nu$  graus de liberdade.

O vetor aleatório  $\mathbf{X}$  obtido por meio da transformação

$$\mathbf{X} = \boldsymbol{\Sigma}^{1/2} \mathbf{Y} + \boldsymbol{\mu},$$

possui distribuição  $t$  multivariada elíptica com  $\nu$  graus de liberdade e parâmetros  $\boldsymbol{\mu}$  e  $\boldsymbol{\Sigma}$  (Ferreira, 2008).

Deve-se repetir os processos anteriores  $n$  vezes a fim de se obter amostras aleatórias de tamanho  $n$  com a distribuição almejada. A geração de variáveis aleatórias com distribuição qui-quadrado pode ser feita a partir do teorema fundamental da transformação de probabilidades mencionado anteriormente.

### 3 METODOLOGIA E VALIDAÇÃO MONTE CARLO

Neste trabalho foram propostos três métodos baseados em estatística computacional intensiva, *bootstrap*, para testar as seguintes hipóteses:

$$\begin{cases} H_0 : \rho_k^2 \geq \rho_{k0}^2 \\ H_1 : \rho_k^2 < \rho_{k0}^2, \end{cases} \quad (3.1)$$

em que  $0 \leq \rho_{k0}^2 \leq 1$  é um valor hipotético, relativo à proporção de variação total explicada pelos  $k$  primeiros componentes principais,  $k \leq p$ ,  $p$  o número de variáveis e  $\rho_k^2$  a proporção de variação total explicada pelos  $k$  primeiros componentes principais populacionais, isto é,

$$\rho_k^2 = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i},$$

sendo  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ , os autovalores da matriz de covariância populacional  $\Sigma$  de uma variável aleatória  $\mathbf{Y} \in \mathbb{R}^p$  de uma distribuição qualquer com média  $\mu$ .

O estimador da proporção de explicação da variação total pelos  $k$  primeiros autovalores  $\rho_k^2$  é dado por:

$$R_k^2 = \frac{\sum_{i=1}^k \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i}, \quad (3.2)$$

em que  $\hat{\lambda}_i$  são os autovalores,  $i = 1, 2, \dots, p$ , da matriz de covariâncias amostral  $\mathbf{S}$ , dada por:

$$\mathbf{S} = \frac{1}{n-1} \left[ \sum_{j=1}^n \mathbf{Y}_j \mathbf{Y}_j^\top - \frac{\left( \sum_{j=1}^n \mathbf{Y}_j \right) \left( \sum_{j=1}^n \mathbf{Y}_j \right)^\top}{n} \right], \quad (3.3)$$

sendo  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_j, \dots, \mathbf{Y}_n$ , uma amostra aleatória de uma população qualquer com média  $\boldsymbol{\mu}$  e covariância  $\boldsymbol{\Sigma}$ , considerando  $\mathbf{Y}_j \in \mathbb{R}^p$ .

Além dos três testes *bootstrap* propostos, também foi utilizado o teste de Fujikoshi (1980), para fins de comparação. Esse teste supõe que a distribuição de  $\mathbf{Y}$  é normal multivariada, além de ter a distribuição da estatística do teste como normal padrão assintótica. Utilizando o resultado,  $\sqrt{n}(R_k^2 - \rho_k^2) \sim N(0, \omega^2)$ , apresentado por Fujikoshi (1980), em que

$$\omega^2 = \frac{2\text{tr}(\boldsymbol{\Sigma}^2)}{[\text{tr}(\boldsymbol{\Sigma})]^2} [(\rho_k^2)^2 - 2\beta\rho_k^2 + \beta],$$

e

$$\beta = \frac{\sum_{i=1}^k \lambda_i^2}{\sum_{i=1}^p \lambda_i^2},$$

então a estatística do teste de Fujikoshi, sob  $H_0$ , dada por

$$Z_c = \frac{R_k^2 - \rho_{k0}^2}{\frac{\omega_0}{\sqrt{n}}}, \quad (3.4)$$

tem distribuição assintótica normal padrão.

Utilizando-se estimadores para  $\omega$  e  $\beta$ , que são respectivamente dados por:

$$\omega_0 = \sqrt{\frac{2\text{tr}(S^2)}{[\text{tr}(S)]^2} [(\rho_{k0}^2)^2 - 2\hat{\beta}\rho_{k0}^2 + \hat{\beta}]}$$

e

$$\hat{\beta} = \frac{\sum_{i=1}^k \hat{\lambda}_i^2}{\sum_{i=1}^p \hat{\lambda}_i^2},$$

o teste em questão pôde ser aplicado. A rotina com os comandos do software R (R Development Core Team, 2009) para aplicação deste teste está apresentada no Anexo B.

Os métodos *bootstrap* considerados nesse trabalho foram delineados de duas formas diferentes. Foram propostos dois métodos de *bootstrap* paramétrico e um não-paramétrico para testar a hipótese 3.1 de interesse. Os resultados metodológicos serão apresentados na seção Resultados e Discussão. Para avaliar o desempenho dos três testes propostos e o teste assintótico de Fujikoshi (1980) foram realizadas simulações Monte Carlo.

Foram consideradas as taxas de erro tipo I e poder na avaliação do desempenho dos quatro testes. Para isso foram simuladas amostras aleatórias no espaço  $p$ -dimensional de tamanho  $n$  dados por  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ , em que  $\mathbf{Y}_j \in \mathbb{R}^p$ , com  $j = 1, 2, \dots, n$ . Essas amostras foram simuladas de distribuições com média  $\boldsymbol{\mu}$  e covariância  $\boldsymbol{\Sigma}_{(p \times p)}$  positiva definida. Foi enfatizado o uso da distribuição normal  $p$ -variada nessas simulações, entretanto, também foram consideradas amostras da distribuição  $t$  de Student multivariada para avaliar a robustez dos testes. Neste trabalho as simulações de realizações de variáveis aleatórias foram feitas por meio do software R Development Core Team (2009).

Para a realização das simulações Monte Carlo foram geradas amostras dessas duas distribuições considerando números de variáveis  $p = 2, 3$  e  $10$  e tamanhos amostrais  $n = 15, 30$  e  $100$ . Também foram considerados diferentes valores para a proporção de explicação dos  $k$  primeiros componentes sob a hipótese nula,  $\rho_{k0}^2 = 0,70$  e  $0,90$ . Decidiu-se pelo uso desses valores por indicação dos autores Jolliffe (2002) e Timm (2002), não optando-se por  $0,80$  pelo fato de ser um valor muito próximo à  $0,70$  e  $0,90$ . Diferentes valores para o número de componentes principais foram considerados,  $k = 1$  e  $2$ , que explicariam o valor mínimo  $\rho_{k0}^2$ , da variação total contida nas  $p$  variáveis originais,  $k \leq p$ . Todos os testes foram aplicados a cada uma dessas amostras em cada uma das configurações formadas pela combinação dos valores de  $p, n, \rho_{k0}^2$  e  $k$ , considerando os níveis de significância ( $\alpha$ ) de  $0,10, 0,05$  e  $0,01$ . Em todos os casos foram simuladas  $N = 1000$  amostras de Monte Carlo e as proporções de rejeição de  $H_0$  foram registradas para cada teste.

Foram simuladas amostras de populações multivariadas com parâmetro  $\Sigma$  determinado de tal forma que  $\rho_k^2$  variasse no intervalo entre  $0$  e  $1$ . Assim, quando a amostra gerada for proveniente de uma população em que  $\rho_k^2 < \rho_{k0}^2$ , a proporção de rejeições de  $H_0$  em  $N$  simulações Monte Carlo é uma estimativa de poder. Por outro lado, se a amostra gerada for proveniente de uma população em que  $\rho_k^2 \geq \rho_{k0}^2$ , a hipótese nula 3.1 é verdadeira e a proporção de rejeições de  $H_0$  é uma estimativa do erro tipo I.

Sem perda de generalidade, fixou-se a estrutura de  $\Sigma$  por

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix},$$

em que  $\sigma^2$  foi também fixado em 1 e  $\rho$  determinado de tal forma que  $\rho_k^2$  fosse igual ao valor previamente escolhido. Para matrizes com essa estrutura os autovalores são:

$$\lambda_1 = \sigma^2[1 + (p - 1)\rho] \quad (3.5)$$

$$\lambda_i = \sigma^2(1 - \rho) \quad (3.6)$$

em que  $i = 2, 3, \dots, p$ .

Para que  $\Sigma$  seja uma matriz positiva definida, é necessário que todos os seus autovalores sejam maiores do que zero. Observando a equação 3.6 nota-se que  $\lambda_i$  sempre será positivo, pois  $\rho$  varia entre  $-1$  e  $1$ . Já na equação 3.5 podem ocorrer valores negativos. Assim, fazendo  $\lambda_1 > 0$  tem-se:

$$\rho > \frac{-1}{(p - 1)}. \quad (3.7)$$

Também  $\rho_k^2$  deve ser um valor entre 0 e 1. Assim, utilizando-se 3.5 e 3.6 obtém-se:

$$\rho_k^2 = \frac{\lambda_1 + \sum_{i=2}^k \lambda_i}{p} = \frac{k + (p - k)\rho}{p},$$

como também é necessário que  $\rho_k^2$  seja um valor maior que zero, tem-se:

$$\rho > \frac{-k}{(p - k)}. \quad (3.8)$$

Se  $\rho$  for isolado e expresso em função de  $\rho_k^2$  chega-se a:

$$\rho = \frac{p\rho_k^2 - k}{p - k}. \quad (3.9)$$



Dependendo do valor fixado para  $p$  em 3.5 e usando as restrições 3.7 e 3.8, ainda podem ocorrer valores negativos para o primeiro autovalor ( $\lambda_1$ ). A fim de evitar esses valores negativos é necessário impor a condição  $\rho > 0$ . Dentre as três condições necessárias para que  $\Sigma$  seja positiva definida, a mais forte é  $\rho > 0$ , então usando a expressão 3.9 tem-se:

$$\rho_k^2 > \frac{k}{p}. \quad (3.10)$$

Desse modo, imposta a condição 3.10 para os valores de  $\rho_k^2$ , os valores de  $\rho$  foram determinados por 3.9. Construída a matriz  $\Sigma$ , amostras aleatórias  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  em que  $\mathbf{Y}_j \in \mathbb{R}^p$  foram geradas da distribuição normal  $p$ -variada por

$$\mathbf{Y}_j = \Sigma^{1/2} \mathbf{Z}_j + \boldsymbol{\mu} \quad (3.11)$$

em que  $\Sigma^{1/2}$  é a matriz raiz quadrada de  $\Sigma$ ,  $\boldsymbol{\mu} \in \mathbb{R}^p$  foi fixada como o vetor nulo  $\mathbf{0}_{(p \times 1)}$  e  $\mathbf{Z}_j \in \mathbb{R}^p$ , é um vetor de variáveis aleatórias normais padrão independentemente distribuídas, geradas pelo método polar Box-Müller (Press et al., 1992), para  $j = 1, 2, \dots, n$ .

Simuladas as amostras de diferentes tamanhos  $n$ , diferentes dimensões  $p$  e diferentes números de componentes principais  $k$  explicando a proporção da variação total  $\rho_k^2$ , foram aplicados os quatro testes e computadas as proporções de rejeições da hipótese nula ao longo das  $N = 1000$  simulações Monte Carlo para um determinado nível de significância ( $\alpha$ ) dado por

$$\% \text{ rejeições} = \frac{\sum_{\ell=1}^N I(\text{valor-p}_\ell \leq \alpha)}{N} \quad (3.12)$$

em que  $\text{valor-p}_\ell$  é o valor-p da  $\ell$ -ésima amostra Monte Carlo simulada e  $I(\text{valor-p}_\ell \leq \alpha)$  é a função indicadora que retorna 1 se a condição for verdadeira ou 0 se for falsa na  $\ell$ -ésima simulação. Os valores de  $n$  utilizados foram 15, 30 e 100, foram adotados para  $p = 2$  e 10 e  $k = 1$  e 2.

As taxas de rejeições 3.12 foram computadas para todos os testes em todas as configurações.

## 4 RESULTADOS E DISCUSSÃO

Esta seção será dividida em duas subseções. Na primeira serão apresentados os testes propostos neste trabalho envolvendo métodos computacionalmente intensivos. Na segunda, os resultados referentes a avaliação Monte Carlo do desempenho dos mesmos.

### 4.1 Resultados metodológicos

Nesta seção os três métodos computacionalmente intensivos, que são a contribuição deste trabalho, estão apresentados. Foram descritas todas as etapas necessárias para a implementação de cada um deles. Em todos os casos foi considerada uma amostra aleatória  $p$ -dimensional de tamanho  $n$ , dada por  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ .

Foram propostos dois testes para hipótese 3.1 baseados em *bootstrap* paramétrico e um não paramétrico.

Os três testes apresentaram um excelente desempenho em relação a velocidade de processamento. Usando um computador da marca Acer modelo Aspire 5610, com processador Intel Core Duo, foram gastos em média dois segundos para o processamento de um único teste com 1000 reamostragens *bootstrap*.

#### 4.1.1 Testes de *bootstrap* paramétrico

A partir da amostra aleatória original obtida em  $\mathbb{R}^p$  é estimada a matriz de covariâncias  $\Sigma$  utilizando a expressão 3.3. Usando algum método numérico são obtidos os autovalores de  $\mathbf{S}$  dados por  $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$ . Nesse trabalho utilizou-se a função *eigen* do programa R (R Development Core Team, 2009). Em seguida utilizando-se a expressão 3.2, foi calculado o valor de  $R_k^2$ .

Para a construção dos testes de *bootstrap* paramétrico utilizou-se a distribuição normal  $p$ -variada como referência. A média  $\boldsymbol{\mu}_b$  dessa distribuição foi considerada, sem perda de generalidade, igual ao vetor nulo  $\mathbf{0} = [0 \dots 0]^\top$ . A definição da matriz de covariâncias dessa distribuição  $\boldsymbol{\Sigma}_b$  foi condicionada à imposição da hipótese nula 3.1. Assim,  $B - 1$  amostras de *bootstrap* de tamanho  $n$  no espaço  $p$ -dimensional foram retiradas dessa distribuição. Essas amostras foram retiradas de uma população em que  $\rho_k^2 = \rho_{k0}^2$ , ou seja, estavam sob  $H_0$ .

O parâmetro de covariação da distribuição de *bootstrap*, a matriz  $\boldsymbol{\Sigma}_b$ , foi construído de forma a atender a condição imposta pela hipótese nula 3.1. Inicialmente, considerou-se a decomposição espectral de  $\boldsymbol{\Sigma}_b$  por

$$\boldsymbol{\Sigma}_b = \hat{\boldsymbol{P}} \boldsymbol{\Lambda}_b \hat{\boldsymbol{P}}^\top, \quad (4.1)$$

em que  $\hat{\boldsymbol{P}}$  é a matriz de autovetores de  $\boldsymbol{S}$  e  $\boldsymbol{\Lambda}_b$  uma matriz diagonal, definida de tal forma que a proporção explicada pelos  $k$  primeiros autovalores seja igual a  $\rho_{k0}^2$ .

Seja  $\hat{V}_t = \sum_{i=1}^p \lambda_i$  a variação total contida em  $\boldsymbol{S}$ , então  $\rho_{k0}^2 \hat{V}_t$  representa a proporção que deve ser explicada pelos  $k$  primeiros autovalores. Como  $\sum_{i=1}^k \hat{\lambda}_i < \rho_{k0}^2 \hat{V}_t$ , a quantidade  $\hat{\Delta} = \rho_{k0}^2 \hat{V}_t - \sum_{i=1}^k \hat{\lambda}_i$  foi retirada dos últimos  $p - k$  autovalores de forma proporcional, e adicionada de forma também proporcional aos  $k$  primeiros autovalores. Assim, para os  $k$  últimos autovalores tem-se:

$$\lambda_{ib} = \hat{\lambda}_i - \frac{\hat{\lambda}_i}{\sum_{i=k+1}^p \hat{\lambda}_i} \cdot \hat{\Delta},$$

para  $i = k + 1, k + 2, \dots, p$ , em que  $\hat{\lambda}_{ib}$  é o  $i$ -ésimo elemento da diagonal de  $\boldsymbol{\Lambda}_b$  e  $\hat{\lambda}_i$  é o autovalor de  $\boldsymbol{S}$ .

Os  $k$  primeiros autovalores de  $\boldsymbol{\Lambda}_b$  foram definidos por:

$$\lambda_{ib} = \hat{\lambda}_i + \frac{\hat{\lambda}_i}{\sum_{i=1}^k \hat{\lambda}_i} \cdot \hat{\Delta},$$

para  $i = 1, 2, \dots, k$ . Ver exemplo numérico no Anexo A.

A matriz  $\Sigma_b$  pôde ser construída utilizando-se a expressão 4.1. Definindo os parâmetros da distribuição normal  $p$ -variada  $\mu$  e  $\Sigma_b$ , foram gerados  $B - 1$  amostras aleatórias de tamanho  $n$  no espaço  $p$ -dimensional dessa distribuição. Cada amostra gerada estava sob a hipótese nula 3.1. Considerando a  $\ell$ -ésima amostragem de *bootstrap*  $\mathbf{Y}_{1\ell}^*, \mathbf{Y}_{2\ell}^*, \dots, \mathbf{Y}_{j\ell}^*, \dots, \mathbf{Y}_{n\ell}^*$ , em que  $\mathbf{Y}_{j\ell}^* \in \mathbb{R}^p$ , e  $\ell = 1, 2, \dots, n$ , foram definidas as seguintes quantidades, que possibilitaram a construção dos dois diferentes testes propostos.

A primeira delas foi definida por

$$R_{k\ell}^2 = \frac{\sum_{i=1}^k \lambda_i^*}{\sum_{i=1}^p \lambda_i^*}, \quad (4.2)$$

em que  $\lambda_i^*, i = 1, 2, \dots, p$  são os autovalores da matriz de covariâncias  $\mathbf{S}_\ell^*$  calculada a partir da  $\ell$ -ésima amostra de *bootstrap*.

Os valores  $R_{k\ell}^2$  ao longo das  $B - 1$ , sendo  $B = 1000$ , amostras de *bootstrap* foram computados e referem-se a distribuição nula de *bootstrap*. Assim, o valor-p de um teste *bootstrap* paramétrico baseado nessa distribuição pode ser definido por

$$\text{valor-p} = \frac{\sum_{\ell=1}^B I(R_{k\ell}^2 \leq R_k^2)}{B}, \quad (4.3)$$

ou seja, será dado pela proporção de valores da explicação dos  $k$  primeiros autovalores da distribuição nula de *bootstrap* paramétrico que foram inferiores ao valor obtido na amostra original ( $R_k^2$ ). Esse valor deve ser confrontado com o nível nominal de significância  $\alpha$  para decisão a respeito da rejeição ou não da hipótese nula 3.1. Esse teste foi nomeado com a sigla *BPRk* (Teste *bootstrap* paramétrico baseado em  $R_k^2$ ) e a rotina com os comandos do software R Development Core Team (2009) para sua aplicação está apresentada no Anexo C.

A segunda estatística, para a definição do segundo teste, inspirada pelo trabalho de Fujikoshi (1980), foi definida por:

$$Z_{cl} = \frac{R_{k\ell}^2 - \rho_{k0}^2}{\frac{\omega_{0\ell}}{\sqrt{n}}}, \quad (4.4)$$

em que  $R_{k\ell}^2$  é dado por 4.2 e  $\omega_{0\ell}$  é obtido por

$$\omega_{0\ell} = \sqrt{\frac{2\text{tr}(S_\ell^{*2})}{[\text{tr}(S_\ell^*)]^2} [(\rho_{k0}^2)^2 - 2\hat{\beta}_\ell \rho_{k0}^2 + \hat{\beta}_\ell]}$$

sendo

$$\hat{\beta}_\ell = \frac{\sum_{i=1}^k \lambda_i^{*2}}{\sum_{i=1}^p \lambda_i^{*2}}.$$

O valor  $Z_c$  da amostra original foi computado utilizando a expressão 3.4. Os valores de  $Z_{cl}$ , expressão 4.4, para as  $B - 1$  amostras de *bootstrap* constituem a distribuição nula de *bootstrap* paramétrico.

O valor-p foi computado por

$$\text{valor-p} = \frac{\sum_{\ell=1}^B I(Z_{c\ell} \leq Z_c)}{B}. \quad (4.5)$$

O valor-p foi confrontado com  $\alpha$  para a tomada de decisão a respeito da rejeição ou não da hipótese nula 3.1. Esse teste foi denotado pela sigla *BPZc* (Teste *bootstrap* paramétrico baseado na estatística  $Z_c$ ). A rotina com os comandos do software R Development Core Team (2009) para aplicação deste teste está apresentada no Anexo D.

#### 4.1.2 Teste de *bootstrap* não-paramétrico

A partir da amostra original  $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$  foi obtida a matriz de covariância amostral  $\mathbf{S}_{(pxp)}$  utilizando a expressão 3.3. Em seguida foi estimado  $\rho_k^2$  utilizando a expressão 3.2 na amostra original, fixado o valor  $k$  de interesse.

Na sequência foram realizadas reamostragens com reposição (*bootstrap*) da amostra original, preservando seu tamanho e a estrutura dos vetores  $\mathbf{Y}_j, j = 1, 2, \dots, n$ . Assim, foi obtida a amostra de *bootstrap* de tamanho  $n$ . Como os vetores de observações multivariadas são mantidos intactos, a estrutura de covariação das variáveis não é alterada. Para essa amostra foram calculados a matriz  $\mathbf{S}$  utilizando expressão 3.3, os autovalores  $\hat{\lambda}_i$ , e a porcentagem da explicação da variação total, utilizando a expressão 3.2, sendo nesse caso denotado por  $R_{k\ell}^2$  para indicar a estimativa na  $\ell$ -ésima amostra de *bootstrap*. Esse processo é repetido  $B - 1$  vezes. O valor original de  $R_k^2$  e os  $B - 1$  valores de *bootstrap* foram agrupados para formarem a distribuição de *bootstrap*.

O valor- $p$ , evidência contrária ou a favor de  $H_0$  dada em 3.1, foi determinado da seguinte forma :

$$\text{valor-p} = \frac{\sum_{\ell=1}^B I(R_{k\ell}^2 \geq \rho_{k0}^2)}{B}. \quad (4.6)$$

O valor- $p$  obtido em 4.6 foi confrontado com o valor da significância nominal  $\alpha$  para a tomada de decisão a respeito da rejeição ou não da hipótese nula 3.1. Esse teste foi denotado por  $BNP$  (Teste *bootstrap* não-paramétrico) e a rotina com os comandos do software R Development Core Team (2009) para sua aplicação está apresentada no Anexo E.

## 4.2 Validação Monte Carlo

Nas subseções a seguir estão apresentados os resultados da validação Monte Carlo dos testes propostos. O teste de Fujikoshi, utilizado apenas para comparação, foi denotado por  $Fj$ . Duas etapas foram consideradas. Na primeira os testes foram avaliados sob normalidade multivariada. Na segunda, considerou-se a distribuição  $t$  multivariada para se determinar a robustez.

### 4.2.1 Validação sob normalidade

Na Figura 2 estão apresentadas as curvas de poder dos testes  $Fj$ ,  $BNP$ ,  $BPRk$  e  $BPZc$  considerando  $H_0 : \rho_k^2 \geq 0,7$ , para  $k = 1$ ,  $p = 2$  e  $n = 15$ , sendo  $\alpha = 0,10$ ,  $0,05$  e  $0,01$ . A linha vertical, paralela à ordenada no ponto  $0,70$  representa a separação das regiões sob  $H_1$  e sob  $H_0$ . As linhas horizontais determinam os níveis de significância  $\alpha$  em cada caso. Essa é uma situação de pequenas amostras, portanto espera-se que o poder seja pequeno. Para  $\alpha = 0,10$  (Figura 2a), os valores de poder dos testes  $BPRk$  e  $BPZc$  foram praticamente iguais, iniciando-se em um valor próximo de  $0,30$  para  $\rho_k^2 = 0,50$  e decaindo até aproximadamente  $0,10$  quando o valor nominal paramétrico aproxima-se do valor hipotético  $0,70$ .

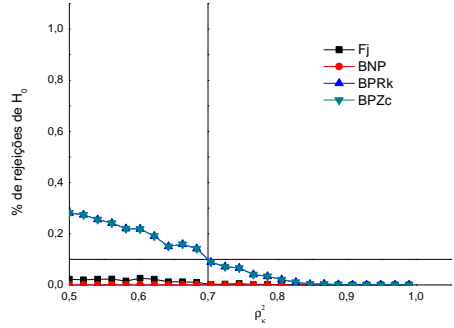


Os valores de poder dos testes  $Fj$  e  $BNP$  também foram semelhantes (Figura 2a), porém estiveram sempre abaixo do nível nominal de significância ( $\alpha = 0,10$ ). Para valores  $\rho_k^2 \geq 0,70$ , os valores das taxas de erro tipo I foram menores ou iguais ao valor nominal adotado. É conveniente chamar a atenção para o fato de que as curvas de poder na região sob  $H_0$  representam os valores das taxas de erro tipo I. Assim, os testes  $BPRk$  e  $BPZc$  foram semelhantes e superiores aos testes  $Fj$  e  $BNP$ , que por sua vez também foram idênticos em desempenho.

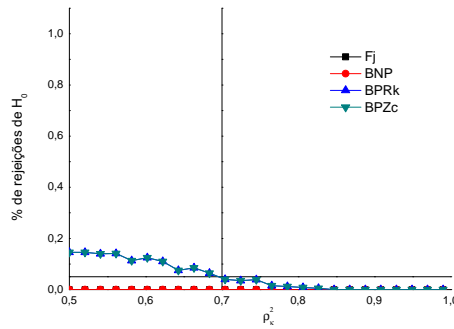
Para  $\alpha = 0,05$  (Figura 2b), foi observado o mesmo padrão quando  $\alpha = 0,10$ , embora com valores de poder menores. Isso já é esperado pela teoria, o que reforça a confiabilidade dos resultados da simulação realizada. Para  $\alpha = 0,01$  (Figura 2c) novamente observou-se o mesmo padrão de resposta, mas com valores de poder muito próximos do nível nominal adotado.

Depreende-se dessas observações que houve uma superioridade dos testes *bootstrap* paramétrico em relação ao teste assintótico ( $Fj$ ) e ao teste *bootstrap* não-paramétrico ( $BNP$ ). Do ponto de vista do controle do erro tipo I, todos eles apresentaram ótimo desempenho, ou seja, controlaram adequadamente o erro tipo I. Apesar dos testes,  $BPRk$  e  $BPZc$ , terem apresentado maiores valores para o poder, ainda assim esses valores são considerados de baixa magnitude, mesmo para valores de  $\rho_k^2$  afastados de  $\rho_{k0}^2 = 0,7$ .

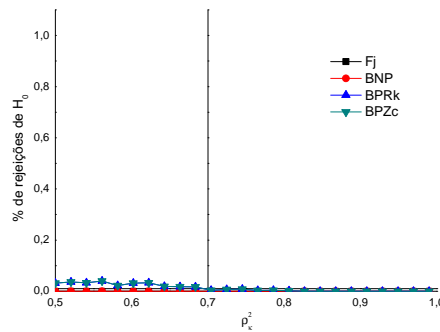
Na Figura 3 estão as curvas de poder dos quatro testes em situações idênticas às da Figura 2, exceto pelo valor hipotético, que nesse caso é igual a 0,90. De maneira geral o que se observa é que os testes  $BPRk$  e  $BPZc$  se aproximam da equivalência. O  $Fj$  apresenta em geral maior poder com  $\alpha \geq 0,05$  em relação ao  $BNP$ , invertendo o resultado para  $\alpha = 0,01$ . Entretanto, os valores de poder, em relação aos observados na Figura 2, foram muito superiores, pois o valor de  $\rho_{k0}^2$  está mais distante dos valores iniciais de  $\rho_k^2$  facilitando a detecção da difer-



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

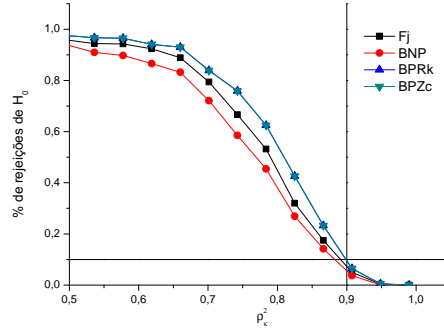
FIGURA 2 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

ença existente entre eles. Novamente, em relação ao controle do erro tipo I, os teste tiveram excelente desempenho. Todos eles apresentaram taxas de erro tipo I menores ou iguais aos valores de significância ( $\alpha$ ) nominais adotados. Em relação ao poder, os testes *BPRk* e *BPZc* foram superiores aos testes *Fj* e *BNP*. Para o nível nominal de significância de 0,01 (Figura 3c), as diferenças relativas em relação ao poder dos testes aumentaram, inclusive com a inversão do desempenho dos testes *Fj* e *BNP*, conforme salientado anteriormente.

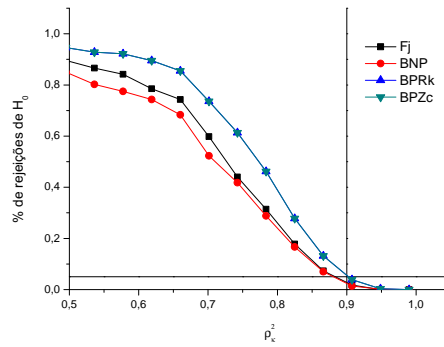
Perez-Neto et al. (2005) fizeram um estudo para comparar o desempenho de algumas regras de determinação do número de componentes principais. Dentre as regras estudadas, métodos *bootstrap* baseados em autovalores foram avaliados. A metodologia aplicada foi semelhante a usada neste trabalho, diferenciando-se pelo uso do método percentil para intervalos de confiança. Nesse estudo os métodos de *bootstrap* foram superados pelos métodos baseados em permutação e autovalores médios.

Nos Anexos F e G estão apresentados os resultados do desempenho dos teste nas mesmas configurações das Figuras 2 e 3, exceto pelo tamanho da amostra, que neste caso é  $n = 30$ . O padrão dos resultados encontrados em ambos os casos foram semelhantes, com os testes *BPRk* e *BPZc* superiores aos demais na maior parte dos casos. O que se deve observar, no entanto, é que houve maiores valores de poder para  $n = 30$  em relação as mesmas situações para  $n = 15$ , como já é preconizado pela teoria (Mood et al., 1974).

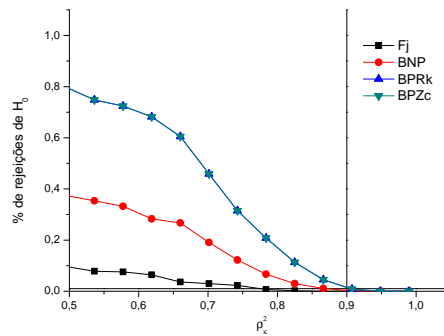
Na Figura 4 estão apresentados os resultados do poder e erro tipo I dos quatro testes estudados, nas mesmas situações utilizadas na Figura 2 e Anexo F, exceto pelo tamanho da amostra que neste caso foi igual a 100. O poder dos quatro testes, nessa situação de  $n = 100$  tendeu a se igualar. Mesmo assim, os testes *BPRk* e *BPZc* foram muito similares entre si e superiores aos demais.



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

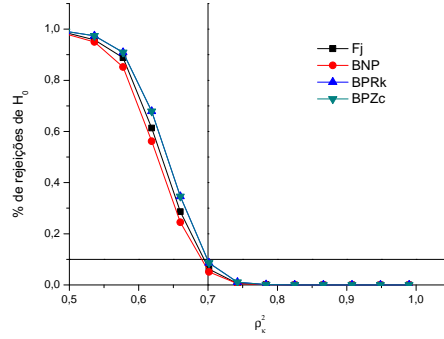
FIGURA 3 Poder dos testes de Fujikoshi ( $Fj$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

Para  $\alpha = 0,01$  (Figura 4c) houve um maior discernimento entre os desempenhos dos testes do que os casos de  $\alpha = 0,10$  e  $\alpha = 0,05$  (Figuras 4 a e b). Em todos os casos houve controle adequado do erro tipo I.

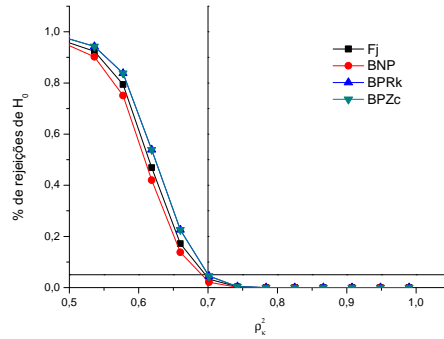
No estudo feito por Amorim et al. (2010) foi apresentado um teste Monte Carlo para a avaliação da unidimensionalidade de painéis sensoriais, o *TMCU*. Este teste tem por base avaliar se a proporção de explicação do primeiro componente principal é igual ou maior a um valor hipotético. O objetivo é avaliar se o primeiro componente principal contém quase toda a informação da variabilidade do painel sensorial, o que seria o mesmo que afirmar que o painel é unidimensional. Sendo assim, foi considerado  $k = 1$  na construção do *TMCU*, o que é uma limitação para sua aplicação em outras áreas. Já os testes propostos no presente trabalho, *BPRk*, *BPZc* e *BNP* podem ser aplicados para qualquer valor de  $k$ , desde que  $k < p$ .

A avaliação do desempenho do *TMCU* também foi feita através do estudo da probabilidade de erro tipo I e poder. Os resultados encontrados por Amorim et al. (2010) em relação ao erro tipo I foram semelhantes aos resultados dos testes aqui propostos. Em todas as situações, sob normalidade, o erro tipo I foi controlado adequadamente, ficando muito próximo ou abaixo do nível de significância considerado.

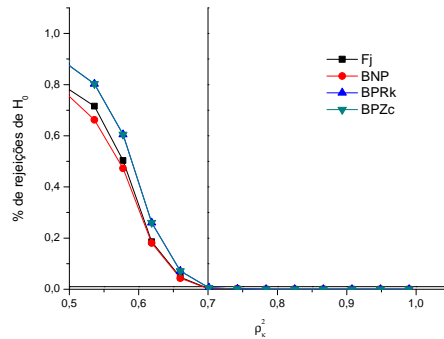
Na Figura 5 estão apresentados os resultados de erro tipo I e poder dos quatro testes para  $n = 100$  em situações similares às apresentadas na Figura 3 e Anexo G. Todos os testes apresentaram desempenho extremamente parecidos em relação ao poder e as taxas de erro tipo I. Isso é um forte indício de que os testes poderiam ser equivalentes assintoticamente. Para valores  $\rho_k^2$  entre 0,50 e 0,75 nos três casos (Figura 5 a, b e c) os valores de poder foram iguais a 100% para os quatro testes comparados. Para valores entre 0,75 e 0,90, os testes *BPRk* e *BPZc* foram



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

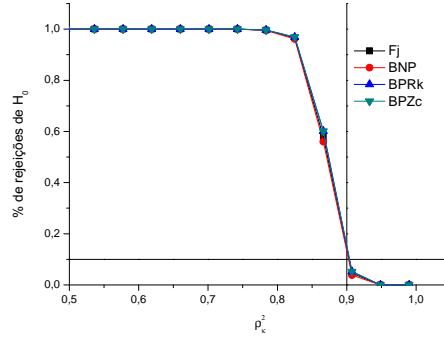
FIGURA 4 Poder dos testes de Fujikoshi ( $Fj$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_c^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

superiores aos outros dois, mas com diferenças desprezíveis. Houve controle do erro tipo I em todas as situações simuladas para os quatro testes.

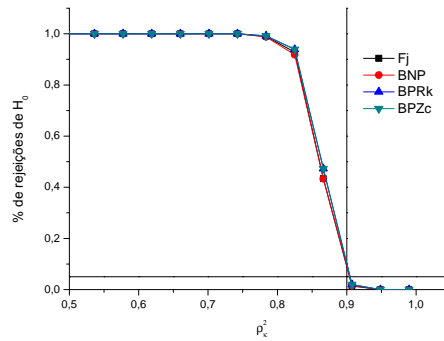
Nos Anexos de H a M, estão apresentados os resultados do desempenho dos testes nas mesmas situações das Figuras 2, 3, Anexos F, G e Figuras 4 e 5 em ordem correspondente, exceto pelo fato do número de variáveis,  $p$ , ser igual a 10 nesses casos. Em todos os casos os quatro testes tiveram desempenhos bem semelhantes, com pequena superioridade para  $BPRk$  e  $BPZc$  em algumas circunstâncias. Comparando o padrão de resposta com as situações anteriores ( $p = 2$ ), verifica-se uma grande semelhança, exceto o fato de que o aumento do número de variáveis teve um efeito benéfico, pois os valores de poder foram superiores e as curvas de rejeições de  $H_0$  apresentaram um comportamento mais adequado em relação ao que se espera, pela teoria, de uma curva de poder. No teste apresentado por Amorim et al. (2010), onde o número de provadores do painel sensorial corresponde ao número de variáveis  $p$ , observa-se que o aumento desse valor também provoca um efeito benéfico no poder do  $TMCU$ .

Os valores de poder encontrados por Amorim et al. (2010) apresentaram um padrão de comportamento semelhante aos testes aqui propostos, sendo que em situações onde o tamanho de amostra é menor, os valores de poder também são menores. Os tamanhos de amostra, para Amorim et al. (2010), correspondem ao número de produtos degustados pelos provadores, sendo assim, os maiores valores de poder foram observados em situações em que o número de produtos degustados e o número de provadores eram maiores.

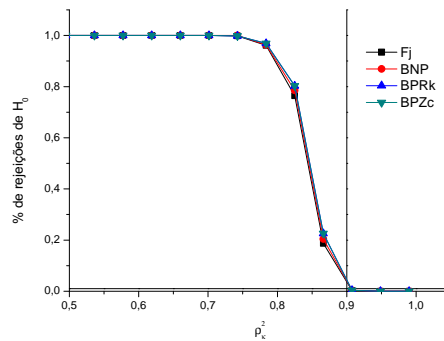
Nas Figuras 6 e 7 estão apresentados os valores de poder e erro tipo I semelhantes às situações apresentadas nas Figuras 4 e 5, com a exceção do número de variáveis ( $p$ ) ser igual a 10 e o número de componentes principais que explica a proporção mínima  $\rho_k^2$  da variação total,  $k$ , ser igual a 3. Da mesma forma que



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

FIGURA 5 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



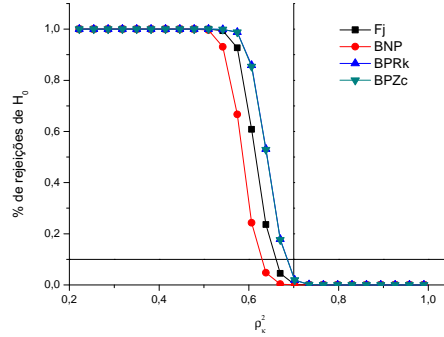
ocorreu anteriormente, os testes  $BPRk$  e  $BPZc$  foram superiores aos seus concorrentes em relação ao poder e todos controlaram o erro tipo I. O aumento de  $\rho_k^2$  de 0,70 para 0,90 teve um efeito de equiparar o desempenho dos testes. O teste  $BNP$  foi novamente o de pior desempenho geral, sendo que em alguns casos a redução de seu poder é bem drástica em relação aos outros três testes.

Nos Anexos de N a Q estão apresentados os casos similares aos das Figuras 6 e 7, afora pelos tamanhos amostrais, que foram iguais a 15 e 30. Nestas figuras, os resultados proporcionaram o mesmo padrão de resposta observado para  $n = 100$ .

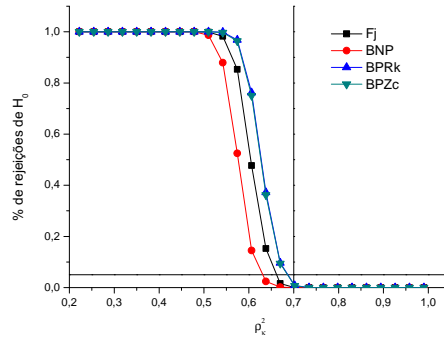
#### 4.2.2 Validação sob a $t$ multivariada

Verificou-se o desempenho dos testes em uma situação em que um dos principais pressupostos para a maioria deles foi violado, que é a normalidade multivariada. Optou-se pela distribuição  $t$  de Student multivariada com grau de liberdade  $\nu = 1$ . Com isso pretendeu-se avaliar a robustez dos quatro testes. Dentre os quatro testes, apenas o  $BNP$  não pressupõe normalidade para a sua aplicação. Assim, antecipa-se que este teste possa ter um melhor desempenho.

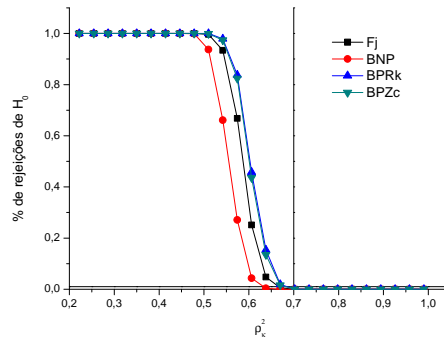
Na Figura 8, estão apresentados os valores de poder e erro tipo I dos testes  $Fj$ ,  $BNP$ ,  $BPRk$  e  $BPZc$  em função dos valores de  $\rho_k^2$ , para  $k = 1$ ,  $p = 10$  e  $n = 30$  com hipótese nula  $H_0 : \rho_k^2 = 0,90$ , considerando ainda  $\alpha = 0,10$ ,  $0,05$  e  $0,01$ . De maneira geral o que se observa é que os testes  $Fj$ ,  $BPRk$  e  $BPZc$  tiveram desempenho praticamente idêntico, como pode ser constatado na Figura 8, observando que suas curvas são indistinguíveis. Esses testes também apresentaram fraco desempenho no que tange ao controle do erro tipo I. Considerando-se todos os níveis nominais de significância estudados (Figuras 8 a, b e c) verificou-se que esses três testes foram liberais, principalmente quando o valor paramétrico  $\rho_k^2$  se



(a)  $\alpha = 0,10$

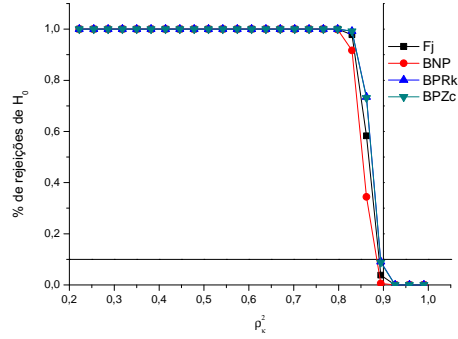


(b)  $\alpha = 0,05$

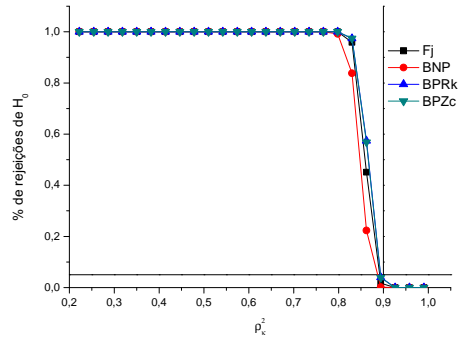


(c)  $\alpha = 0,01$

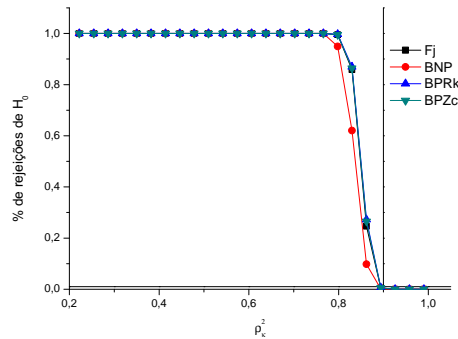
FIGURA 6 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_c^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

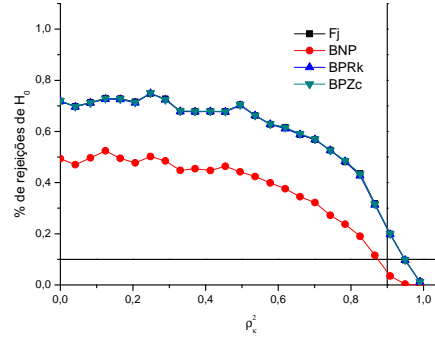
FIGURA 7 Poder dos testes de Fujikoshi ( $F_j$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

situava próximo do valor hipotético  $\rho_{k0}^2$  (0,90). Na medida em que o valor de  $\rho_k^2$  se afasta de  $\rho_{k0}^2$ , mas continua-se sob  $H_0$ , as taxas de erro tipo I se aproximam do valor nominal ou se tornam inferiores a ele. Assim, estes testes passam a controlar adequadamente o erro tipo I quando  $\rho_k^2$  se aproxima de 1. A aproximação das taxas de erro tipo I em relação ao valor de significância nominal  $\alpha$  é mais lenta à medida que o valor de  $\alpha$  diminui.

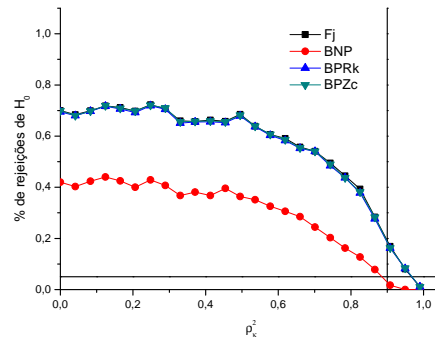
O teste *BNP*, por outro lado, apresentou controle adequado do erro tipo I em todos os casos. Como consequência disso, espera-se que este teste, que é conservativo, seja menos poderoso que os seus três competidores, que foram liberais. Isso realmente aconteceu, como pode ser observado na Figura 8. As perdas relativas mais acentuadas se deram para o nível nominal de significância de 0,01 (Figura 8 c). Isso também era esperado, pois os testes *Fj*, *BPRk* e *BPZc* foram mais liberais nesse caso. Assim, conclui-se que o único teste considerado robusto foi o *BNP*. Esse resultado já havia sido antecipado, uma vez que esse teste é o único que não pressupõe normalidade multivariada. Convém salientar que os testes *Fj*, *BPRk* e *BPZc* apresentaram taxas de erro tipo I aproximadamente iguais a 20%, 16% e 11% para  $\alpha = 0,10, 0,05, \text{ e } 0,01$ , respectivamente, quando  $\rho_k^2 \cong \rho_{k0}^2$  nas simulações realizadas.

Na Figura 9 estão os valores de poder e erro tipo I dos quatro testes com  $n = 100$ , nas mesmas configurações apresentadas na Figura 8. O que se observa nesse caso, é o mesmo padrão de resposta encontrado para  $n = 30$ . O que foi mais surpreendente é que o caráter liberal dos testes *Fj*, *BPRk* e *BPZc* se acentuou com o aumento do tamanho amostral. Não houve grandes ganhos de poder do teste *BNP* quando  $n$  aumentou de 30 para 100. O aumento do poder dos outros três testes pode ser atribuído ao aumento das taxas de erro tipo I.

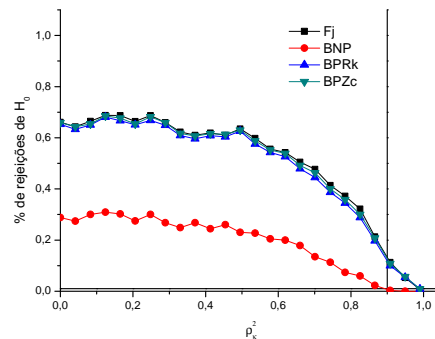
Como o desempenho dos testes propostos variou em função da distribuição



(a)  $\alpha = 0,10$

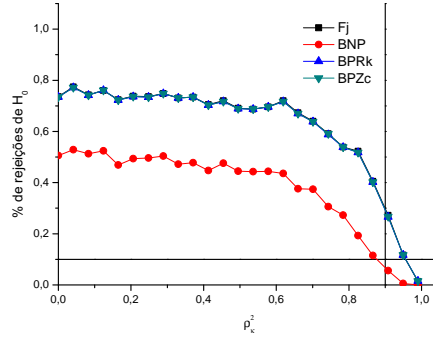


(b)  $\alpha = 0,05$

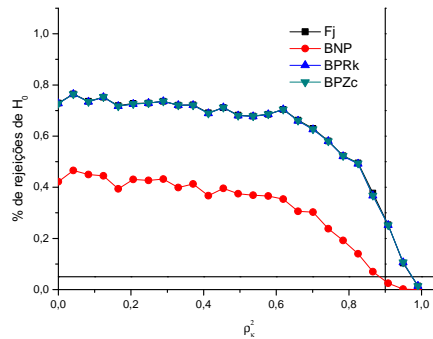


(c)  $\alpha = 0,01$

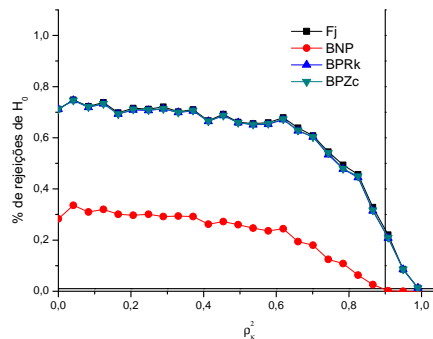
FIGURA 8 Poder dos testes de Fujikoshi ( $Fj$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_c^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



(a)  $\alpha = 0,10$



(b)  $\alpha = 0,05$



(c)  $\alpha = 0,01$

FIGURA 9 Poder dos testes de Fujikoshi ( $Fj$ ), *bootstrap* não paramétrico ( $BNP$ ), *bootstrap* paramétrico com  $R_k^2$  ( $BPRk$ ) e *bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) em função de  $\rho_c^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

de probabilidade de onde a amostra foi retirada, é necessário que se aplique um teste de normalidade multivariada aos dados amostrais (Cirillo & Ferreira, 2003) (Royston, 1983b) antes de se decidir qual teste recomendar. Sob normalidade multivariada os testes  $BPRk$  e  $BPZc$  são recomendados e em caso contrário, o teste  $BNP$  é indicado.

## 5 CONCLUSÕES

O presente trabalho tem as seguintes conclusões:

1. Os três novos testes *bootstrap* para a proporção de explicação de componentes principais foram propostos com sucesso;
2. Sob normalidade multivariada os testes *bootstrap* paramétrico com  $R_k^2$  ( $BPk$ ) e *Bootstrap* paramétrico com  $Z_c$  ( $BPZc$ ) são recomendados, pois tiveram os maiores valores de poder e controlaram o erro tipo I;
3. Sob não-normalidade o teste *Bootstrap* não-paramétrico  $BNP$  é recomendado por ser robusto, uma vez que foi o único a controlar o erro tipo I;
4. Os testes propostos superaram o teste de Fujikoshi na maioria das situações.



## REFERÊNCIAS BIBLIOGRÁFICAS

- AFIFI, A. A. **Computer aided multivariate analysis**. 3. ed. Los Angeles: Chapman & Hall, 1996. 455 p.
- AMORIM, I. S.; FERREIRA, E. B.; LIMA, R. R.; PEREIRA, R. G. F. A. Monte carlo based test for inferring about the unidimensionality of a brazilian coffee sensory panel. **Food Quality and Preference**, Barking, v. 21, n. 3, p. 319–323, 2010.
- ANDERSON, T. W. **An itroduction to multivariate statistical analysis**. New York: J. Wiley, 1971. 675 p.
- BANZATTO, D. A.; KRONKA, S. N. **Experimentação agrícola**. 4. ed. Jaboticabal: Fealq, 2006. 237 p.
- BARON, M. **Probability and statistics for computer scientists**. Boca Raton: Chapman & Hall/CRC, 2007. 413 p.
- CECCHETTI, D. **Poder e taxas de erro tipo I de quatro critérios multivariados para o teste de igualdade de efeitos de tratamentos avaliados por meio do método de Monte Carlo**. 1999. 56 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras.
- CHERNICK, M. R. **Bootstrap methods: a guide for practitioners and researchers**. 2. ed. Newtown: Wiley-Interscience, 2008. 369 p.
- CIRILLO, M. A. **Teste de normalidade e avaliação por simulação Monte Carlo do problema de Behrens-Fisher no espaço p-dimensional**. 2003. 128 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras.
- CIRILLO, M. A.; FERREIRA, D. F. Extensão do teste para normalidade univariado baseado no coeficiente de correlação quantil-quantil para o caso multivariado. **Revista de Matemática e Estatística**, Marília, v. 21, n. 3, p. 57–75, 2003.
- COSTA, S. F. **Introdução ilustrada à estatística (com muito humor)**. São Paulo: Harbra, 1992. 303 p.
- DACHS, J. N. **Estatística computacional: uma introdução em turbo pascal**. Rio de Janeiro: Livros Técnicos e Científicos, 1988. 236 p.

- DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. Cambridge: Cambridge University, 2008. 582 p.
- DUNTEMAN, G. H. **Principal components analysis**. Newbury Park: SAGE, 1989. 96 p.
- EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistics**, Hayward, v. 7, n. 1, p. 1–26, Jan. 1979a.
- EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. Boca Raton: Chapman & Hall, 1993. 436 p.
- FERREIRA, D. F. **Estatística básica**. Lavras: UFLA, 2005. 664 p.
- FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 662 p.
- FLECK, M. P. A.; BOURDEL, M. C. Método de simulação e escolha de fatores na análise dos principais componentes. **Revista de Saúde Pública**, São Paulo, v. 32, n. 3, p. 267–272, Jun. 1998.
- FUJIKOSHI, Y. Asymptotic expansions for the distributions of the sample roots under nonnormality. **Biometrika**, London, v. 67, n. 1, p. 45–51, 1980.
- GENTLE, J. E. **Random number generation and Monte Carlo methods**. 2. ed. New York: Springer, 2003. 381 p. (Statistics and computing).
- GIVENS, G. H.; HOETING, J. A. **Computational statistics**. Hoboken: Wiley-Interscience, 2005. 418 p.
- GROSSMAN, G. D.; NICKERSON, D. M.; FREEMAN, M. C. Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. **Bulletin of Ecological Society of America**, Tempe, v. 72, n. 1, p. 341–347, 1991.
- HARRIS, R. J. **A Primer of multivariate statistics**. 3. ed. Mahwah: Lawrence Erlbaum Associates, 2001. 609 p.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, Arlington, v. 24, p. 417–441, Sept. 1933.
- JOLLIFFE, I. T. **Principal components analysis**. 2. ed. New York: Springer Verlag, 2002. 487 p.

- KUNSCH, H. R. The jackknife and the bootstrap for general stationary observations. **Annals of Statistics**, Hayward, v. 17, n. 3, p. 1217–1241, Mar. 1989.
- MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. 2. ed. São Paulo: Edusp, 2006. 428 p.
- MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. Singapore: McGraw-Hill, 1974. 564 p.
- MORRISON, D. F. **Multivariate statistical methods**. 3. ed. New York: McGraw-Hill, 1990. 495 p.
- OLIVEIRA, M. S.; BEARZOTI, E.; BOAS, F. L. V.; NOGUEIRA, D. A.; NICOLAU, L. A. **Introdução à estatística**. Lavras: UFLA, 2009. 334 p.
- PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, London, v. 6, n. 2, p. 559–572, 1901.
- PEREZ-NETO, P.; JACKSON, D. A.; SOMERS, K. M. How many principal components? stopping rules for determining the number of non-trivial axes revisited. **Computational Statistics & Data Analysis**, Amsterdam, v. 49, n. 4, p. 974–997, June. 2005.
- PRESS, W. H.; FLANNERY, B. P.; TEUKOLSKYY, S. A.; VETTERLING, W. T. **Numerical recipes in fortran: the art of scientific computing**. Cambridge: Cambridge University, 1992. 994 p.
- R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <<http://www.R-project.org>>. Acesso em: 20 dez. 2009.
- RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2000. 303 p.
- RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. New York: J. Wiley, 2002. 708 p.
- ROYSTON, J. P. Some techniques for assessing multivariate normality based on the shapiro - wilk w. **Journal of the Royal Statistical Society: Series C, Applied Statistics**, London, v. 32, n. 2, p. 121–133, Jan. 1983b.

SILVA, E. C. **Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo.** 1998. 54 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian, and MCMC: methods in quantitative analysis.** New York: Springer Verlag, 2002. 740 p.

TIMM, N. H. **Applied multivariate analysis.** New York: Springer Verlag, 2002. 693 p.

## ANEXOS

ANEXO A *Exemplo numérico para a construção da matriz  $\mathbf{\Lambda}_b$ .*

Seja  $\mathbf{Y}_{(n \times p)}$  uma matriz de dados e  $\mathbf{S}_{(p \times p)}$  a sua matriz de covariâncias, sendo  $p = 4$  o número de variáveis. Considera-se os autovalores da matriz de covariâncias como  $\hat{\lambda}_1 = 6$ ,  $\hat{\lambda}_2 = 4$ ,  $\hat{\lambda}_3 = 3$  e  $\hat{\lambda}_4 = 2$  e a partir da decomposição  $\mathbf{\Sigma} = \hat{\mathbf{P}}\mathbf{\Lambda}\hat{\mathbf{P}}^\top$ , tem-se que:

$$\hat{\mathbf{\Lambda}} = \begin{pmatrix} \hat{\lambda}_1 & 0 & 0 & 0 \\ 0 & \hat{\lambda}_2 & 0 & 0 \\ 0 & 0 & \hat{\lambda}_3 & 0 \\ 0 & 0 & 0 & \hat{\lambda}_4 \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Suponha que o pesquisador deseja que dois, dos primeiros componentes principais, detenham 80% da informação das variáveis originais, ou seja,  $\rho_{k0}^2 = 0,80$  e  $k = 2$ . É sabido que  $\sum_{i=1}^p \hat{\lambda}_i = 15$  é a variação total.

Verifica-se que os dois primeiros componentes principais têm uma variação total igual a 10 e os últimos igual a 5.

$$(\hat{\lambda}_1 = 6) + (\hat{\lambda}_2 = 4) = 10$$

$$(\hat{\lambda}_3 = 3) + (\hat{\lambda}_4 = 2) = 5.$$

O objetivo é fazer com que a soma dos dois primeiros autovalores seja igual a 12, pois 12 representa 80% da variação total 15. Para que isso seja feito necessita-se que duas unidades da variação total sejam retiradas dos últimos autovalores e acrescentadas aos primeiros.

O terceiro autovalor ( $\hat{\lambda}_3 = 3$ ), representa 60% da soma dos últimos autovalores, obviamente  $\hat{\lambda}_4 = 2$  representa 40%. Deve-se retirar 1,2 do valor de  $\hat{\lambda}_3$ , ou seja, retira-se 60% de 2 do terceiro autovalor. Do mesmo modo, retira-se 0,8 do valor de  $\hat{\lambda}_4$ , que representa 40% das duas unidades retiradas dos últimos autovalores. Retiradas essas duas unidades deve-se acrescentá-las aos primeiros autovalores da mesma forma, proporcionalmente. Assim, como  $\hat{\lambda}_1$  representa 60% da soma dos dois primeiros autovalores, acrescenta-se 1,2 ao seu valor, já para  $\hat{\lambda}_2$  deve-se acrescentar 0,8.

Desse modo, os novos autovalores são:

$$\begin{aligned}(\hat{\lambda}_{1b} = 7,2) + (\hat{\lambda}_{2b} = 4,8) &= 12 \\(\hat{\lambda}_{3b} = 1,8) + (\hat{\lambda}_{4b} = 1,2) &= 3,\end{aligned}$$

sendo que os dois primeiros têm 80% da variação total.

Logo,

$$\mathbf{\Lambda}_b = \begin{pmatrix} 7,2 & 0 & 0 & 0 \\ 0 & 4,8 & 0 & 0 \\ 0 & 0 & 1,8 & 0 \\ 0 & 0 & 0 & 1,2 \end{pmatrix}$$

é uma matriz reformulada de forma que os dois primeiros componentes principais detenham 80% da informação das variáveis originais.

ANEXO B *Rotina no R utilizada para a aplicação do teste de Fujikoshi.*

```
# A função deve receber:
# k: número almejado dos primeiros componentes principais;
# rho2k0: valor hipotético para o percentual de explicação dos
#       k primeiros componentes;
# n: tamanho da amostra;
# Y: matriz de dados (n x p).

Fujikoshi <- function(k, rho2k0, n, Y)
{
  p <- ncol(Y)
  if (k >= p) stop("K deve ser menor que p!")
  S <- var(Y)
  eig <- eigen(S)
  betah <- sum(eig$values[1:k]*eig$values[1:k])/
           sum(eig$values*eig$values)
  w0 <- 2 * sum(eig$values^2)/sum(eig$values)^2 *
        (rho2k0^2 - 2 * betah * rho2k0 + betah)
  w0 <- w0^0.5
  R2k <- sum(eig$values[1:k])/sum(eig$values)
  Zc <- sqrt(n) * (R2k - rho2k0) / w0
  valor.p <- pnorm(Zc)
  return(list(R2k = R2k, Zc = Zc, valor.p = valor.p))
}
```

ANEXO C *Rotina no R utilizada para a aplicação do teste bootstrap paramétrico baseado na estatística  $R_k^2$ .*

```
# A função deve receber:
# k: número almejado dos primeiros componentes principais;
# rho2k0: valor hipotético para o percentual de explicação dos
#       k primeiros componentes;
# n: tamanho da amostra;
# B: número de simulações bootstrap;
# Y: matriz de dados (n x p)

PCBootP <- function(k, rho2k0, n, B, Y)
{
  p <- ncol(Y)
  if (k >= p) stop("K deve ser menor que p!")
  S <- var(Y)
  eig <- eigen(S)
  R2k <- sum(eig$values[1:k])/sum(eig$values)
  Vt <- sum(eig$values)
  Delta <- rho2k0 * Vt - sum(eig$values[1:k])
  lamb <- eig$values[1:k] + eig$values[1:k] /
    sum(eig$values[1:k]) * Delta
  lamb2 <- eig$values[(k+1):p]
  lamb2 <- lamb2 - lamb2 /sum(eig$values[(k+1):p]) * Delta
  lamb <- c(lamb,lamb2)
  Sigb <- eig$vectors %*% diag(as.vector(lamb)) %*%
    t(eig$vectors)
  R2kb <- R2k
  for (ib in 1:(B-1))
  {
    Yfake <- mvrnorm(n, mu=rep(c(0),times=p), Sigma=Sigb)
    Sfake <- var(Yfake)
  }
}
```



```
eigfake <- eigen(Sfake)
R2kfake <- sum(eigfake$values[1:k])/sum(eigfake$values)
R2kb    <- rbind(R2kb, R2kfake)
}
valor.p <- length(R2kb[R2kb <= R2k]) / B
return(list(R2k=R2k, valor.p=valor.p))
}
```

**ANEXO D** *Rotina no R utilizada para a aplicação do teste bootstrap paramétrico baseado na estatística  $Z_c$ .*

```
# A função deve receber:
# k: número almejado dos primeiros componentes principais;
# rho2k0: valor hipotético para o percentual de explicação dos
#       k primeiros componentes;
# n: tamanho da amostra;
# B: número de simulações bootstrap;
# Y: matriz de dados (n x p)

PCBootPZc <- function(k, rho2k0, n, B, Y)
{
  p <- ncol(Y)
  if (k >= p) stop("K deve ser menor que p!")
  S <- var(Y)
  eig <- eigen(S)
  R2k <- sum(eig$values[1:k])/sum(eig$values)
  Vt <- sum(eig$values)
  Delta <- rho2k0 * Vt - sum(eig$values[1:k])
  lamb <- eig$values[1:k] + eig$values[1:k] /
    sum(eig$values[1:k]) * Delta
  lamb2 <- eig$values[(k+1):p]
  lamb2 <- lamb2 - lamb2 /sum(eig$values[(k+1):p]) * Delta
  lamb <- c(lamb,lamb2)
  Sigb <- eig$vectors %*% diag(as.vector(lamb)) %*%
    t(eig$vectors)
  betah <- sum(eig$values[1:k]*eig$values[1:k])/
    sum(eig$values*eig$values)
  w0 <- 2 * sum(eig$values^2)/sum(eig$values)^2 *
    (rho2k0^2 - 2 * betah * rho2k0 + betah)
  w0 <- w0^0.5
  Zc <- sqrt(n) * (R2k - rho2k0) / w0
}
```

```

Zcb <- Zc
for (ib in 1:(B-1))
{
  Yfake <- mvrnorm(n, mu=rep(c(0),times=p), Sigma=Sigb)
  Sfake <- var(Yfake)
  eigfake <- eigen(Sfake)
  R2kfake <- sum(eigfake$values[1:k])/sum(eigfake$values)
  betah <- sum(eigfake$values[1:k]*eigfake$values[1:k]) /
            sum(eigfake$values*eigfake$values)
  w0b <- 2 * sum(eigfake$values^2)/sum(eigfake$values)^2 *
          (rho2k0^2 - 2 * betah * rho2k0 + betah)
  w0b <- w0b^0.5
  Zcfake <- sqrt(n) * (R2kfake - rho2k0) / w0b
  Zcb <- rbind(Zcb, Zcfake)
}
valor.p <- length(Zcb[Zcb <= Zc]) / B
return(list(R2k = R2k, Zc = Zc, valor.p = valor.p))
}

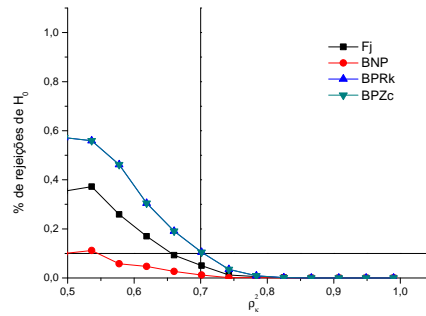
```

ANEXO E *Rotina no R utilizada para a aplicação do teste bootstrap não-paramétrico.*

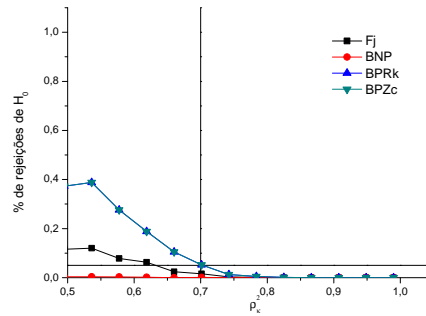
```
# A função deve receber:
# k: número almejado dos primeiros componentes principais;
# rho2k0: valor hipotético para o percentual de explicação dos
#       k primeiros componentes;
# n: tamanho da amostra;
# B: número de simulações bootstrap;
# Y: matriz de dados (n x p)

PCBootNP <- function(k, rho2k0, n, B, Y)
{
  p <- ncol(Y)
  if (k >= p) stop("K deve ser menor que p!")
  S <- var(Y)
  eig <- eigen(S)
  R2k <- sum(eig$values[1:k])/sum(eig$values)
  R2kb <- R2k
  for (ib in 1:(B-1))
  {
    rowy <- sample(1:n,replace = T)
    Yfake <- Y[rowy,]
    Sfake <- var(Yfake)
    eigfake <- eigen(Sfake)
    R2kfake <- sum(eigfake$values[1:k])/sum(eigfake$values)
    R2kb <- rbind(R2kb, R2kfake)
  }
  valor.p <- length(R2kb[R2kb >= rho2k0]) / B
  return(list(R2k=R2k, valor.p=valor.p))
}
```

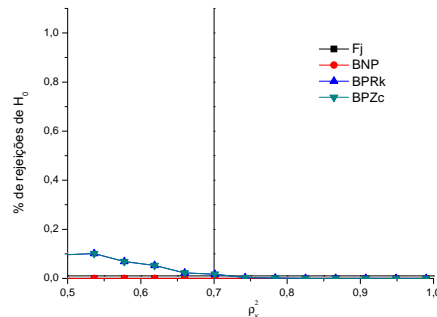
ANEXO F Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



( $\alpha = 0,10$ )

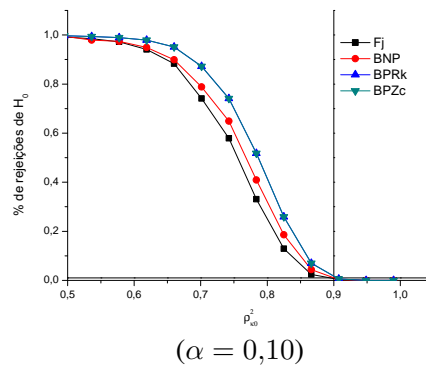
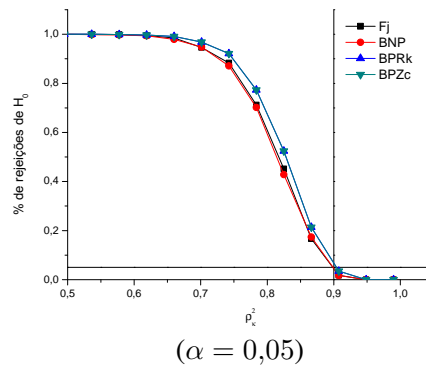
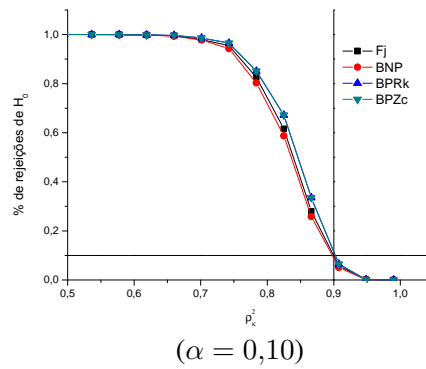


( $\alpha = 0,05$ )

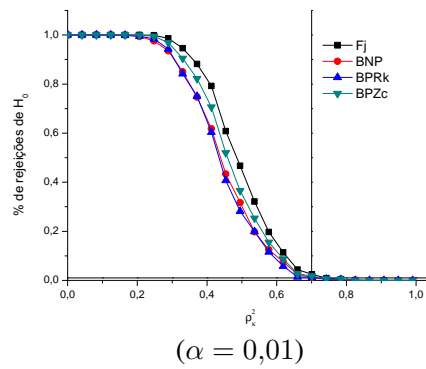
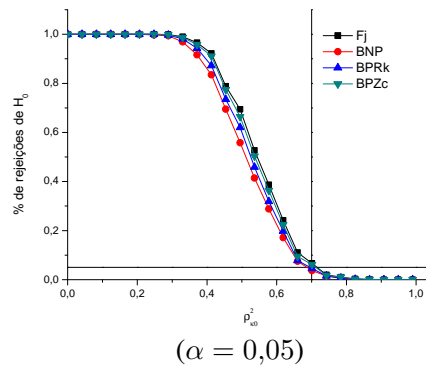
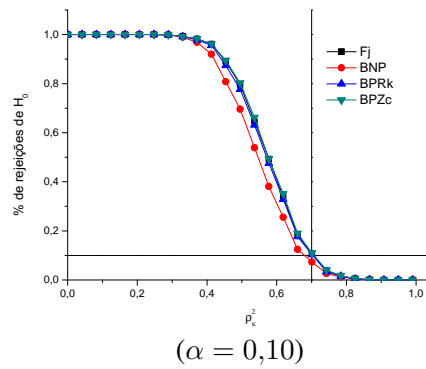


( $\alpha = 0,01$ )

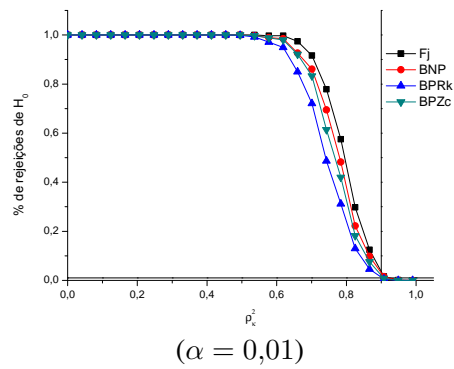
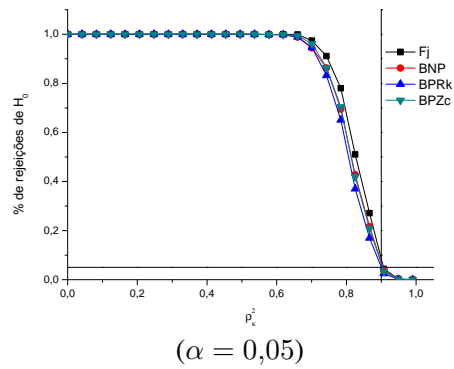
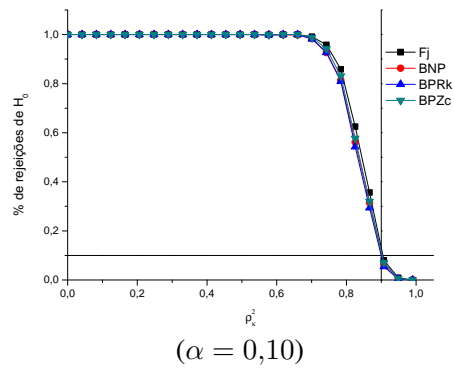
ANEXO G Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 2$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



ANEXO H Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_{k0}^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

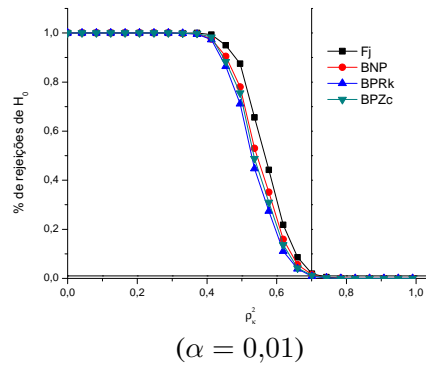
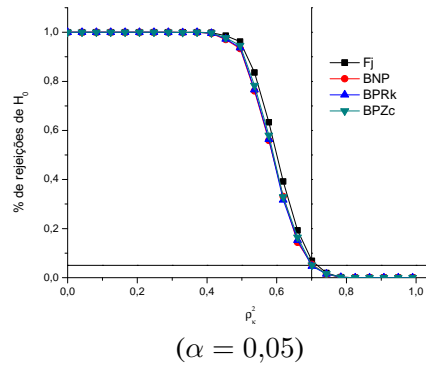
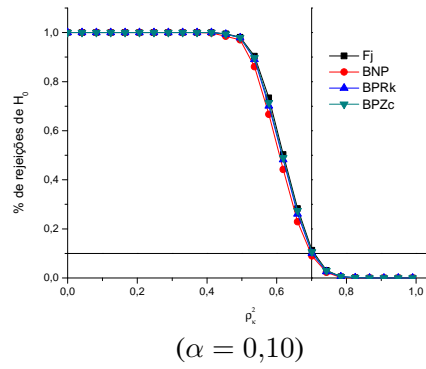


ANEXO I Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_{k0}^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

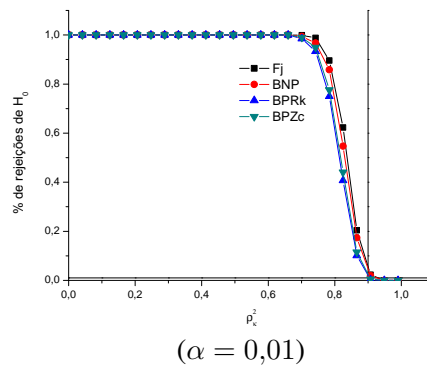
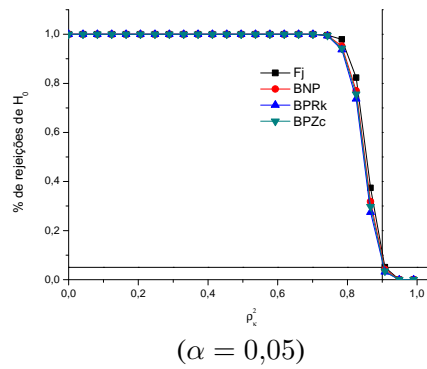
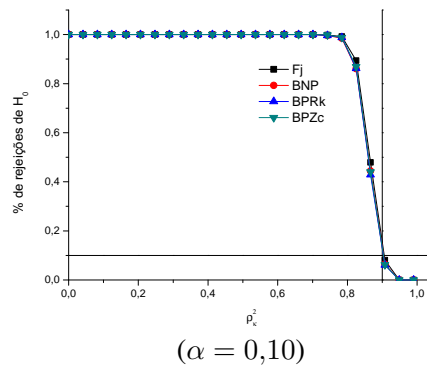




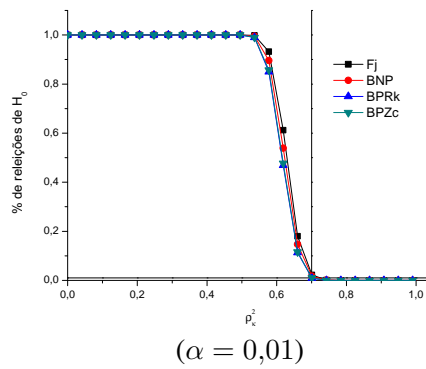
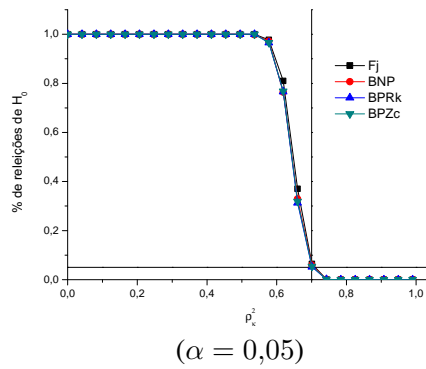
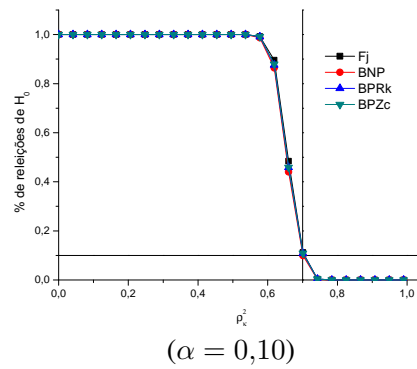
ANEXO J Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



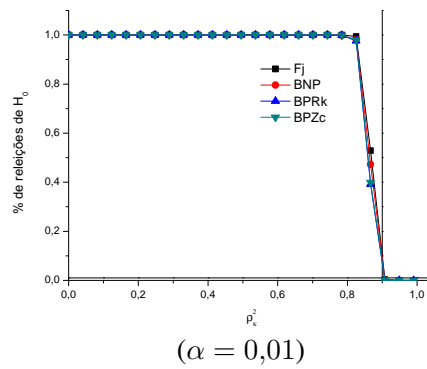
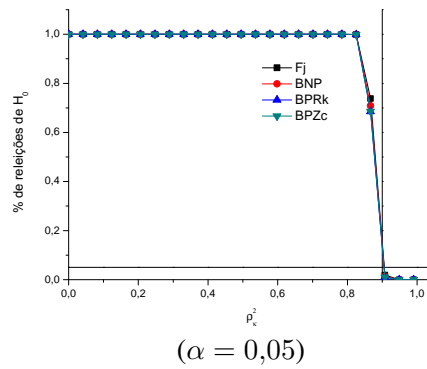
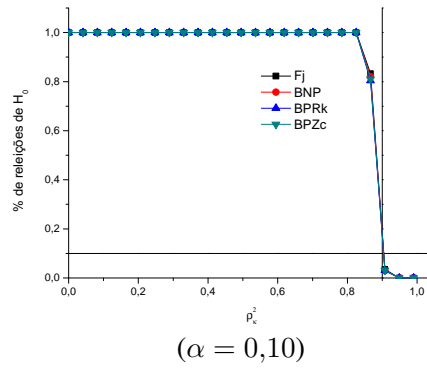
ANEXO K Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



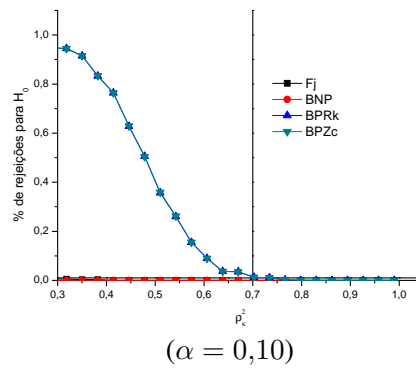
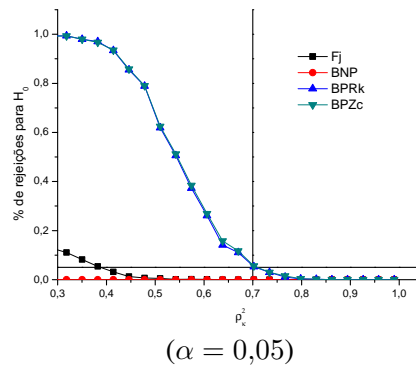
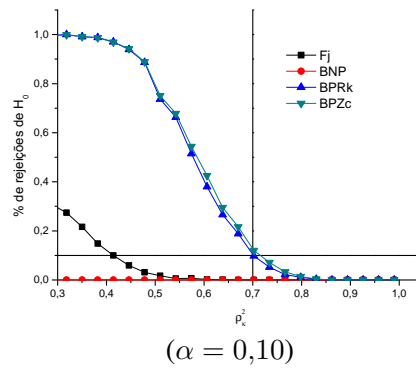
ANEXO L Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



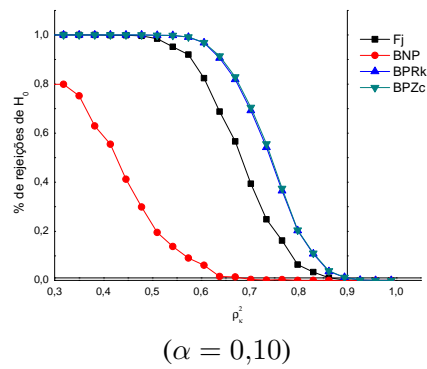
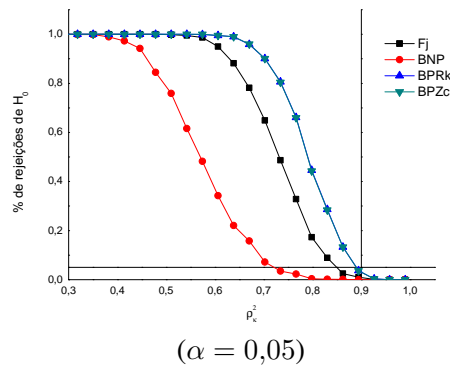
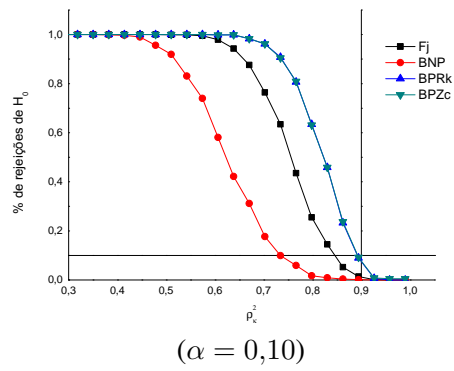
ANEXO M Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 1$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 100$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



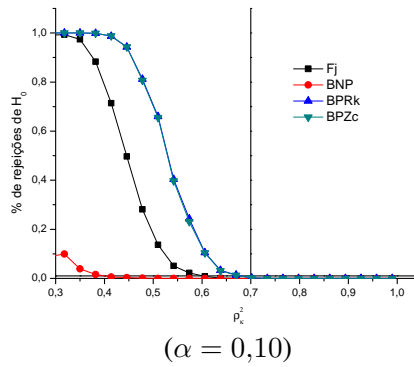
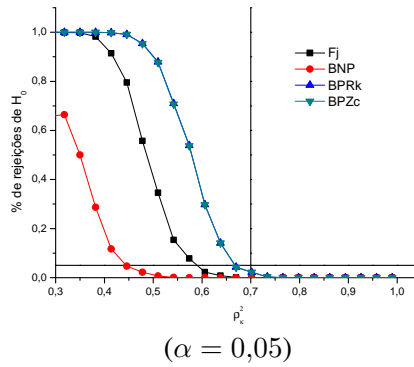
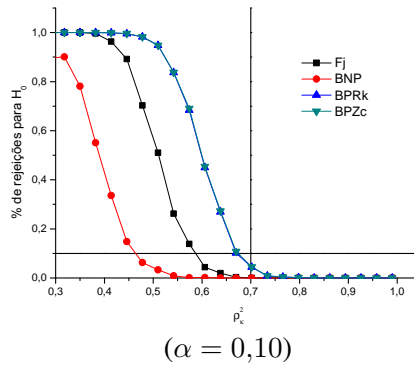
ANEXO N Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



ANEXO O Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_{k0}^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 15$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



ANEXO P Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_{k0}^2$  para  $H_0 : \rho_{k0}^2 \geq 0,70$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .



ANEXO Q Poder dos testes de Fujikoshi ( $F_j$ ), bootstrap não paramétrico (BNP), bootstrap paramétrico com  $R_k^2$  (BPRk) e bootstrap paramétrico com  $Z_c$  (BPZc) em função de  $\rho_k^2$  para  $H_0 : \rho_{k0}^2 \geq 0,90$ , considerando  $k = 3$  componentes,  $p = 10$  variáveis e amostra de tamanho  $n = 30$  para (a)  $\alpha = 0,10$ , (b)  $\alpha = 0,05$  e (c)  $\alpha = 0,01$ .

