ELCIO DO NASCIMENTO CHAGAS

EFICIÊNCIA DE ESTIMADORES ROBUSTOS A OBSERVAÇÕES DISCREPANTES EM REGRESSÃO MULTIVARIADA COM APLICAÇÃO NA ANÁLISE SENSORIAL DE CAFÉ

LAVRAS-MG

2011

ELCIO DO NASCIMENTO CHAGAS

EFICIÊNCIA DE ESTIMADORES ROBUSTOS A OBSERVAÇÕES DISCREPANTES EM REGRESSÃO MULTIVARIADA COM APLICAÇÃO NA ANÁLISE SENSORIAL DE CAFÉ

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

Orientador

Dr. Augusto Ramalho de Morais

Coorientador

Dr. Marcelo Angelo Cirillo

LAVRAS-MG

2011

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

Chagas, Elcio do Nascimento.
Eficiência de estimadores robustos a observações discrepantes em regressão multivariada com aplicação na análise sensorial de café / Elcio do Nascimento Chagas. – Lavras : UFLA, 2011.
94 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2011.
Orientador: Augusto Ramalho de Morais.
Bibliografia.

1. Análise de regressão. 2. Regressão robusta. 3. Distância de Mahalanobis. 4. Distância robusta. I. Universidade Federal de Lavras. II. Título.

CDD – 519.536

ELCIO DO NASCIMENTO CHAGAS

EFICIÊNCIA DE ESTIMADORES ROBUSTOS A OBSERVAÇÕES DISCREPANTES EM REGRESSÃO MULTIVARIADA COM APLICAÇÃO NA ANÁLISE SENSORIAL DE CAFÉ

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

APROVADA em 03 de Agosto de 2011.

Dr. Carlos Tadeu Santos Dias	ESALQ - USP
Dr. Fabyano Fonseca e Silva	UFV
Dr. Joel Augusto Muniz	UFLA
Dr. Marcelo Angelo Cirillo	UFLA

Dr. Augusto Ramalho de Morais
Orientador

LAVRAS-MG

2011

À memória de meus pais: Pedro Francisco das Chagas e Maria do Nascimento Chagas.

À minha esposa Lana, pela paciência e auxílio.

Às minhas filhas, Liz e Maria, pelo apoio, carinho e dedicação.

À toda minha família, pelo incentivo. Em especial, aos meus irmãos Anna e Hélio, pelo exemplo de força e determinação.

DEDICO

AGRADECIMENTOS

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, juntamente com seus docentes, pela oportunidade concedida para o aprendizado e realização do doutorado.

Ao Instituto Federal do Espírito Santo - Campus Alegre, pela oportunidade de crescimento, aprendizado e realização profissional.

À CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pelo suporte financeiro concedido para a realização desse trabalho.

À FAPEMIG - Fundação de Amparo à Pesquisa do Estado de Minas Gerais, por conceder apoio financeiro ao projeto de pesquisa.

Ao meu orientador professor Dr. Augusto Ramalho de Morais, pelo apoio, confiança, ensinamentos e acompanhamento desse trabalho. Agradeço muito

Ao Professor, Amigo e co-orientador, Dr. Marcelo Angelo Cirillo , pela amizade, por toda a sua disposição em colaborar, pelo apoio em todas as horas, ao qual tenho muita satisfação e orgulho de termos trabalhado juntos. Muito obrigado, as palavras são poucas para agradecer.

Aos professores do DEX - UFLA, pelos ensinamentos. Em especial ao apoio e confiança do professor Dr. Joel Augusto Muniz, que desde o mestrado se mostrou sempre atencioso e preocupado com os alunos.

Aos funcionários do DEX - UFLA, pelo carinho, boa vontade e eficiência com que sempre me atenderam. Especialmente à Josi, Selminha, Edila e Maria.

Aos professores e funcionários do Instituto Federal do Espírito Santo - Campus Alegre, pelo incentivo e companheirismo.

Aos meus colegas de turma, aos meus amigos Ana Lúcia, Augusto, Edcarlos, Manoel, Paulo, Paulo Emiliano e a todos que, de alguma forma, contribuíram para a realização desse trabalho, meu muito obrigado!!!

"O saber a gente aprende com os mestres e os livros. A sabedoria se aprende é com a vida e com os humildes."

Cora Coralina.

RESUMO

Dada a sensibilidade do método de mínimos quadrados à presença de observações discrepantes, é sabido que as estimativas de mínimos quadrados são afetadas pela presença de uma ou mais dessas observações. Frente ao exposto, esse trabalho foi realizado com o objetivo de pesquisar o efeito de observações discrepantes no ajuste de modelos de regressão multivariada. Com esse propósito, medidas de eficiência entre estimadores robustos denominados Covariância de Determinante Mínimo (Minimum Covariance Determinant - *MCD*) e Elipsoide de Volume Mínimo (Minimum Volume Ellipsoid - *MVE*), foram empregadas em simulação Monte Carlo, considerando distribuições que apresentassem desvios de simetria e excesso de curtose em relação à distribuição normal multivariada. Nesse contexto, uma nova medida foi proposta e sua validação se deu na realização de estudos de simulação Monte Carlo, assumindo diferentes configurações paramétricas, especificadas por estrutura da matriz de covariância, número de variáveis e tamanho amostral. Após a validação dos estimadores, modelos de regressão multivariada foram ajustados considerando variáveis relacionadas ao perfil sensorial e químico de genótipos de cafeeiro arábica (*Coffea arabica* L.). Os resultados relativos à simulação, indicaram que estimador *MVE* foi mais eficiente quando o efeito das observações foi proveniente de distribuições com desvio de simetria e excesso de curtose. Em se tratando da aplicação, os modelos de regressão ajustados pelos métodos *MCD* e *MVE* foram mais adequados para o estudo das variáveis físico-químicas e sensoriais, sendo condizente com os resultados experimentais.

Palavras-chave: Análise de regressão. Regressão robusta. Distância de Mahalanobis. Distância robusta.

ABSTRACT

Given the sensibility of the least squares method to the outliers presence it is known that the estimates of least squares are affected by the presence of one or more of those observations. This way, the objective of this work was to search the effect of outliers in the adjustment of models of multivariate regression. With this purpose, efficiency measures among robust estimators Minimum Covariance Determinant - MCD and Minimum Volume Ellipsoid - MVE were used in simulation Monte Carlo considering distributions to present deviations symmetry and excess kurtosis in relation to the multivariate normal distribution. In this context, a new measure was proposed and its validation made by simulation studies Monte Carlo, assuming different parametric configurations specified by structure of the covariance matrix, number of variables and sample size. After the validation of the estimators, multivariate regression models were adjusted considering variables concerning the sensorial and chemical profile of genotypes of arabica coffee (Coffea arabica L.). The relative results to the simulation indicated that MVE estimator was more efficient when the effect of the observations came from distributions with deviation symmetry and excess kurtosis. About the application, the regression models adjusted to the MCD and MVE methods were more appropriate for the study of the physiochemical and sensory variables, being consistent with the experimental results.

Key-words: Regression analysis. Robust regression. Mahalanobis distance. Robust distance

SUMÁRIO

1 INTRODUÇÃO

A análise de regressão é uma técnica estatística amplamente utilizada nos mais diversos tipos de pesquisas. Na estimação dos parâmetros de um modelo de regressão linear, em geral, utiliza-se o método de mínimos quadrados devido à facilidade de cálculos e implementação computacional.

A validação de um modelo de regressão é feita pela análise dos resíduos, sob a suposição que os resíduos são variáveis aleatórias independentes e identicamente distribuídos com distribuição normal, média nula e variância constante. No entanto, essas pressuposições nem sempre são atendidas completamente, os dados podem apresentar violações ao modelo tais como caudas mais longas ou podem ser que não venham da mesma distribuição, violando a condição de que os erros sejam identicamente distribuídos. Sendo assim, devido ao não cumprimento das pressuposições acima citadas, torna-se necessário em alguns casos, utilizar transformações nas variáveis para que a normalidade dos resíduos seja alcançada, bem como estabilizar a variância e linearizar a relação entre variáveis.

Entretanto, convém salientar que a estimação levando em consideração a transformação em um conjunto de dados que contenham possíveis observações discrepantes também requer alguns cuidados, pois em algumas situações essas observações são acomodadas pela transformação, ou seja, elas não serão discrepantes na escala transformada.

Dada a sensibilidade do método de mínimos quadrados à presença de observações discrepantes, as estimativas de mínimos quadrados podem ser completamente afetadas pela presença de uma ou mais dessas observações, que ocorrem com frequência em um conjunto de dados reais, cujos efeitos podem afetar seriamente o ajuste do modelo, de tal forma que as estimativas obtidas e suas

respectivas inferências poderão estar influenciadas. Diante do exposto, diversos testes estatísticos e técnicas de diagnósticos são utilizados para a identificação das observações atípicas.

No caso univariado, uma observação discrepante é caracterizada como aquela que não obedece a um mesmo padrão do conjunto de dados ao qual pertence e, geralmente, é identificada por uma simples inspeção visual. Em conjuntos de dados uni e bivariados, a identificação é feita com o auxílio dos gráficos de dispersão, porém em dimensões maiores, a detecção dessas observações é mais complicada, uma vez que para um número maior de variáveis, torna-se necessário considerar a possibilidade de que as observações atípicas possam se manifestar em direções diferentes das que são detectadas na plotagem de variáveis.

Uma alternativa para solucionar o problema é a utilização de métodos que não tenham suas estimativas facilmente comprometidas por tais observações, visto que sua influência no ajuste do modelo faz com que algumas técnicas de diagnósticos deixem de detectar tais observações, constituindo, assim, o efeito de mascaramento.

Na literatura, têm sido propostas algumas regras para rejeição de observações discrepantes, como promover a sua retirada do conjunto de dados antes de proceder às próximas análises. A eliminação dessas observações poderá acarretar prejuízos aos resultados, pois elas, supostamente, poderão conter informações relevantes sobre os dados e serem decisivas no conhecimento da população. Uma alternativa é a utilização de métodos robustos de estimação que contemplem o efeito dessas observações no sentido de amenizar o impacto em alguma estatística de interesse, exemplificada pelas estimativas dos parâmetros e seus erros padrão ou matrizes de covariâncias.

Tendo por base essas informações, esse trabalho foi realizado com o ob-

jetivo de pesquisar o efeito de observações discrepantes em modelos de regressão multivariada, por meio do cálculo de medidas de eficiência entre estimadores robustos, denominados Covariância de Determinante Mínimo (*Minimum Covariance Determinant - MCD*) e Elipsóide de Volume Mínimo (*Minimum Volume Ellipsoid - MVE*).

Objetivou-se também, propor uma medida de eficiência com o propósito de avaliar o desempenho dos métodos robustos, considerando o fato de que as observações discrepantes foram caracterizadas por distribuições que apresentassem desvios de simetria e excesso de curtose em relação à distribuição normal multivariada. Para que tal medida pudesse ser validada, estudos de simulação Monte Carlo foram realizados, assumindo diferentes configurações paramétricas especificadas por uma estrutura da matriz de covariância, por um número de variáveis e um tamanho amostral.

Para ilustrar a metodologia proposta, procedeu-se a uma aplicação na análise sensorial, referente ao perfil sensorial e químico de genótipos de cafeeiro arábica (*Coffea arabica* L.), na região sul do estado de Minas Gerais e na região Mogiana do estado de São Paulo.

2 REFERENCIAL TEÓRICO

A análise de regressão amplamente estudada em estatística é aplicada nas diversas áreas de pesquisa. As técnicas de regressão são utilizadas tanto na literatura como em problemas práticos. No entanto, segundo Nogueira (2007) têm sido predominantemente utilizadas em modelos univariados, em que a variável resposta é única e está associada a um conjunto de variáveis preditoras. Já, no contexto, em que a variável resposta é multivariada, poucas informações são encontradas.

A técnica clássica de análise de variância e de regressão, conforme elucidado em textos clássicos de estatística, como em Cochran e Cox (1992) e Draper e Smith (1998) entre outros, considera que na análise de um conjunto de dados, esse conjunto deve ter algumas características ou propriedades tais como a suposição de que os erros tenham distribuição de probabilidade normal. No entanto, às vezes, essas pressuposições não são satisfeitas. Isso ocorre por vários motivos e, entre esses, a presença de observações atípicas. Na realidade, não é raro tais observações - que são conhecidas em inglês como *outliers* - e doravante denominadas observações ou dados discrepantes, serem detectadas em relação ao conjunto de dados.

Define-se uma observação discrepante como aquela que se encontra longe da maior parte dos dados ou que se desvia do padrão estabelecido pela maioria. Pode ser proveniente de erros de registro, de medida ou outro tipo de erro, mas pode conter informação relevante. Merece atenção especial e não deve ser posta de lado ou descartada, sem ao menos uma prévia análise. Como essa observação pode passar despercebida e ter pouco efeito sobre a análise de regressão ou exercer grande influência sobre os estimadores dos parâmetros, distorcendo os resultados da análise e, consequentemente, as conclusões, faz-se necessária a busca por um

método de análise, que leve em conta essas considerações.

Uma alternativa factível de ser utilizada pelo pesquisador, refere-se à utilização de transformação nos dados. Entretanto, a aplicação de transformação implica na mudança de escala dos valores, o que poderá dificultar a interpretação dos resultados pelo pesquisador. Por outro lado, existe a possibilidade de se ajustarem os modelos por meio de métodos robustos, limitados a uma fração de observações discrepantes presentes na amostra. Tais métodos proporcionam as seguintes vantagens: descrição de uma estrutura que melhor se ajusta a um conjunto de dados, identificação de pontos que se desvie da maioria e que tenham grande influência sobre o restante do conjunto (HUBERT; ROUSSEEUW; AELST, 2008).

A inferência robusta preocupa-se com a construção de procedimentos que forneçam resultados confiáveis, em situações nas quais o modelo não esteja em conformidade com os dados porque esses podem apresentar algum tipo de desvio, como, por exemplo, arredondamento de observações, ocorrência de erros grosseiros e/ou registro errado de observações (HAMPEL, 1971).

2.1 Aplicações de métodos robustos na análise estatística

Na análise estatística de um conjunto de dados ou mesmo antes de se aplicação de qualquer técnica estatística recomenda-se que o pesquisador faça uma análise exploratória no conjunto de dados. Eventualmente, algumas observações poderão apresentar um comportamento discrepante em relação à maioria, em qualquer domínio de aplicação. A maioria das pesquisas científicas, industriais e econômicas lidam com grandes quantidades de variáveis, o que aumenta a possibilidade de se encontrarem dados discrepantes.

A identificação desses dados não é um processo trivial para dados univariados e isso se acentua em um contexto multidimensional. Rousseeuw e Leroy

(1987) afirmam que para dados univariados e bivariados, a localização dessas observações pode ser feita, na maioria das vezes, por uma inspeção visual, o que se torna inviável para maiores dimensões ($k > 2$, em que k é o número de variáveis regressoras). Segundo Rocke e Woodruff (1996), detectar dados atípicos em um conjunto de dados que contenham mais que uma pequena fração dessas observações têm sido impraticáveis em altas dimensões, porque quanto maior o número de variáveis de um modelo, mais difícil se torna a identificação de observações discrepantes com o uso das técnicas de regressão clássica.

Em se tratando de inferências robustas, o enfoque deve ser a obtenção de estimativas "resistentes" a uma fração razoável de observações discrepantes, com a qual, sugere-se então a formulação ou pesquisa de métodos de estimação robusta, dentre os quais podem ser citados o estimador de Mínimos Quadrados dos Resíduos Aparados (*Least Trimmed of Squares - LTS*) e o estimador Menor Mediana dos Quadrados dos Resíduos (*Least Median of Squares - LMS*) propostos por Rousseeuw (1984).

Na inferência multivariada, foram investigados os estimadores robustos do vetor de médias $(\boldsymbol{\mu})$ e matriz de covariâncias $(\boldsymbol{\Sigma})$ tais como o estimador M (MARONNA, 1976), o Elipsoide de Volume Mínimo (*Minimum Volume Ellipsoid - MVE*) e o estimador Covariância de Determinante Mínimo (*Minimum Covariance Determinant - MCD*) de Rousseeuw (1984,).

Estimadores robustos do vetor de médias e matriz de covariâncias em altas dimensões foram estudados por Rocke e Woodruff (1996). Um procedimento que utiliza propriedades de componentes principais e requer pouco esforço computacional na identificação de observações discrepantes em altas dimensões, foi apresentado por Filzmoser, Maronna e Werner (2008).

Em termos práticos, nota-se que as técnicas de análise de dados utilizando

inferência robusta, têm sido aplicados regularmente a problemas que surgem em finanças, nas ciências físicas, sociais, médicas, entre outras. Alguns estudos são destacados a seguir.

Damião (2007) examinou as características de carteiras compostas por ações e otimizadas, segundo o critério da média de retorno e matriz de covariâncias, formadas por meio de estimativas robustas de risco e retorno. Os resultados, nesse estudo, mostraram que as carteiras obtidas por meio de estimativas robustas , apresentaram melhoras em sua estabilidade e variabilidade.

Paula (2006) avaliou o uso da técnica Análise Condicionada da Demanda via regressão robusta (estimadores - MM) em contrapartida à utilização da regressão clássica, na estimação do consumo de energia elétrica por uso final do setor residencial. Os resultados de sua pesquisa mostraram que a regressão linear via estimadores robustos foi a mais indicada para esse tipo de análise.

Mendes e Leal (June 2005) propuseram uma estimativa robusta da matriz de covariâncias, comumente aplicadas em portifólios, utilizando o estimador Covariância de Determinante Mínimo no cálculo das estimativas robustas. Experimentos de simulação indicaram que a estimativa proposta apresentou bom desempenho, sob modelos Normal contaminados e sob a distribuição-t multivariada.

Cunha, Machado e Figueiredo Filho (2002) utilizaram a análise exploratória de dados e regressão robusta por meio do estimador de Mínimos Quadrados dos Resíduos Aparados, para modelar o crescimento em diâmetro e área basal de árvores sobre uma mesma área de 576 ha de floresta tropical primária, localizada na Floresta Nacional do Tapajós no Pará. A pesquisa mostrou que o uso da análise exploratória de dados e da regressão robusta viabilizou a análise e a determinação dos incrementos periódicos em diâmetro e área basal, em bases consistentes.

2.2 Modelo de regressão linear

A análise de regressão é uma das mais importantes técnicas estatísticas, utilizada em aplicações de diversas áreas. Pode-se utilizar um modelo de regressão linear múltipla para avaliar a relação entre uma variável resposta em relação a k variáveis independentes ou regressoras, representado por

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \text{ para } i = 1, ..., n \tag{2.1}$$

sendo: n o tamanho amostral, Y_i a resposta referente à i-ésima unidade amostral; $\boldsymbol{X_i} = (X_{i1}, X_{i2}, ..., X_{ik})^t$ vetor de observações das variáveis regressoras para a i-ésima unidade amostral, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^t$ vetor de coeficientes de regressão, parâmetros desconhecidos, que se quer estimar; $\boldsymbol{\varepsilon_i}$ componente de erro aleatório associado à observação Y_i, e tal que,

$$E(\boldsymbol{\varepsilon}_i) = 0; Var(\boldsymbol{\varepsilon}_i) = \boldsymbol{\sigma}^2$$

2.2.1 Estimador de mínimos quadrados dos parâmetros do modelo de regressão linear

Considerando o modelo descrito em (2.1), o método de mínimos quadrados (MMQ) pode ser utilizado para estimar os coeficientes de regressão. Com esse propósito, supondo que $n > k + 1$ observações sejam avaliadas, em que k é o número de variáveis regressoras, Y_i a i-ésima variável resposta observada, X_{ij} a i-ésima observação da j-ésima variável regressora $(i = 1, ..., n, j = 1, ..., k)$ e assumindo que $\varepsilon_i \sim N\left(0, \sigma^2\right)$, a função quadrática $S(\boldsymbol{\beta})$, que representa a soma

dos quadrados dos erros, é dada por

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_{ij} \right)^2 \tag{2.2}$$

em que a função $S(\boldsymbol{\beta})$ terá o valor crítico, quando as derivadas parciais em relação à $\beta_0, \beta_1, ..., \beta_k$ forem nulas. Derivando a função S em relação aos parâmetros

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_{ij} \right)$$

e

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_{ij} \right) X_{ij}$$

Igualando-se a zero, tem-se

$$-2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \sum_{j=1}^{k} \widehat{\beta_j} X_{ij} \right) = 0 \tag{2.3}$$

e

$$-2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \sum_{j=1}^{k} \widehat{\beta_j} X_{ij} \right) X_{ij} = 0 \tag{2.4}$$

Desenvolvendo-se as expressões (2.3) e (2.4), obtêm-se as equações de mínimos quadrados

$$n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} X_{i1} + \widehat{\beta_2} \sum_{i=1}^{n} X_{i2} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} X_{ik} = \sum_{i=1}^{n} Y_i$$

$$\widehat{\beta_0} \sum_{i=1}^{n} X_{i1} + \widehat{\beta_1} \sum_{i=1}^{n} X_{i1}^2 + \widehat{\beta_2} \sum_{i=1}^{n} X_{i1}X_{i2} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} X_{i1}X_{ik} = \sum_{i=1}^{n} X_{i1}Y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} X_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} X_{ik}X_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} X_{ik}X_{i2} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} X_{ik}^2 = \sum_{i=1}^{n} X_{ik}Y_i$$

É possível notar que há $p = k + 1$ equações, uma para cada coeficiente de regressão. Portanto, as soluções dessas equações resultam nos estimadores de mínimos quadrados (EMQ) para o modelo descrito em (2.1)

2.2.2 Propriedades dos estimadores de mínimos quadrados dos parâmetros do modelo de regressão linear

Considere o modelo linear geral (2.1), descrito em forma matricial (DRAPER; SMITH, 1998)

$$\boldsymbol{Y}_{(n\times 1)} = \boldsymbol{X}_{(n\times(k+1))}\boldsymbol{\beta}_{((k+1)\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)} \tag{2.5}$$

em que $E(\boldsymbol{\varepsilon}_i) = 0; Var(\boldsymbol{\varepsilon}_i) = I\boldsymbol{\sigma}^2$.

O método de mínimos quadrados permite determinar os estimadores de $\boldsymbol{\beta}$, considerando-se como condição que $S(\boldsymbol{\beta})$ seja mínima.

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}^t\boldsymbol{Y} - 2\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{Y} + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \tag{2.6}$$

Derivando-se parcialmente (2.6) em relação a $\boldsymbol{\beta}$ e igualando a zero, obtém-se

$$\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{Y} \tag{2.7}$$

Se a matriz $\boldsymbol{X}$ for de posto coluna completo, então $\boldsymbol{X}^t\boldsymbol{X}$ é uma matriz positiva definida e, dessa forma, é não singular. Portanto, existe a matriz inversa $\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}$

e o estimador de mínimos quadrados de $\boldsymbol{\beta}$ é obtido a partir da equação

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y} \tag{2.8}$$

Substituindo-se (2.5) em (2.8) e calculando-se a esperança, obtém-se

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon} \\ E[\hat{\boldsymbol{\beta}}] &= E\left[\boldsymbol{\beta} + \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon}\right] \\ E[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} \end{aligned} \tag{2.9}$$

De maneira semelhante, de (2.9), tem-se

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon} \tag{2.10}$$

e, usando-se o fato de que $E[\varepsilon\varepsilon^t] = \sigma^2 I$, a variância de $\hat{\boldsymbol{\beta}}$ é dada por

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\right] \\ Var(\hat{\boldsymbol{\beta}}) &= \sigma^2\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} \end{aligned} \tag{2.11}$$

Assim, a solução da equação (2.8) apresenta uma solução única, correspondente ao estimador linear não tendencioso e de variância mínima para $\boldsymbol{\beta}$.

2.3 Distribuições multivariadas

As distribuições multivariadas podem ser construídas por meio de uma distribuição univariada qualquer. Apresenta-se a seguir, sucintamente, um resumo das distribuições multivariadas normal, normal contaminada, log-normal e *t-student*.

2.3.1 Distribuição normal multivariada

Dado o vetor $\boldsymbol{X} = [X_1, \ldots, X_k]^t$ em que cada componente X_i, $i = 1, \ldots, p$, é uma variável aleatória com distribuição normal de probabilidade com média μ e variância σ^2, a distribuição conjunta destes componentes gera a distribuição normal multivariada. A distribuição de $\boldsymbol{X}$ é denotada por $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\mu}$ é um vetor de médias k-dimensional sendo $\mu_i = \mathrm{E}(X_i)$ e a matriz de covariância $\boldsymbol{\Sigma}$ de dimensão $k \times k$, sendo o (i,j)-ésimo elemento a $Cov(X_i, X_j)$, $i = 1,...,k$ e $j = 1,...,k$. Segundo Johnson e Wichern (2007) se a matriz $\boldsymbol{\Sigma}$ for singular, então a distribuição de probabilidade de $\boldsymbol{X}$ estará confinada no subespaço em $\mathbb{R}^k$. Se a matriz $\boldsymbol{\Sigma}$ tem posto completo k, então a função densidade de probabilidade de $\boldsymbol{X}$ será definida como

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right], \tag{2.12}$$

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

$$\boldsymbol{X} = A\boldsymbol{Y} + \boldsymbol{\mu} \tag{2.13}$$

sendo A qualquer matriz de dimensão $(k \times k)$ para o qual $A'A = \boldsymbol{\Sigma}$. De acordo com Johnson (1987) muitas aplicações com simulação Monte Carlo, o caso em que $\boldsymbol{\Sigma}$ tem posto completo, a densidade citada na equação 2.12 será suficiente.

2.3.2 Distribuição normal multivariada contaminada

Outra distribuição importante é a da normal multivariada contaminada. Dado o vetor aleatório $\boldsymbol{X} = [X_1, \ldots, X_k]^t \in \mathbb{R}^k$ com distribuição normal multi-

variada contaminada, sua função densidade de probabilidade será

$$f(\boldsymbol{x}) = \delta(2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right] + \\ +(1-\delta)(2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] \quad (2.14)$$

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

Segundo Johnson (1987), a geração de variáveis estatísticas a partir da equação 2.14 é fácil e pode ser realizada como a seguir:

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

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

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

O problema em usar a equação 2.14, de acordo com Johnson (1987), não está na geração da variável, mas sim na seleção dos parâmetros. O número de parâmetros na equação 2.14 é $m^2 + 3m + 1$, os quais correspondem a $2m$ médias, $2m$ variâncias, $m^2 - m$ correlações e a probabilidade δ de mistura.

No caso bivariado, a equação 2.14 pode ser escrita como

$$f(x,y) = \delta f_1(x,y) + (1-\delta)f_2(x,y), \quad (2.15)$$

em que f_i é definida com na equação 2.12, com médias μ_{i1} e μ_{i2}, desvios padrão σ_{i1} e σ_{i2}, e correlação ρ_i, $i = 1{,}2$.

2.3.3 Distribuição log-normal multivariada

Seja $\boldsymbol{X} = [X_1, X_2, \ldots, X_k]$ um vetor k-dimensional, $X \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Considere a transformação $Y_i = \exp(X_i)$ e defina $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_k]$. A densidade de Y é a distribuição log-normal multivariada e tem a seguinte forma

$$f(\boldsymbol{y}) = \boldsymbol{y}^{-1}(2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\ln \boldsymbol{y} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\ln \boldsymbol{y} - \boldsymbol{\mu})\right] \qquad (2.16)$$

em que $ln\boldsymbol{y} = [ln\boldsymbol{y_1}, ln\boldsymbol{y_2}, \ldots, ln\boldsymbol{y_k}]$ é um vetor coluna k-dimensional e $y_i = \exp(x_i)$.

2.3.4 Distribuição *t-student* multivariada

Além das anteriores, uma distribuição de igual importância é a *t-student* multivariada. Dado um vetor aleatório $\boldsymbol{X} = [X_1, \ldots, X_k]^t \in \mathbb{R}^k$ com função de densidade de probabilidade

$$f(\boldsymbol{x}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}\Gamma[(\nu + p)/2]}{[\Gamma(1/2)]^p\Gamma(\nu/2)\nu^{p/2}}\left[1 + \frac{(\boldsymbol{x} - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}} \qquad (2.17)$$

diz-se que $\boldsymbol{X}$ tem distribuição de probabilidade *t* multivariada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ e com ν graus de liberdade, com a notação $\boldsymbol{X} \sim t_k(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$. Nota-se ainda que a distribuição *t* multivariada se aproxima da distribuição normal multivariada com matriz de covariância $\boldsymbol{\Sigma}$ quando $\nu \to \infty$ (LANGE; RODERICK; TAYLOR, 1989).

2.3.4.1 Propriedades do Modelo de Regressão com distribuição Normal

Assumindo duas amostras aleatórias X_1,X_2,..., X_k e X_{k+1},X_{k+2},..., X_p que formam o vetor aleatório $\boldsymbol{X} = [X_1,...,X_p]^t = \left[\boldsymbol{X}_{(1)}^t, \boldsymbol{X}_{(2)}^t\right]^t \in \mathbb{R}^p$ em que $\boldsymbol{X}_{(1)} = [X_1,...,X_k]^t$ e $\boldsymbol{X}_{(2)} = [X_{k+1},...,X_p]^t$, a seguir são apresentadas

importantes propriedades, por meio de teoremas que podem ser provados usando resultados encontrados em Johnson e Wichern (2007).

Teorema 2.1 *(a) Se* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, e se sua matriz de covariâncias é uma diagonal, então os componentes do vetor* $\boldsymbol{X}$ *são variáveis normais independentemente distribuídas.*

(b) Seja $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *e seja uma matriz* $\boldsymbol{A}_{(p \times p)}$*. A combinação linear* $\boldsymbol{Y} = \boldsymbol{AX} \sim N_p\left(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^t\right)$*. Se* $\boldsymbol{a}_{(p \times 1)}$ *é um vetor de constantes, então* $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$

Teorema 2.2 *(a) Seja* $\boldsymbol{X} = \left[\boldsymbol{X}_{(1)}^t, \boldsymbol{X}_{(2)}^t\right]^t \in \mathbb{R}^p$*, sendo* $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_{(1)}^t, \boldsymbol{\mu}_{(2)}^t\right]^t$ *o vetor de médias particionado e a matriz de covariâncias dada por* $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11(k \times k)} & \boldsymbol{\Sigma}_{12\,(k \times (p-k))} \\ \hline \boldsymbol{\Sigma}_{21((p-k) \times k)} & \boldsymbol{\Sigma}_{22((p-k) \times (p-k))} \end{array}\right]$*. Se* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *com* $\boldsymbol{\Sigma}_{12} = 0$*, então* $\boldsymbol{X}_{(1)} \sim N_k\left(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11}\right)$ *e* $\boldsymbol{X}_{(2)} \sim N_{(p-k)}\left(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma}_{22}\right)$ *são independentementes distribuídos.*

(b) Se $\boldsymbol{X}_{(1)} \sim N_k\left(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11}\right)$ *e* $\boldsymbol{X}_{(2)} \sim N_{(p-k)}\left(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma}_{22}\right)$ *são independentes, então*

$$\begin{pmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right)$$

Teorema 2.3 *Seja um vetor aleatório* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, então a distribuição condicional de* $\boldsymbol{X}_{(1)}$ *dado que* $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$*, é normal p-variada com média* $\boldsymbol{\mu}_1^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ *e covariância* $\boldsymbol{\Sigma}_{11}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

A distribuição condicional de qualquer subvetor de variáveis aleatórias normais multivariadas dado os valores do subvetor remanescente é normal multivariada, com média e covariância especificadas no teorema (2.3) supondo que

todas as variáveis Y, X_1, X_2,..., X_k na seção (2.2) sejam aleatórias formando o vetor $\left[Y,\boldsymbol{X}^t\right]^t \in \mathbb{R}^{(k+1)}$, de forma que

$$\boldsymbol{\mu} = \left[\begin{array}{c} \mu_{Y(1\times1)} \\ \hline \boldsymbol{\mu}_{X(k\times1)} \end{array}\right] \quad e \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \Sigma_{YY(1\times1)} & \boldsymbol{\Sigma}_{YX\,(1\times k)} \\ \hline \boldsymbol{\Sigma}_{XY(k\times1)} & \boldsymbol{\Sigma}_{XX(k\times k)} \end{array}\right] \tag{2.18}$$

Tendo por base os resultados mencionados na equação (2.6) e no teorema (2.3)

$$E\left(Y|\boldsymbol{X}\right) = \mu_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}_X\right) \tag{2.19}$$

Os estimadores de mínimos quadrados para $\boldsymbol{\beta}, \beta_0,$ e $S\left(\boldsymbol{\beta}\right)$ podem ser escritos da seguinte forma

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\Sigma}}_{XX}^{-1}\widehat{\boldsymbol{\Sigma}}_{XY} \\ \widehat{\beta}_0 &= \widehat{\mu}_Y - \widehat{\boldsymbol{\beta}}^{\boldsymbol{t}}\widehat{\boldsymbol{\mu}}_X \\ S\left(\widehat{\boldsymbol{\beta}}\right) &= \widehat{\boldsymbol{\Sigma}}_{YY} - \widehat{\boldsymbol{\beta}}^{\boldsymbol{t}}\widehat{\boldsymbol{\Sigma}}_{XX}\widehat{\boldsymbol{\beta}} \end{aligned} \tag{2.20}$$

2.4 Modelo de regressão multivariada

O modelo de regressão multivariada generaliza o modelo linear de regressão, no sentido de que na realização de um experimento, mensuram-se, de forma independente em cada unidade amostral, n indivíduos em relação a p variáveis respostas. Nesse caso, o vetor n-dimensional das observações da j-ésima variável resposta é da forma $\boldsymbol{Y}_j = [Y_{1j},...,Y_{ij},....,Y_{nj}]^t$, em que Y_{ij} é a i-ésima observação da j-ésima variável resposta, sendo $i = 1,2,...,n$ e $j = 1,2,...,p$.

A cada um desses p vetores está associado um grupo de variáveis regres-

soras, em geral não aleatórias. Assim, pode-se entender que o vetor $\boldsymbol{Y_1}$ tem um modelo linear de regressão descrito por Johnson e Wichern (2007), com as seguintes equações

$$\begin{aligned} Y_{11} &= \beta_{01} + \beta_{11}X_{11} + ... + \beta_{k1}X_{1k} + \varepsilon_{11} \\ Y_{21} &= \beta_{01} + \beta_{11}X_{21} + ... + \beta_{k1}X_{2k} + \varepsilon_{21} \\ &\vdots \\ Y_{n1} &= \beta_{01} + \beta_{11}X_{n1} + ... + \beta_{k1}X_{nk} + \varepsilon_{n1} \end{aligned}$$

Mediante o exposto, nota-se que o modelo linear geral pode ser representado por

$$\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \tag{2.21}$$

em que, $\boldsymbol{Y}_j$ corresponde a um vetor n-dimensional, contendo os valores da j-ésima variável resposta para os n elementos da amostra; $\boldsymbol{X}$ é a matriz $n \times (k+1)$ de observações das variáveis regressoras associadas à j-ésima variável resposta, $\boldsymbol{\beta}_j$ vetor $(k+1) \times 1$ de coeficientes de regressão, parâmetros desconhecidos, que se quer estimar; $\boldsymbol{\varepsilon}_j$ vetor n-dimensional, contendo os componentes do erro aleatório associado à observação Y_j.

Em notação matricial, a matriz de variáveis respostas pode ser representada por

$$\boldsymbol{Y}_{(n\times p)} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2j} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ Y_{i1} & Y_{i2} & \ldots & Y_{ij} & \ldots & Y_{ip} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nj} & \cdots & Y_{np} \end{bmatrix} = \left[\boldsymbol{Y}_1 \vdots \boldsymbol{Y}_2 \vdots \cdots \vdots \boldsymbol{Y}_j \vdots \cdots \vdots \boldsymbol{Y}_p\right]$$

sendo que cada linha de $\boldsymbol{Y}$ conterá as p variáveis respostas para uma determinada unidade amostral, enquanto cada coluna, um vetor n-dimensional das observações de uma determinada variável resposta. Nesse contexto, a matriz $\boldsymbol{Y}$ será escrita, de forma alternativa, como $[\boldsymbol{Y}_1,\boldsymbol{Y}_2,...,\boldsymbol{Y}_n]^t$, na qual a amostra aleatória corresponderá a $\boldsymbol{Y}_1,\boldsymbol{Y}_2,...,\boldsymbol{Y}_n$. Nessa situação, o vetor p-dimensional da i-ésima unidade amostral pode ser escrito como

$$\boldsymbol{Y}_i = [Y_{i1},Y_{i2},...,Y_{ip}]^t$$

em que os vetores $\boldsymbol{Y}_i \quad (i = 1,2,...,n)$, correspondentes às linhas de $\boldsymbol{Y}$, são independentes. Dessa forma, $\boldsymbol{Y}_i \sim N_p\left(\boldsymbol{\beta}^t\boldsymbol{X}_i,\boldsymbol{\Sigma}\right)$ com $\boldsymbol{X}_i$ sendo descrito como a i-ésima linha de $\boldsymbol{X}$, e esta, a matriz de variáveis regressoras (ou de planejamento) dada por

$$\boldsymbol{X}_{(n\times(k+1))} = \begin{bmatrix} X_{10} & X_{11} & \cdots & X_{1k} \\ X_{20} & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i0} & X_{i1} & \ldots & X_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & \cdots & X_{nk} \end{bmatrix} \tag{2.22}$$

em que $X_{i0} = 1$ para todo i.

Em conformidade com a dimensão da matriz de planejamento, a matriz de parâmetros pode ser escrita da seguinte forma

$$\boldsymbol{\beta}_{((k+1)\times p)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{i1} & \beta_{i2} & \cdots & \beta_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kp} \end{bmatrix} = \left[\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2 \vdots \cdots \vdots \boldsymbol{\beta}_p\right]$$

com $\boldsymbol{\beta}_j = [\beta_{0j},\beta_{1j},...,\beta_{kj}]^t$ e a matriz dos erros não observáveis é representada por

$$\boldsymbol{\varepsilon}_{(n\times p)} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix} = \left[\boldsymbol{\varepsilon}_1 \vdots \boldsymbol{\varepsilon}_2 \vdots \cdots \vdots \boldsymbol{\varepsilon}_p\right] \tag{2.23}$$

com $\boldsymbol{\varepsilon}_j = [\varepsilon_{1j},...,\varepsilon_{nj}]^t$.

Os elementos de $\boldsymbol{\beta}$ podem ser estimados separadamente. No entanto, devido à correlação existente entre as p variáveis respostas, a estimação conjunta dos parâmetros torna-se mais adequada. Assim, o modelo de regressão multivariada pode ser descrito por

$$\boldsymbol{Y}_{(n\times p)} = \boldsymbol{X}_{(n\times(k+1))}\boldsymbol{\beta}_{((k+1)\times p)} + \boldsymbol{\varepsilon}_{(n\times p)} \tag{2.24}$$

$E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_{\varepsilon}$.

Convém ressaltar que as p observações na i-ésima unidade amostral têm

matriz de covariância $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_{ik}\}$, mas observações de diferentes ensaios são não correlacionadas.

2.4.1 Estimador de mínimos quadrados dos parâmetros do modelo de regressão multivariada

Consoante às seções anteriores, considere o caso em que se dispõe de p variáveis respostas Y_1,...,Y_p e k variáveis explicativas X_1,..., X_k, supondo-se que todas as variáveis Y_1,...,Y_p,X_1,...,X_k formam o vetor aleatório, $\left[\boldsymbol{Y}^t,\boldsymbol{X}^t\right]^t \in \mathbb{R}^{(p+k)}$. O modelo de regressão multivariada, apresentado em Rousseeuw *et al.* (2004), é dado por

$$\boldsymbol{Y}_{(p\times 1)} = \boldsymbol{\beta}^t_{(k\times p)}\boldsymbol{X}_{(p\times 1)} + \boldsymbol{\beta}_{0(p\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)} \tag{2.25}$$

Seja $\boldsymbol{\mu}$ o vetor de médias e a matriz de covariâncias $\boldsymbol{\Sigma}$ particionados de forma que

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{Y(p\times 1)} \\ \hline \boldsymbol{\mu}_{X(k\times 1)} \end{array}\right] \text{ e } \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{YY(p\times p)} & \boldsymbol{\Sigma}_{YX\,(p\times k)} \\ \hline \boldsymbol{\Sigma}_{XY(k\times p)} & \boldsymbol{\Sigma}_{XX(k\times k)} \end{array}\right] \tag{2.26}$$

em que $\boldsymbol{\mu}_Y$ é o vetor de médias de Y, $\boldsymbol{\mu}_X$ é o vetor de médias de X; $\boldsymbol{\Sigma}_{YY}$ é a matriz de covariâncias de Y, $\boldsymbol{\Sigma}_{XY}$ é a matriz de covariâncias entre X e Y e $\boldsymbol{\Sigma}_{XX}$ é a matriz de covariâncias de X.

Tendo por base os resultados apresentados na seção (2.2.1), verifica-se que

$$E\left(\boldsymbol{Y}|X_1,\ldots,X_k\right) = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}^{-1}_{XX}\left(\boldsymbol{X} - \boldsymbol{\mu}_X\right) \tag{2.27}$$

em que $E\left(\boldsymbol{Y}|X_1,\ldots,X_k\right)$ é a esperança para o vetor de variáveis resposta, dado as k variáveis explicativas. Entende-se também que esta esperança condicional é

composta de p regressões univariadas, sendo o primeiro componente do vetor de médias condicional dado pela equação (2.19).

Os estimadores de mínimos quadrados $\boldsymbol{\beta}, \boldsymbol{\beta}_0$ e $\boldsymbol{\Sigma}_\varepsilon$ podem ser escritos como função dos componentes dos vetores em (2.26), da seguinte forma

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\Sigma}}_{XX}^{-1} \widehat{\boldsymbol{\Sigma}}_{XY} \\ \widehat{\boldsymbol{\beta}}_0 &= \widehat{\boldsymbol{\mu}}_Y - \widehat{\boldsymbol{\beta}}^{\boldsymbol{t}} \widehat{\boldsymbol{\mu}}_X \\ \widehat{\boldsymbol{\Sigma}}_\varepsilon &= \widehat{\boldsymbol{\Sigma}}_{YY} - \widehat{\boldsymbol{\beta}}^{\boldsymbol{t}} \widehat{\boldsymbol{\Sigma}}_{XX} \widehat{\boldsymbol{\beta}} \end{aligned} \tag{2.28}$$

Olive (2008, p.309-311) , apresenta uma demonstração para esses resultados.

2.5 Principais conceitos de robustez

Segundo Damião (2007) A estatística robusta pode ser descrita como uma generalização da estatística clássica que leva em consideração a possibilidade de especificações incorretas do modelo e da distribuição dos dados em estudo. Essa teoria e seus resultados são válidos tanto dentro do modelo especificado como nas proximidades dele, nesse caso, por exemplo, quando a amostra em estudo está contaminada por observações discrepantes.

2.5.1 Ponto de ruptura

A noção de ponto de ruptura foi apresentada de forma concisa, formalmente definida e discutida por Hampel *et al.* (1986). Desde então, tornou-se uma medida bastante utilizada nas pesquisas sobre estatística robusta. Originalmente, o ponto de ruptura é um conceito assintótico, mas que, no entanto, esse conceito foi adaptado para amostras finitas por Donoho e Huber ().

O ponto de ruptura $(\delta_n^*(T, Z))$ de um estimador para amostras finitas é

definido, informalmente, como a menor fração de contaminação da amostra por observações discrepantes, a qual possa comprometer a qualidade ou veracidade das estimativas a serem obtidas, ou seja, a maior fração de contaminação que um estimador possa suportar e, ainda assim, fornecer informação confiável sobre o parâmetro considerado.

Considere um conjunto de dados, que se relacionam segundo o modelo especificado em (2.5), $\boldsymbol{Z} = \{(X_{11},...,X_{1k}, Y_1), ..., (X_{n1},..., X_{nk}, Y_n)\}$. Sendo T um estimador de regressão e aplicando T à amostra Z, isso produz um vetor de coeficientes de regressão $T(Z) = \widehat{\boldsymbol{\beta}}$ em que $\widehat{\boldsymbol{\beta}} = (\beta_1,..., \beta_k)$.

Defina como $\boldsymbol{Z}^*$ o conjunto de todas as possíveis amostras contaminadas, obtidas pela substituição de m observações quaisquer dos dados originais por valores discrepantes. Então, a fração de contaminação será dada por $\delta = \dfrac{m}{n}$.

A diferença entre o valor esperado e o valor verdadeiro, conhecida como vício, pode ser expressa por $b(m; T, Z)$. O vício máximo causado por alguma contaminação δ pode ser definido por

$$b(m; T, Z) = \sup_{\boldsymbol{Z^*}} \|T(Z^*) - T(Z)\|$$

em que o supremo é calculado em relação a todos os possíveis conjuntos $\boldsymbol{Z}^*$ resultantes da substituição de m observações do conjunto de dados $\boldsymbol{Z}$. Se $b(m; T, Z)$ é finito, os m valores discrepantes podem produzir um efeito arbitrário em T, de forma que aconteça uma "ruptura" para uma dada quantidade deles. Dessa forma, o ponto de ruptura do estimador T é definido por

$$\delta_n^* (T, \boldsymbol{Z}) = \min_{1 \leqslant m \leqslant n} \{b(m; T, Z) = \infty\} \tag{2.29}$$

Exemplificando para a média amostral, considerando uma amostra finita,

verifica-se que uma única observação discrepante afeta fortemente a estimativa de mínimos quadrados, pois se uma única observação assumir um valor tendendo ao infinito, a média amostral tenderá ao infinito. Assim, nesse contexto, o ponto de ruptura desse estimador é dado por

$$\delta_n^* \left(\widehat{\boldsymbol{\mu}}, \boldsymbol{X}\right) = \frac{1}{n} \quad e \quad \lim_{n \to \infty} \frac{1}{n} = 0 \tag{2.30}$$

Então, pode-se afirmar que o método dos mínimos quadrados possui um ponto de ruptura de 0%, ou seja, apenas uma observação afastada da distribuição principal já altera o estimador T.

De forma similar, o ponto de ruptura relacionado ao estimador da matriz de covariância $\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{X}\right)$ é definido como a menor fração de observações discrepantes que possa fazer com que o maior autovalor $\lambda_1\left(\widehat{\boldsymbol{\Sigma}}\right)$ seja próximo do infinito ou com que o menor autovalor $\lambda_k\left(\widehat{\boldsymbol{\Sigma}}\right)$ autovalor assuma valores próximos de zero (HUBERT; ROUSSEEUW; AELST, 2008).

2.5.2 Função influência

A função influência (*IF*) mede o efeito no estimador de uma contaminação infinitesimal em uma observação qualquer na amostra e nos dá a idéia de como ficaria um estimador sob contaminações pontuais.

Pode-se definir a distribuição G, seja $T(G)$ um estimador de regressão, a partir da mistura de duas distribuições (F e H) da seguinte forma

$$G = (1 - \delta)F + \delta H, \quad \text{sendo} \quad \delta \in [0, 1] \tag{2.31}$$

A distribuição G pode ser considerada uma distribuição mista entre a distribuição principal F e uma contaminação dada pela distribuição H, com uma probabilidade

δ, que produz observações, com elevada probabilidade $1 - \delta$ da distribuição F, contaminadas por observações da distribuição H, com pequena probabilidade δ. No caso, um estimador será dito robusto, se permanecer estável no conjunto de distribuições G formadas a partir de F.

Um caso particular da função G ocorre quando Δ_h é uma distribuição na qual o valor h ocorre com probabilidade 1. Dessa forma, se X possui uma distribuição Δ_h, então $P(X \leq x) = 0$ se $x < h$, e a média de X é $E(X) = h$. Portanto, a distribuição resultante $G_{h,\delta}$ será um modelo simplificado, para uma situação de contaminação de uma amostra por uma proporção δ de observações discrepantes em h. Em que

$$G_{h,\delta} = (1 - \delta)F + \delta\Delta_h$$

O interesse em definir a função $G_{h,\delta}$ é o de observar como o valor h afeta o valor de uma função ou estimador da distribuição F quando h ocorre com probabilidade δ. Pode-se notar que quando δ é suficientemente pequeno, as distribuições $G_{h,\delta}$ e F são bastante semelhantes. A influência relativa do valor h em um estimador $T(F)$, é dada por

$$\frac{T\left(G_{h,\delta}\right) - T\left(F\right)}{\delta} \tag{2.32}$$

A função influência é a influência relativa de h em um estimador $T(F)$, quando a contaminação por h, ou seja, a probabilidade de contaminação por h tende a zero, é dada por

$$IF = \lim_{\delta \to 0} \frac{T\left(G_{h,\delta}\right) - T\left(F\right)}{\delta} \tag{2.33}$$

nos pontos h em que o limite existe e, ainda, mede o efeito que uma pequena fração

de contaminação por observações discrepantes no ponto h provoca no estimador.

2.5.3 Equivariância

A propriedade de equivariância é importante no âmbito de estudos aplicados e segundo Olive (2008) tem recebido considerável atenção na literatura, visto que essa propriedade envolve transformação nos dados, permitindo que a escala da variável original possa ser alterada, sem que haja perda de coerência nas conclusões, baseadas nos resultados estimados por regressão. Além do mais, conhecendo-se as propriedades de equivariância e invariância, sob transformações lineares nas variáveis, torna-se possível encontrar limites superiores para o ponto de ruptura dos estimadores que possuem essas propriedades.

O trabalho de Koenker (2005) ilustra a propriedade de equivariância com um exemplo de um modelo que analisa a temperatura de um líquido, Y, em relação às variáveis regressoras X. Uma vez que as medidas de temperatura de Y estão mensuradas em graus Fahrenheit, seria mais natural mudar a escala para graus Celsius, pois é a unidade de medida mais usual. Assim, pela propriedade de equivariância, efetuando a mudança de grau Fahrenheit para grau Celsius, espera-se que as estimativas de regressão mudem, porque mudou a escala, mas que sua interpretação permaneça invariante.

Mediante ao exposto sobre essas propriedades, são enunciadas as seguintes suposições: considerando que $\boldsymbol{X}$ e $\boldsymbol{Y}$ representam os dados originais e que se relacionam segundo o modelo (2.24). Então $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})$ é o vetor de estimativas dos coeficientes de regressão; $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})$ é o vetor das respostas estimadas e $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ é o vetor de resíduos.

Seja $\boldsymbol{b}$ um vetor de constantes, k-dimensional e a transformação nas variáveis de forma que $\boldsymbol{W} = \boldsymbol{X}$ e $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{b}$, afirma-se que um estimador $\widehat{\boldsymbol{\beta}}$ é

equivariante de regressão se

$$\widehat{\boldsymbol{\beta}}\left(\boldsymbol{W},\boldsymbol{Z}\right)=\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X},\boldsymbol{Y}+\boldsymbol{X}\boldsymbol{b}\right)=\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X},\boldsymbol{Y}\right)+\boldsymbol{b} \qquad \forall\ \boldsymbol{b}\in\mathbb{R}^{k} \tag{2.34}$$

Observe que $\widehat{\boldsymbol{Z}}=\widehat{\boldsymbol{Y}}+\boldsymbol{X}\boldsymbol{b}$ e $\boldsymbol{r}\left(\boldsymbol{W},\boldsymbol{Z}\right)=\boldsymbol{Z}-\widehat{\boldsymbol{Z}}=\boldsymbol{r}\left(\boldsymbol{X},\boldsymbol{Y}\right)$. Dessa maneira, os resíduos são invariantes acerca desse tipo de transformação.

Considere a transformação nos dados, dada por $\boldsymbol{W}=\boldsymbol{X}$ e $\boldsymbol{Z}=c\boldsymbol{Y}$, sendo c um escalar qualquer, então, um estimador $\widehat{\boldsymbol{\beta}}$ é equivariante de escala se

$$\widehat{\boldsymbol{\beta}}\left(\boldsymbol{W},\boldsymbol{Z}\right)=\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X},c\boldsymbol{Y}\right)=c\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X},\boldsymbol{Y}\right) \qquad \forall\ \boldsymbol{b}\in\mathbb{R}^{k} \tag{2.35}$$

Isso implica que o ajuste é independente da escolha da unidade de medida da variável resposta, pois $\widehat{\boldsymbol{Z}}=c\widehat{\boldsymbol{Y}}$ e $\boldsymbol{r}\left(\boldsymbol{W},\boldsymbol{Z}\right)=\boldsymbol{Z}-\widehat{\boldsymbol{Z}}=c\boldsymbol{r}\left(\boldsymbol{X},\boldsymbol{Y}\right)$ favorecendo essa interpretação.

Considere a transformação nos dados, de forma que $\boldsymbol{W}=\boldsymbol{X}\boldsymbol{A}$ e $\boldsymbol{Z}=\boldsymbol{Y}$, definindo $\boldsymbol{A}_{k\times k}$ uma matriz qualquer não singular, diz-se que um estimador $\widehat{\boldsymbol{\beta}}$ é afim equivariante se

$$\widehat{\boldsymbol{\beta}}\left(\boldsymbol{W},\boldsymbol{Z}\right)=\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X}\boldsymbol{A},\boldsymbol{Y}\right)=\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}\left(\boldsymbol{X},\boldsymbol{Y}\right) \tag{2.36}$$

Em se tratando dos estimadores robustos do vetor de médias $(\boldsymbol{\mu})$ e matriz de covariâncias $(\boldsymbol{\Sigma})$, assuma que X representa uma matriz dos dados que foram observados com $\boldsymbol{X}_{i}$, sendo descrito como a i-ésima linha de $\boldsymbol{X}$. Considere a transformação linear $\boldsymbol{Z}=\boldsymbol{W}\boldsymbol{A}+\boldsymbol{B}$, definindo $\boldsymbol{A}_{k\times k}$ como uma matriz qualquer não singular e $\boldsymbol{B}=\boldsymbol{1}\boldsymbol{b}^{t}$, em que $\boldsymbol{1_{n}}$ é um vetor n-dimensional e $\boldsymbol{b}$ é um vetor de constantes k-dimensional .

Define-se $\widehat{\boldsymbol{\mu}}\left(\boldsymbol{Z}\right)\in\mathbb{R}^{k}$ como um estimador de localização afim equivari-

ante, se e somente se, para qualquer matriz $\boldsymbol{A}_{k\times k}$ não singular

$$\widehat{\boldsymbol{\mu}}\left(\boldsymbol{X}_1\boldsymbol{A}+\boldsymbol{b}^t,...,\boldsymbol{X}_n\boldsymbol{A}+\boldsymbol{b}^t\right)=\widehat{\boldsymbol{\mu}}\left(\boldsymbol{X}_1,...,\boldsymbol{X}_n\right)\boldsymbol{A}+\boldsymbol{b}^t \qquad \forall\ \boldsymbol{b}\in\mathbb{R}^k \tag{2.37}$$

Diz-se que $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$ será um estimador equivariante de escala, se e somente se,

$$\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{X}_1\boldsymbol{A}+\boldsymbol{b},...,\boldsymbol{X}_n\boldsymbol{A}+\boldsymbol{b}\right)=\boldsymbol{A}^t\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{X}_1,...,\boldsymbol{X}_n\right)\boldsymbol{A} \tag{2.38}$$

Note que se um estimador é afim equivariante, as estimativas para $\boldsymbol{Y}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,...,\boldsymbol{Y}_n)$ com $\boldsymbol{Y_i}=\boldsymbol{A}\boldsymbol{X_i}^t+\boldsymbol{b}$, permanecem inalteradas. Essa propriedade permite-nos mudar o sistema de coordenadas para os preditores, sem afetar as previsões.

2.5.4 Conjuntos elementares

Segundo Machado (1997), o algoritmo de reamostragem dos conjuntos elementares foi proposto por Theil (1950) e consiste, basicamente, na retirada sem reposição de todas os subconjuntos possíveis (subamostras), de tamanho k de um conjunto de dados. O tamanho deve ser o mínimo suficiente para que os parâmetros do modelo possam ser estimados.

Uma das vantagens desse algoritmo é a possibilidade de ser utilizado em métodos que são resistentes a observações atípicas, já que foram propostos como um método computacional, com o objetivo de calcular estimadores para a média e matriz de covariância com alto ponto de ruptura.

Seja o conjunto de dados $\boldsymbol{Z}_{n\times(k+p)}=(\boldsymbol{X}_{n\times k}\ \ \boldsymbol{Y}_{n\times p})$ que se relacionam

segundo o modelo apresentado em (2.24). Particionando a matriz Z de forma que

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X_J} & \boldsymbol{Y_J} \\ \boldsymbol{X_{n-J}} & \boldsymbol{Y_{n-J}} \end{pmatrix} \tag{2.39}$$

em que $\boldsymbol{Y_J}$ e $\boldsymbol{X_J}$ representam respectivamente, a submatriz, $k \times p$, da variável resposta e a submatriz, de ordem k, das variáveis preditoras referentes às observações desse conjunto.

Considere $S = \{\{1,...,k\},...,\{j_1,...,j_k\},...,\{(n-k+p),...,n\}\}$ o conjunto de todos possíveis subconjuntos de observações que se podem obter com n e k fixados, em que $\boldsymbol{J} = \{j_1,...,j_k\}$ representa um subconjunto de índices com k observações dos dados originais. Cada subconjunto $\boldsymbol{Z}_J = \begin{pmatrix} \boldsymbol{X_J} & \boldsymbol{Y_J} \end{pmatrix}$, $\boldsymbol{J} \subset S$ é denominado conjunto elementar, em que

$$\boldsymbol{Z}_J = \begin{pmatrix} X_{J_1 1} & \cdots & X_{J_1 k} & Y_{J_1 1} & \cdots & Y_{J_1 p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{J_k 1} & \cdots & X_{J_k k} & Y_{J_k 1} & \cdots & Y_{J_k p} \end{pmatrix} \tag{2.40}$$

Diversos autores, dentre os quais Rousseeuw (1984), utilizaram essa metodologia no método de estimação menor mediana dos quadrados dos resíduos, Rousseeuw *et al.* (2004) também a utilizaram para o cálculo de estimadores multivariados da média e matriz de covariância.

2.6 Métodos robustos de estimação da matriz de covariâncias

2.6.1 Regressão covariância de determinante mínimo

A medida de dispersão nos dados multivariados é baseada na estrutura de covariância que é registrada em uma matriz de covariâncias $\boldsymbol{\Sigma}$. São duas as medidas de dispersão multivariada mais conhecidas, mencionadas por Herwindiati

e Isa (2009), sendo essas a variância total (VT) e a variância generalizada (VG). Em se tratando da estimação de uma variância generalizada, representada pelo determinante da matriz de covariância, utiliza-se o método *MCD* (covariância de determinante mínimo) com o propósito de estimar o vetor de localização $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$, obtidos por mínimos quadrados nas expressões dadas na equação (2.18), que apresente o menor determinante entre todos subconjuntos pesquisados. Portanto, nota-se que este método contempla aspectos da inferência robusta com propriedade de apresentar elevado ponto de ruptura (CROUX; HAESBROECK, 1999).

Contextualizando os modelos de regressão multivariada, o procedimento a ser adotado para garantir as propriedades de equivariância exigidas em um estimador de regressão multivariada é dado considerando um conjunto de dados $\boldsymbol{X} = (\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_n)$. O método de regressão *MCD*, com o objetivo de minimizar o determinante da estimativa de covariância que é determinada por $\widehat{\boldsymbol{\Sigma}}\,(\boldsymbol{X})$, investigará o subconjunto $(X_{i1},...,X_{ih})$ de tamanho h, impondo a restrição que $\left(\dfrac{n+k+1}{2}\right) \leqslant h \leqslant n$. No caso de pequenas amostras, adota-se um fator de correção recomendado por Pison, Aelst e Willems (Apr. 2002).

Com as especificações de h mencionadas, Rousseeuw *et al.* (2004) garantem que as estimativas *MCD* apresentarão maior ponto de ruptura possível $(\delta^* = 50\%)$. Entretanto, se considerarmos $\delta_n^*(\widehat{\boldsymbol{\mu}}) = \delta_n^*(\widehat{\boldsymbol{\Sigma}}) \cong \left(\dfrac{n-h}{n}\right)$ o ponto de ruptura terá um valor dado entre $0 \leq \delta^* \leq 0{,}5$. Convém ressaltar que para muitos estimadores, $\boldsymbol{\delta_n^*}\,(\widehat{\boldsymbol{\mu}},\boldsymbol{X})$ varia ligeiramente apenas com $\boldsymbol{X}$ e n, o que de certa forma condiz denotar esse limite por $\delta^*(\widehat{\boldsymbol{\mu}})$, quando $n \to \infty$.

Tendo como suporte teórico o teorema *C-step*, Rousseeuw e Driessen 1999) asseguram que dado um conjunto de h observações fornecendo as estimativas $\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{k}},\widehat{\boldsymbol{\Sigma}}_k\right)$. Quando as estimativas $\left(\widehat{\boldsymbol{\mu}}_{k+1},\widehat{\boldsymbol{\Sigma}}_{k+1}\right)$ são computadas com as

h primeiras observações, ordenadas de acordo com as distâncias de Mahalanobis $d_i\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{k}},\widehat{\boldsymbol{\Sigma}}_k\right)$ para $i = 1,...,n$ então $\det\left(\widehat{\boldsymbol{\Sigma}}_{k+1}\right) \leqslant \det\left(\widehat{\boldsymbol{\Sigma}}_k\right)$. A principal vantagem desse teorema é que ele permite obter uma seqüência de matrizes, cujo determinante é menor do que o determinante da matriz de partida.

2.6.2 Elipsoide de volume mínimo

Utiliza-se o método elipsoide de volume mínimo (*Minimum Volume Ellipsoid - MVE*) proposto por Rousseeuw () com o propósito de estimar o vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Esse estimador possui ponto de ruptura $0{,}5$, é equivariante e definido como o Elipsoide de menor volume, cobrindo pelo menos h pontos de $\boldsymbol{X}$. O *MVE* investigará o subconjunto $(X_{i1},...,X_{ih})$ de tamanho $h = \left[\frac{n}{2}\right] + 1$ que apresente o Elipsoide de menor volume entre todos subconjuntos pesquisados, em que $\left[\frac{n}{2}\right]$ é o maior inteiro $\leq \frac{n}{2}$.

Rousseeuw e Zomeren (1990) assumiram $h = \left(\frac{n+k+1}{2}\right)$ ao utilizarem o estimador *MVE* no cálculo de distâncias robustas, de forma que o vetor de médias $\boldsymbol{\mu}$ foi estimado como o centro do Elipsoide e a matriz de covariâncias $\boldsymbol{\Sigma}$ como o próprio Elipsoide multiplicado por um fator de correção para pequenas amostras.

2.6.3 Vetor mínima variância

O método vetor mínima variância (*Minimum Vector Variance - MVV*), foi proposto por Herwindiati, Djauhari e Mashuri (2007) para detectar observações discrepantes multivariadas. A distância robusta de Mahalanobis é calculada a partir do estimador *MVV*, utilizando-se um algoritmo similar ao usado no método de regressão *MCD*. Em sua execução, a distância de Mahalanobis é calculada pelo critério do vetor de variância mínima (VV), definido como $tr\left(\boldsymbol{\Sigma}^2\right)$ e introduzido

por Djauhari (2005. 1 CD-ROM.), ao invés do critério da variância generalizada VG, definida como $|\mathbf{\Sigma}|$.

Uma relação entre variância total VT, variância generalizada VG e vetor variância VV pode ser descrita supondo que $\boldsymbol{x}$ seja um vetor aleatório da matriz de covariâncias $\mathbf{\Sigma}$, de ordem k, em que $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_k \geqslant 0$ são os autovalores de $\mathbf{\Sigma}$. Então,

$$VT = tr\left(\mathbf{\Sigma}\right) = \lambda_1 + \lambda_2 + ... + \lambda_k \tag{2.41}$$

$$VG = |\mathbf{\Sigma}| = \lambda_1 \cdot \lambda_2 \cdot ... \cdot \lambda_k \tag{2.42}$$

$$VV = tr\left(\mathbf{\Sigma}^2\right) = \lambda_1^2 + \lambda_2^2 + ... + \lambda_k^2 \tag{2.43}$$

Herwindiati, Djauhari e Mashuri (2007) destacam que o método *MVV*, ao utilizar o $tr\left(\mathbf{\Sigma}^2\right)$, destaca-se em relação aos métodos *MCD* e *MVE*, pelo fato de poder ser usado mesmo no caso em que $\mathbf{\Sigma}$ seja singular e, nesse caso, $|\mathbf{\Sigma}| = 0$. Enquanto uma condição necessária e suficiente para a existência da inversa de $\mathbf{\Sigma}$ é que $|\mathbf{\Sigma}| \neq 0$, restrigindo-se então, o uso da variância generalizada.

2.7 Identificação de observações discrepantes

Barnett e Lewis (1994) definiram uma observação discrepante em um conjunto de dados como sendo aquela que parece ser inconsistente ao conjunto de dados remanescentes.

Uma observação é denominada discrepante na direção y ou *outlier* de regressão, se ela se "afasta do padrão linear" definido pelas outras ou pela maioria das outras observações. Por outro lado, quando uma observação se destaca das

demais, no espaço das variáveis independentes, é denominada ponto de alavanca.

Considerando uma matriz de dados $\boldsymbol{X}$ e $\boldsymbol{X}_i$, $i = 1, ..., n$, sendo descrito como a observação da i-ésima unidade amostral, um método clássico utilizado na identificação de observações que apresentam um comportamento discrepante do padrão seguido pela maioria, consiste em calcular para cada observação $\boldsymbol{X}_i$ a distância de Mahalanobis d_i computada por

$$d_i = \sqrt{(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}(\boldsymbol{X}))^t \widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})^{-1} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}(\boldsymbol{X}))} \tag{2.44}$$

em que $\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) = \dfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ é um estimador do vetor, k-dimensional, de médias amostral, e $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}) = \dfrac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}(\boldsymbol{X})).(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}(\boldsymbol{X}))^t$ é o estimador da matriz, de ordem k, de covariância amostral.

De modo geral, o valor crítico que permite classificar se uma observação é discrepante é definido como um ponto que apresente uma distância d_i da forma

$$d_i \geqslant \sqrt{\chi^2_{k;(1-\alpha)}} \tag{2.45}$$

sendo k o número de variáveis. Para essas distâncias, os estimadores são obtidos pelo método de mínimos quadrados, dessa forma, segundo Johnson (1987) as estimativas podem sofrer mascaramento, sendo influenciadas quando na amostra houver observações discrepantes. Assim, torna-se necessária a obtenção de estimadores robustos para $\widehat{\boldsymbol{\mu}}(\boldsymbol{X})$ e $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})$, de forma que amenizem esse efeito, mas que mantenham as propriedades de equivariância sob transformações afins.

3 METODOLOGIA

Em concordância com os objetivos propostos, a metodologia a ser utilizada nesse trabalho considerou a notação e formulação matricial do modelo de regressão multivariada dado na expressão (2.24), seção (2.4). Dessa forma, para a obtenção dos resultados, sequencialmente os tópicos dessa seção encontram-se estruturados da seguinte forma: 3.1 valores paramétricos utilizados na simulação Monte Carlo e na geração de observações discrepantes; 3.2 construção dos algoritmos *FAST-MCD* e *MVE*; 3.3 avaliação da eficiência relativa entre os estimadores robustos e não robusto da matriz de covariância e 3.4 aplicação das medidas de eficiência na análise sensorial da qualidade de café.

3.1 Valores paramétricos utilizados na simulação Monte Carlo e na geração de observações discrepantes

Com o propósito de gerar variáveis respostas Y com observações discrepantes, desenvolveu-se um programa de simulação utilizando o *software* R(R DEVELOPMENT CORE TEAM, 2010) para a obtenção dos resultados e ajuste do modelo. Nesse processo, foram simuladas amostras aleatórias no espaço p-dimensional de tamanho n dados por $\boldsymbol{Y}_1$,$\boldsymbol{Y}_2$,...,$\boldsymbol{Y}_n$, com $\boldsymbol{Y}_i \in \mathbb{R}^p$, para $i = 1,2,...,n$. Dessa forma, o vetor p-dimensional da i-ésima unidade amostral foi escrito como $\boldsymbol{Y}_i = [Y_{i1},Y_{i2},...,Y_{ip}]^t$.

As simulações Monte Carlo foram realizadas para tamanhos amostrais fixados em n igual a $20, 50, 100,150$ e 200 observações, número de variáveis dependentes fixadas em $p = 5$ e $p = 10$, taxas de misturas fixadas em δ igual a $0{,}05; 0{,}10; 0{,}20;$ $0{,}30$ e $0{,}40$, interpretadas como fração média de observações discrepantes presentes na amostra.

Os valores paramétricos assumidos foram definidos nos vetores de médias $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ p-dimensional, considerando-se, também, as matrizes de covariâncias $\boldsymbol{\Sigma_1}$ e $\boldsymbol{\Sigma_2}$, de ordem p. As amostras multivariadas, representadas pelas matrizes A e B, foram geradas alternando-se os valores paramétricos entre $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ e entre $\boldsymbol{\Sigma_1}$ e $\boldsymbol{\Sigma_2}$ especificados por

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.1}$$

$$\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & \rho^{1-2} & \cdots & \rho^{1-p} \\ \rho^{2-1} & 1 & \cdots & \rho^{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{3.2}$$

em que assumiu-se $\rho = 0{,}5$ como o coeficiente de correlação, diferentes probabilidades de misturas e distribuições *log-normais* e *t-student*.

Mantendo uma probabilidade de mistura fixada para cada unidade amostral, aleatoriamente, gerou-se um valor distribuído por uma $U\,(0, 1)$ e representado por u. Assim sendo, os critérios estabelecidos para compor a amostra de tamanho n, considerando a presença de observações discrepantes, encontram descritos a seguir:

(i) Se $u_j > \delta$, então os dados assumirão valores de uma distribuição normal p-variada com os valores paramétricos $\boldsymbol{\mu_1}$ e $\boldsymbol{\Sigma_1}$ de tal forma que $A_i \sim N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$; se $u_j \leq \delta$, os dados assumirão valores de uma *log-normal* p-variada com a configuração $A_i \sim LN_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$, para $i = 1, ..., n$ e $j =$

$1, ..., n$.

(ii) Se $u_j > \delta$, então os dados assumirão valores de uma distribuição normal p-variada com os valores paramétricos $\boldsymbol{\mu_2}$ em $\boldsymbol{\Sigma_2}$ de tal forma que $B_j \sim N_p(\boldsymbol{\mu_2},\boldsymbol{\Sigma_2})$; se $u_j \leq \delta$, os dados assumirão valores de uma distribuição *t-student* p-variada com a configuração $B_j \sim t_p(\boldsymbol{\Sigma_2},\nu = 5)$, em que ν são os graus de liberdade.

3.2 Construção dos algoritmos *FAST-MCD* e *MVE*

Após gerarem as amostras multivariadas em diferentes níveis de contaminação por observações discrepantes, procedeu-se à estimação robusta do vetor de médias e da matriz de covariância.

Dado o propósito de se obter estimativas robustas por meio dos métodos *MCD* e *MVE* (seção 2.6), assumindo na geração dos dados, diferentes quantidades de observações discrepantes, o algoritmo *FAST-MCD*, proposto por Rousseeuw e Driessen (1999), foi computado assumindo o método de reamostragem de conjuntos elementares descrito na seção (2.5.4), considerando um conjunto de dados observados $\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,...,\boldsymbol{Y}_n\}$ com $\boldsymbol{Y}_i \in \mathbb{R}^p$, para $i = 1,2,...,n$.

Assim sendo, os subconjuntos $(\boldsymbol{J})$ de tamanho $h = \left(\dfrac{n+p+1}{2}\right)$ foram obtidos e o conjunto S, definido na seção (4.8) conteve todos os possíveis subconjuntos de tamanho h a partir da combinação $C_{n,h}$. Para cada $\boldsymbol{J} \subset S$, o algoritmo *FAST-MCD* foi executado nos seguintes passos:

(i) Calcularam-se todos os possíveis subconjuntos $(\boldsymbol{J} \subset S)$ de tamanho h

(ii) Definiu-se, arbitrariamente, um subconjunto $\boldsymbol{H_1}$ de tamanho h.

(iii) Calcularam-se as estimativas para o vetor de médias $\boldsymbol{\mu}_{H_1}$ e matriz de covariâncias $\boldsymbol{S}_{H_1}$.

(iv) Calcularam-se as distâncias de Mahalanobis dadas por

$$d_{H_1}(i) = \sqrt{\left(\boldsymbol{Y}_i - \boldsymbol{\mu_{H_1}}\right)^t \boldsymbol{S}_{\boldsymbol{H_1}}^{-1} \left(\boldsymbol{Y_i} - \boldsymbol{\mu_{H_1}}\right)}$$

.

(v) Ordenaram-se essas distâncias de forma crescente.

(vi) Definiu-se o subconjunto $\boldsymbol{H_2}$, cujos os elementos foram as distâncias ordenadas no passo (v)

(vii) Calcularam-se $\boldsymbol{\mu_{H_2}}$, $\boldsymbol{S_{H_2}}$ e $d_{\boldsymbol{H_2}}(i)$.

(viii) Caso o $\det(\boldsymbol{S_{H_2}}) = 0$, os passos de (ii) a (vi) seriam repetidos porém, se o $\det(\boldsymbol{S_{H_2}}) = \det(\boldsymbol{S_{H_1}})$, o processo seria interrompido. Caso contrário, com o não atendimento dessa condição, o processo seria continuado até a k-ésima iteração em que $\det(\boldsymbol{S_{H_k}}) = \det\left(\boldsymbol{S_{H_{(k+1)}}}\right)$. E, naturalmente, $\det(\boldsymbol{S_{H_1}}) \geq \det(\boldsymbol{S_{H_2}}) \geq \ldots \geq det(\boldsymbol{S_{H_k}}) = \det\left(\boldsymbol{S_{H_{(k+1)}}}\right)$.

Já, a execução do algoritmo *MVE* proposto por Rousseeuw () foi dada nos seguintes passos:

(i) Definiu-se o conjunto S com todos os possíveis subconjuntos de tamanho h.

(ii) Definiu-se um subconjunto $\boldsymbol{J} \subset S$.

(iii) Calculou-se as estimativas para o vetor de médias $\boldsymbol{\mu_J}$ e matriz de covariâncias $\boldsymbol{S_J}$.

(iv) As distâncias dadas por $d_{\boldsymbol{J}}(i) = \sqrt{\text{med}\left(\boldsymbol{Y}_i - \boldsymbol{\mu_J}\right)^t \boldsymbol{S_J}^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu_J}\right)}$ foram calculadas, nas quais a mediana para $i = 1, ..., n$ é representa por med.

(v) Calcularam-se o quadrado do volume do elipsóide resultante, representado por $P^2_{\boldsymbol{J}}$, proporcional a $d^2_{\boldsymbol{J}} \det\left(\boldsymbol{S}_{\boldsymbol{J}}\right)$, ou seja, $P^2_{\boldsymbol{J}} = d^2_{\boldsymbol{J}} \det\left(\boldsymbol{S}_{\boldsymbol{J}}\right)$.

(vi) O passo (v) foi repetido para todos subconjuntos $\boldsymbol{J}$, selecionando-se aquele com menor valor de $P_{\boldsymbol{J}}$.

Obtidas as estimativas robustas das matrizes de covariâncias em (3.2), investigou-se o desempenho dos estimadores robustos de regressão multivariada *MCD* e *MVE* em relação aos estimadores de mínimos quadrados. Computou-se a medida de eficiência relativa ($\mathbf{ER_D}$), tendo por base o emprego dos determinantes das matrizes de covariância utilizados na composição de uma medida de eficiência. O que possibilitará ao pesquisador comparar um estimador robusto em relação ao estimador de mínimos quadrados, mediante a diferentes quantidades de observações discrepantes contidas na amostra. Tal medida é definida em (3.3).

3.3 Avaliação da eficiência relativa entre os estimadores robustos e não robusto da matriz de covariância

Então, a medida definida como

$$\mathbf{ER_D}(\widehat{\boldsymbol{\Sigma}}_{R_i}, \widehat{\boldsymbol{\Sigma}}) = \left(\frac{\left|\widehat{\boldsymbol{\Sigma}}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{R_i}\right|} \right)^{\frac{1}{p}} \qquad i = 1{,}2 \tag{3.3}$$

em que $\widehat{\boldsymbol{\Sigma}}_{R_i}$ correspondeu ao estimador robusto da matriz de covariâncias obtida, sendo que $\widehat{\boldsymbol{\Sigma}}_{R_1}$ representou os estimadores *MCD*, $\widehat{\boldsymbol{\Sigma}}_{R_2}$ representou os estimadores *MVE* e $\widehat{\boldsymbol{\Sigma}}$ correspondeu à estimativa da matriz de covariâncias obtida pelo método dos mínimos quadrados.

Com o propósito de comparar o desempenho dessa medida, substituindo-se o determinante pelo traço das matrizes de covariância, redefiniu-se $\mathbf{ER_D}$ dada

na expressão (3.3), que será descrita na expressão (3.4)

$$\mathbf{ER}_T(\widehat{\mathbf{\Sigma}}_{R_i}, \widehat{\mathbf{\Sigma}}) = \left(\frac{tr\left(\widehat{\mathbf{\Sigma}}\right)}{tr\left(\widehat{\mathbf{\Sigma}}_{R_i}\right)} \right)^{\frac{1}{p}} \qquad i = 1{,}2 \tag{3.4}$$

3.4 Aplicação das medidas de eficiência na análise sensorial da qualidade de café

A aplicação das medidas de eficiência foi realizada na comparação das matrizes de covariâncias estimadas pelos métodos *MCD* e *MVE*, sendo essas utilizadas na construção de um modelo de regressão multivariada, utilizado para descrever o perfil sensorial e químico de genótipos de cafeeiro Bourbon *Coffea arabica L.*, no Sul do estado de Minas Gerais e na região Mogiana do estado de São Paulo, visando à identificação de genótipos promissores para a produção de cafés de qualidade superior.

Com esse propósito, as variáveis físico-químicas, teores finais de trigonelina, ácidos clorogênicos e cafeína, dados em porcentagem de matéria seca (% m.s), foram consideradas independentes e os atributos sensoriais fragrância, sabor, acidez, corpo e equilíbrio como variáveis dependentes (Anexo A). Mantendo-se essas especificações, a descrição do experimento utilizado para a obtenção das respostas é dada a seguir por meio da caracterização dos experimentos (3.4.1).

3.4.1 Caracterização dos experimentos

De acordo com a metodologia utilizada por Figueiredo (2010), os resultados experimentais foram obtidos por meio da avaliação de quatorze genótipos de café arábica *Coffea arabica* L., sendo onze pertencentes ao grupo da cultivar Bourbon, conhecida pelo elevado potencial para a produção de cafés especiais, e três representantes de cultivares comerciais, amplamente cultivadas no Brasil

(Tabela 1). O grupo de genótipos Bourbon constitui uma população segregante, obtida a partir de sementes coletadas em diferentes regiões cafeeiras do Brasil.

Tabela 1 Relação dos genótipos presentes nos experimentos.

	Genótipo/local de origem
1	Bourbon Amarelo/Epamig - Machado
2	Mundo Novo IAC 502/9/Epamig - Machado
3	Catuaí Vermelho IAC 144/Epamig - Machado
4	Icatu Precoce IAC 3282/Procafé - Varginha
5	Bourbon Amarelo/Procafé - Varginha
6	Bourbon Amarelo/Fazenda Bom Jardim - Sto. Antônio do Amparo
7	Bourbon Vermelho/Fazenda São João Batista - Campos Altos
8	Bourbon Amarelo LCJ 9/IAC - Campinas
9	Bourbon Amarelo/Faz. Toriba - São Sebastião do Paraíso
10	Bourbon Amarelo LCJ 10/Fazenda São Paulo - Oliveira
11	Bourbon Amarelo/Aluízio Castro - Carmo de Minas
12	Bourbon Amarelo/Fazenda Paixão - Carmo de Minas
13	Bourbon Trigo/Fazenda Monte Alegre - Alfenas
14	Bourbon Amarelo/Fazenda Samambaia - Sto. Antônio do Amparo

As outras três cultivares Mundo Novo IAC 502/9, Catuaí Vermelho IAC 144 e Icatu Precoce IAC 3282, utilizadas como referência para avaliação da qualidade, são provenientes de instituições de pesquisa e respondem por mais de 90% das lavouras comerciais do Brasil.

As populações foram estabelecidas na forma de experimento em campo, no Sul do estado de Minas Gerais - abrangendo os municípios de Lavras e Santo Antônio do Amparo - e na região Mogiana do estado de São Paulo em São Sebastião da Grama (Tabela 6). Os experimentos foram instalados em dezembro de 2005, no espaçamento de $3{,}5 \times 0{,}8$m. Os dados referiram-se à segunda colheita (safra 2009). Foram adotadas todas as práticas de manejo e recomendação de adubação, usualmente, empregadas na cultura do cafeeiro.

Foram avaliados 14 genótipos de café arábica, em três ambientes de pro-

Tabela 2 Relação dos locais de instalação dos experimentos.

Cidade	Região do Estado	Local
Lavras	Sul de Minas	Setor de cafeicultura DAG/UFLA
Santo Antônio do Amparo	Sul de Minas	Fazenda Cerrado Grande
São Sebastião da Grama	Mogiana Paulista	Fazenda Recreio

dução. Os três experimentos foram instalados no delineamento de blocos casualizados, com três repetições, sendo a parcela experimental constituída por dez plantas. Maiores detalhes podem ser encontrados em Figueiredo (2010).

3.4.2 Ajuste do modelo de regressão multivariada

Em concordância com os objetivos da aplicação (seção 3.4) e com a descrição dos experimentos (seção 3.4.1), o modelo de regressão multivariada proposto foi definido por

$$\boldsymbol{Y}_{(42\times5)} = \boldsymbol{X}_{(42\times5)}\boldsymbol{\beta}_{(4\times5)} + \boldsymbol{\varepsilon}_{(42\times5)} \tag{3.5}$$

em que, para os 14 genótipos de café referentes às diferentes altitudes, $\boldsymbol{Y}$ correspondeu a uma matriz (42×5), contendo os valores médios dos atributos sensoriais fragrância (Y_1), sabor (Y_2), acidez (Y_3), corpo (Y_4) e equilíbrio (Y_5); $\boldsymbol{X}$ correspondeu a uma matriz (42×4) contendo os valores médios dos teores de trigonelina (X_1), ácidos clorogênicos (X_2) e cafeína (X_3) ($\%$ m.s); $\boldsymbol{\beta}$ correspondeu a uma matriz (4×5) de coeficientes de regressão que foram parâmetros desconhecidos a serem estimados e $\boldsymbol{\varepsilon}$ correspondeu a uma matriz (42×5), contendo os componentes do erro aleatório associado aos valores médios dos atributos sensoriais.

Após o ajuste do modelo, ambas as medidas $\mathbf{ER}_D$ e $\mathbf{ER}_T$ foram interpretadas mantendo-se o valor unitário como referência, de modo que quando os resultados dessas medidas foram superior a 1, considerou-se, então, o estimador

robusto especificado em cada medida, mais eficiente que o estimador de mínimos quadrados.

Seguindo as recomendações de Peña e Prieto (2001), Atkinson (1994) e Rousseeuw e Leroy(1987), a identificação de observações discrepantes foi feita por meio da construção de gráficos constituídos dos resultados das distâncias robustas, obtidas com a execução dos algoritmos *FAST-MCD* e *MVE* (3.2).

De forma a identificar as observações classificadas como pontos de alavanca pelos procedimentos clássico e robusto, plotou-se o gráfico *DD plot* considerando as distâncias clássicas *versus* as distâncias robustas. Nessa situação, se os dados não forem contaminados, todos os pontos se localizarão no retângulo onde ambas distância são regulares, enquanto os pontos discrepantes se localizarão mais acima ou à direita da linha de corte. Quatro regiões serão delimitadas pelo valor crítico descrito na equação (2.45) e usadas para identificar os pontos de alavanca (ROUSSEEUW; DRIESSEN, 1999).

4 RESULTADOS E DISCUSSÃO

Nas figuras a seguir são apresentados os resultados gráficos da eficiência relativa ($\boldsymbol{ER}$) dos estimadores usual e robustos das matrizes de covariâncias as quais são resultantes da aplicação dos métodos *MCD* e *MVE* descritos na seção 3.2. As observações discrepantes simuladas foram provenientes das distribuições multivariadas *t-student* (Figuras A e C) e *log-normal* (Figuras B e D), cujas características essencialmente correspondem, em comparação com a distribuição normal multivariada, a desvios de curtose e desvios de simetria com excesso de curtose. A interpretação desses gráficos foi dada de forma comparativa, mantendo-se as especificações paramétricas mencionadas na seção 3.1.

Por meio dos resultados descritos nas Figuras 1A e 1B, em se tratando da estimativa *MCD* da matriz de covariância, na situação em que foi simulada uma porcentagem mínima de observações discrepantes ($\delta = 0{,}05$) e dado o efeito dessas observações provenientes de distribuições com desvios de curtose, observa-se que para $n > 50$ a eficiência relativa ($\boldsymbol{ER_D}$) não foi afetada. Isso pode ser explicado pelo fato de que quanto maior a dimensão da amostra, mais a distribuição *t-student* se aproxima da distribuição Normal. Em termos práticos, pode-se afirmar que há evidências de que a estimativa da matriz de covariância usual poderá ser utilizada nessa condição, sugerindo ao pesquisador não haver necessidade de se utilizar estimativas robustas.

Segundo Todorov e Filzmoser (2009), inicialmente, o *MCD* foi negligenciado em favor do *MVE* porque o algoritmo de reamostragem simples foi mais eficiente para o *MVE*. O ponto de ruptura desse estimador coincide com o do *MVE* (0,5) e apresenta baixa convergência, tornando-o o mais utilizado, entretanto esse estimador não é muito eficiente em se tratando de modelos normais

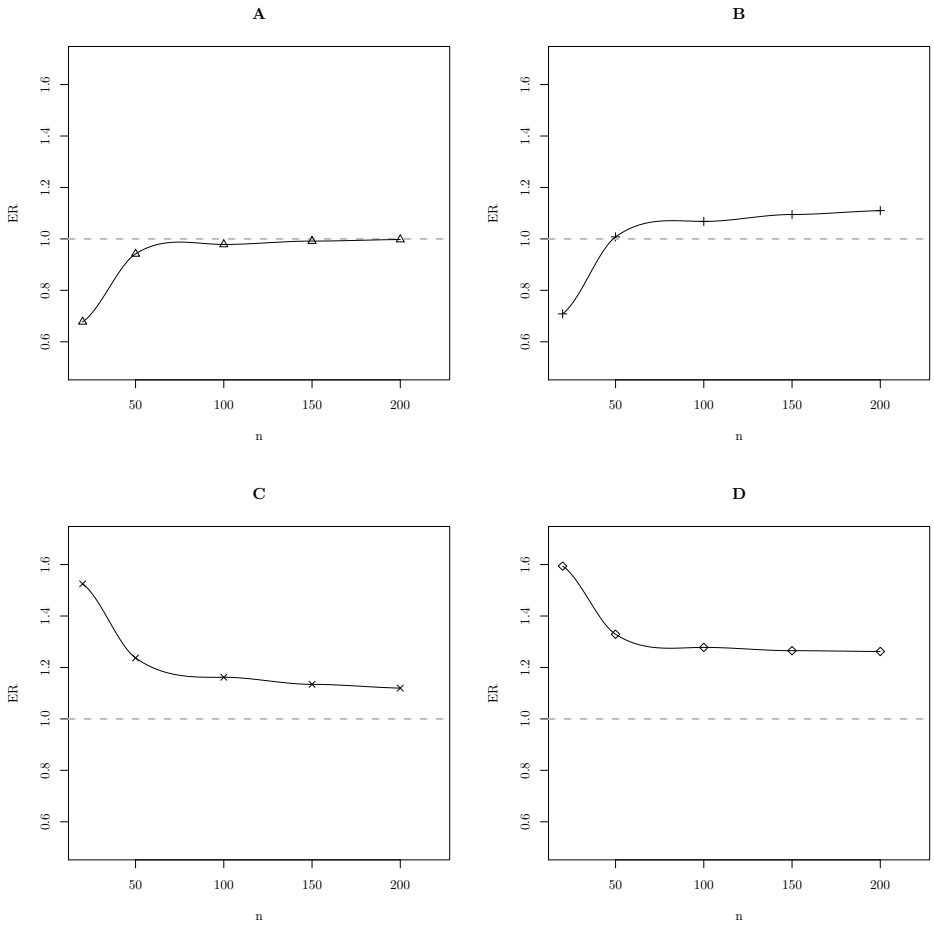

Figura 1 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}05$ das distribuições *t-student* e *log-normal*, utilizando os algoritmos *MCD* (A e B) e *MVE* (C e D)

ou aproximadamente normais, especialmente, se h for selecionado de forma que o maior ponto de ruptura seja alcançado.

A partir dos resultados ilustrados nas Figuras 1C e 1D, pode-se observar que a $\boldsymbol{ER_D}$, em que a estimativa robusta da matriz de covariância foi obtida pelo método *MVE*, apresentou maior sensibilidade em relação aos desvios de curtose,

de tal forma que, à medida que o tamanho amostral foi aumentando, notou-se uma tendência na redução da eficiência relativa do estimador da matriz de covariância *MVE*, na situação em que as observações discrepantes foram geradas por meio da distribuição *t-student* (Figura 1C). Esse resultado foi perceptível pela aproximação do resultado em relação ao valor da unidade de referência, representada pela linha tracejada.

Especificamente, na situação em que as observações discrepantes apresentaram desvios de simetria (Figura 1D), para $n > 50$ a eficiência do estimador *MVE*, aparentemente, não registrou uma tendência à redução, mantendo-se constante à medida que o tamanho amostral foi aumentando, além de apresentar valores superiores em relação aos resultados ilustrados na Figura 1C. Assim, no caso em que os dados apresentaram desvio de simetria com excesso de curtose, pode-se afirmar que o estimador *MVE* foi mais eficiente.

De acordo com Rousseeuw e Zomeren (1990), os estimadores *MVE* são afim equivariantes, possuem ponto de ruptura $0{,}5$, mas são computacionalmente ineficientes, convergindo para a distribuição normal a uma taxa menor que a usual. Entretanto, o algoritmo apresentado para esse método, utilizando processos de reamostragem por meio de conjuntos elementares, é satisfatório em relação ao custo computacional e, com o avanço desses recursos, começa a ser mais utilizado e explorado na análise de regressão.

Na perspectiva de que a amostra apresenta uma quantidade expressiva de observações discrepantes, simulada em $\delta = 0{,}40$ e $p = 5$, os resultados relativos à eficiência dos estimadores robustos, mediante aos efeitos dos desvios de curtose e desvios de simetria, foram mais diferenciados em relação ao método de estimação. Nesse aspecto, comparando-se as Figuras 2A e 2B, notou-se que a estimativa da matriz de covariância do método *MCD* foi mais eficiente quando essas observações

apresentaram maiores desvios de simetria (Figura 2B). Analogamente, observou-se o mesmo efeito para o método *MVE* (Figuras 2C e 2D). Nessa situação, os dados simulados apresentaram uma porcentagem de observações discrepantes próximas do maior ponto de ruptura possível ($\delta^* = 50\%$).

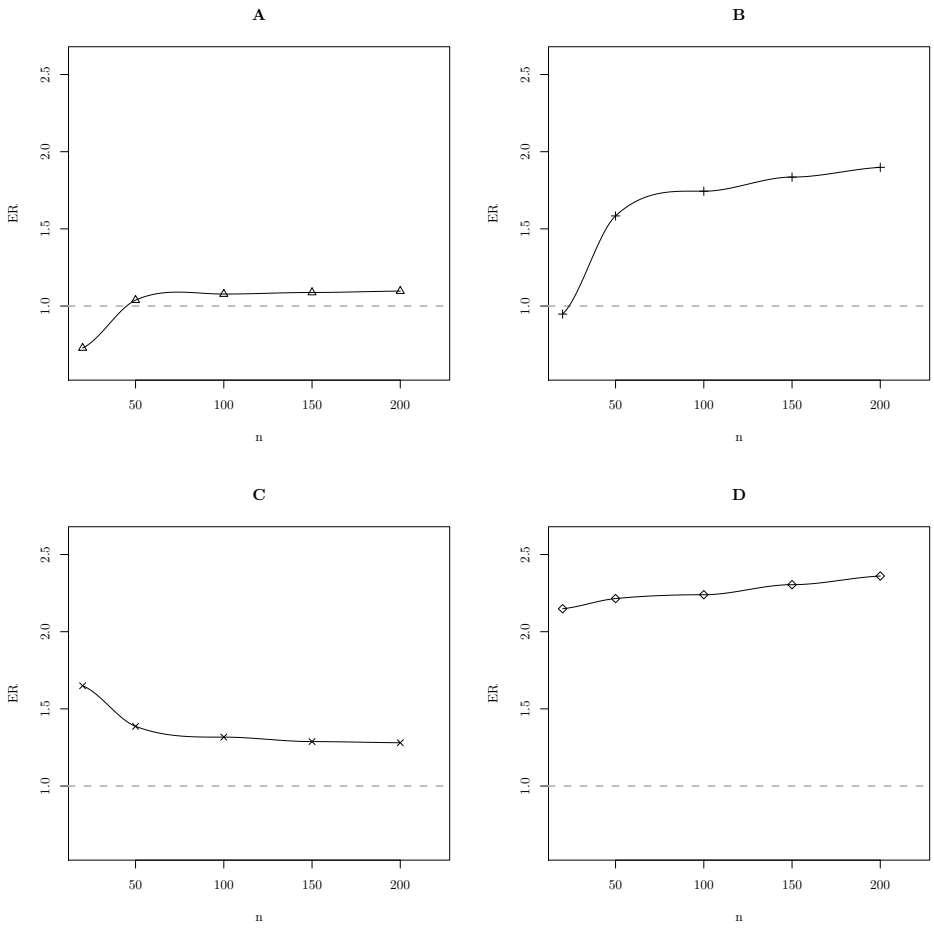

Figura 2 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}40$ das distribuições *t-student* e *log-normal*, utilizando os algoritmos *MCD* (A e B) e *MVE* (C e D)

Na Figura 3, quando se considera o aumento no número de variáveis ($p = 10$), pode-se observar que as características citadas com relação à Figura 1 e à Figura 2 são mantidas, mesmo considerando um aumento na porcentagem de observações discrepantes. O estimador *MVE* foi mais eficiente quando se

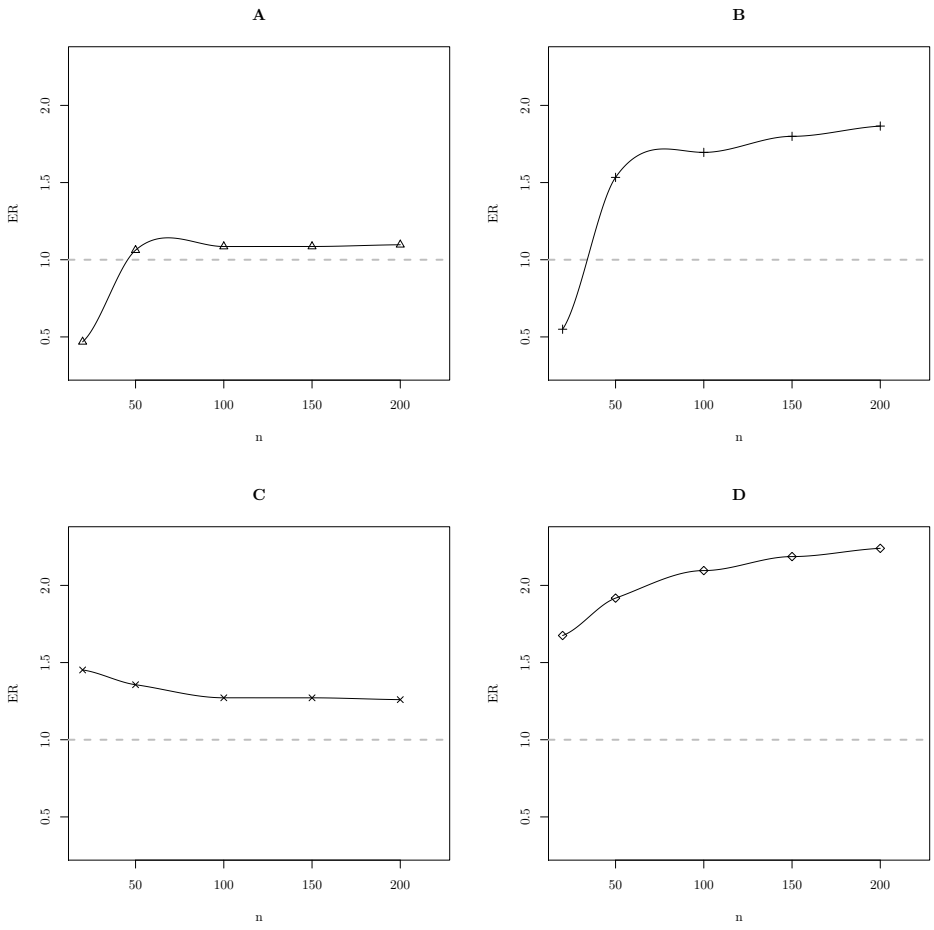

Figura 3 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ em função de diferentes tamanhos amostrais n, considerando $p = 10$ variáveis, taxa de mistura fixada em $\delta = 0{,}40$ das distribuições *t-student* e *log-normal*, utilizando os algoritmos *MCD* (A e B) e *MVE* (C e D)

considera o efeito das observações provenientes de distribuições com desvio de simetria e excesso de curtose. Os resultados, apresentados no Anexo B, permitiram verificar que o efeito de simetria tende a melhorar a eficiência do método *MVE* e que a eficiência relativa não foi afetada com o aumento do número de variáveis.

Nas Figuras de 4 a 7 são apresentados os resultados gráficos para a comparação do cálculo da eficiência relativa dos estimadores usual e robustos das matrizes de covariâncias, resultantes da aplicação dos métodos *MCD* e *MVE*, considerando o uso do determinante ($\boldsymbol{ER_D}$ - Figuras A e B). A interpretação desses gráficos foi dada de forma comparativa à proposta de cálculo da eficiência relativa por meio do uso do traço ($\boldsymbol{ER_T}$ - Figuras C e D), apresentada na seção 3.3. As observações discrepantes simuladas foram provenientes das distribuições multivariadas *t-student* (Figuras A e C) e *log-normal* (Figuras B e D).

Nesse contexto, na Figura 4, ao se comparar os resultados referentes às eficiências $\boldsymbol{ER_D}$ (Figuras 4A e 4B) e $\boldsymbol{ER_T}$ (Figuras 4C e 4D), via método *MCD*, pode-se observar que os resultados foram similares na situação em que foi simulada uma porcentagem mínima de observações discrepantes ($\delta = 0{,}05$). Considerando também o fato de que as observações discrepantes foram provenientes de distribuições com desvios de curtose e desvios de simetria para todos os tamanhos amostrais (n).

De forma semelhante, pode-se verificar o mesmo comportamento quando se considera o cálculo da eficiência relativa, via método *MVE* (Figura 5). Nesse caso, quando se compara o uso do traço (Figuras 5C e 5D) com o uso do determinante (Figuras 5A e 5B), observa-se que a $\boldsymbol{ER_T}$ apresentou valores mais próximos do valor de referência, porém a eficiência do estimador não evidenciou uma tendência à redução, mantendo-se constante à medida que o tamanho amostral foi aumentando. Além disso, no caso em que os dados apresentaram desvios de

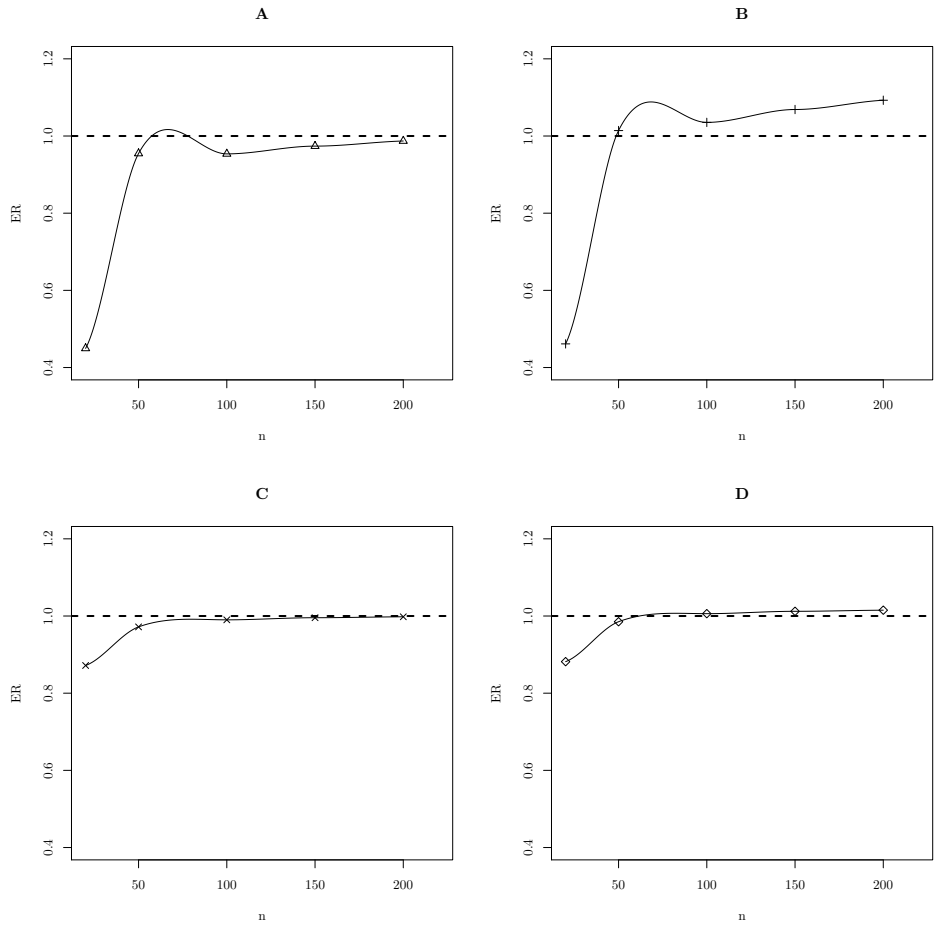

Figura 4 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ (A e B), em comparação com a eficiência relativa $\boldsymbol{ER_T}$ (C e D) em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}05$ das distribuições *t-student* e *log-normal*, utilizando o algoritmo *MCD*

curtose e desvios de simetria, a $\boldsymbol{ER_T}$ apresentou valores superiores em relação aos resultados ilustrados nas Figuras 4C e 4D. Nessa condição, os resultados confirmaram que o estimador *MVE* foi mais eficiente.

Quando a amostra apresentou uma quantidade expressiva de observações

discrepantes, simulada em $\delta = 0{,}40$ e $p = 5$, ao se comparar os resultados referentes às eficiências $\boldsymbol{ER_D}$ (Figuras 6A e 6B) e $\boldsymbol{ER_T}$ (Figuras 6C e 6D), via método *MCD*, os resultados da $\boldsymbol{ER_T}$ confirmaram que, mediante aos efeitos dos desvios de curtose e desvios de simetria, foram mais diferenciados. Nesse as-

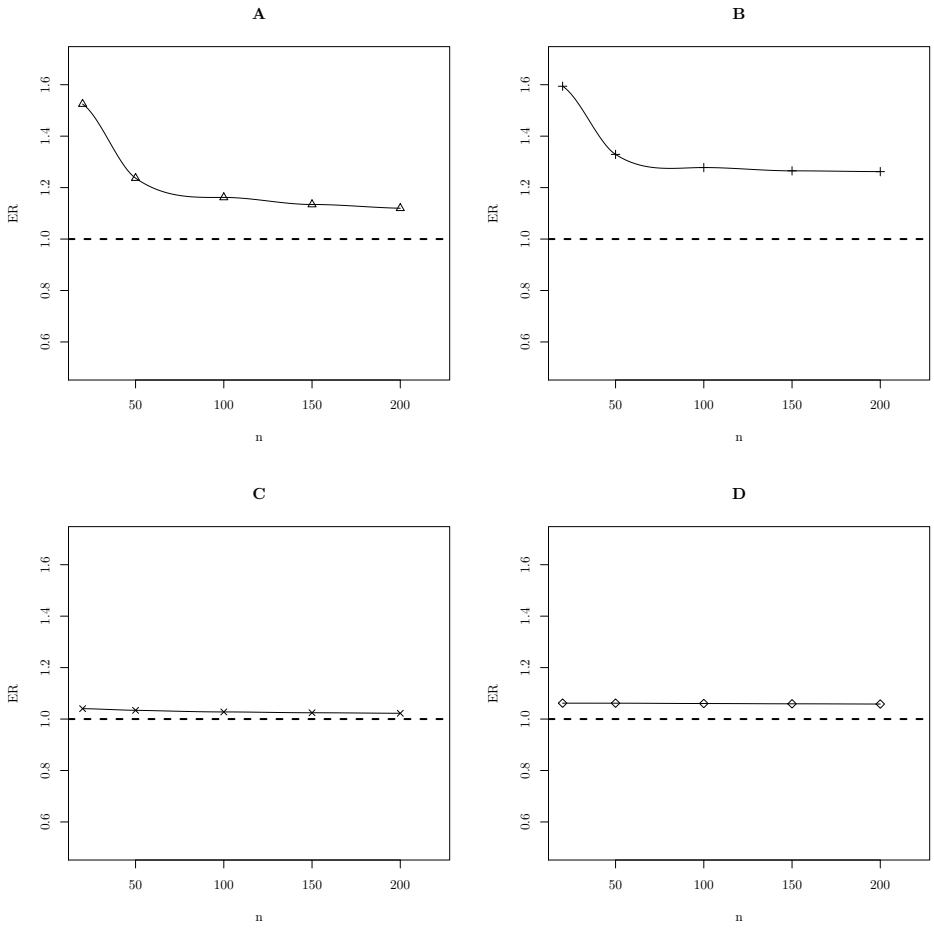

Figura 5 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ (A e B), em comparação com a eficiência relativa $\boldsymbol{ER_T}$ (C e D) em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}05$ das distribuições *t-student* e *log-normal*, utilizando o algoritmo *MVE*

pecto, comparando-se as Figuras 6C e 6D, verificou-se que a estimativa da matriz de covariância obtida foi mais eficiente, quando essas observações apresentaram maiores desvios de simetria (Figura 6D).

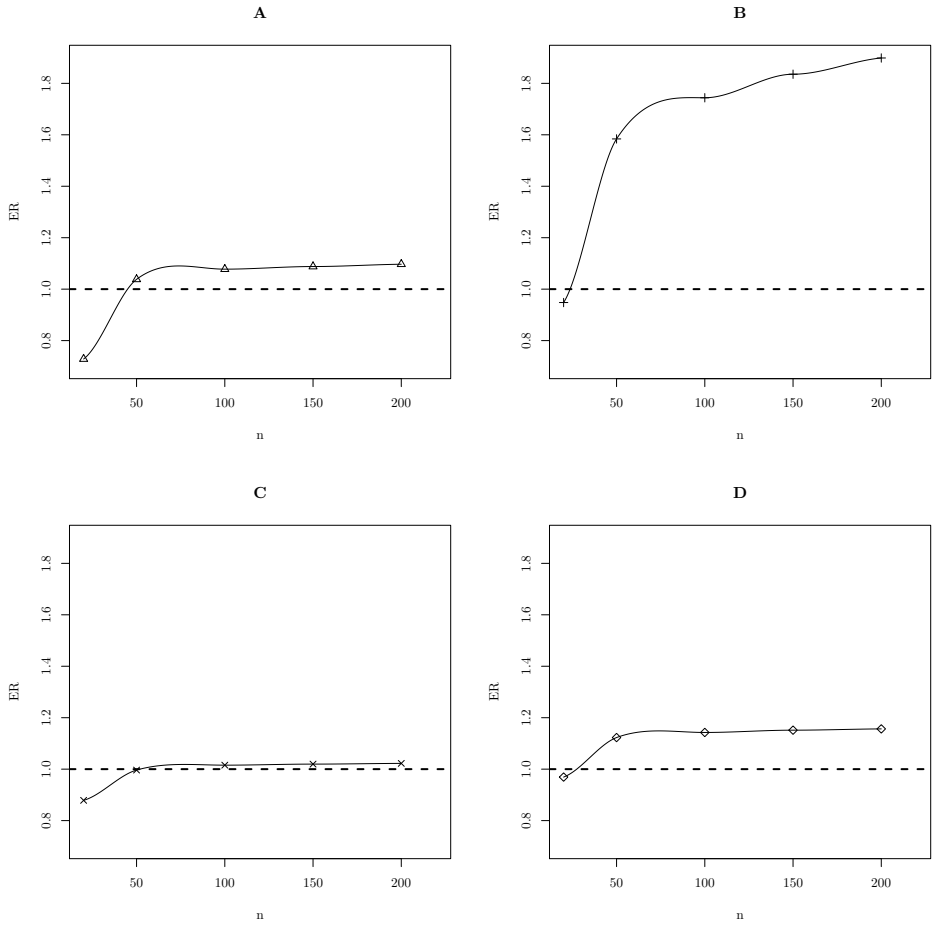

Figura 6 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ (A e B), em comparação com a eficiência relativa $\boldsymbol{ER_T}$ (C e D) em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}40$ das distribuições *t-student* e *log-normal*, utilizando o algoritmo *MCD*

Pode-se ainda notar que a $\boldsymbol{ER_T}$ (Figuras 6C e 6D) apresentou valores

mais próximos do valor de referência, quando se compara o uso do traço com o uso do determinante no cálculo da $\boldsymbol{ER}$, porém a eficiência do estimador manteve-se constante à medida que o tamanho amostral foi aumentando. Similarmente, na situação em que a amostra apresentou uma quantidade expressiva de observações

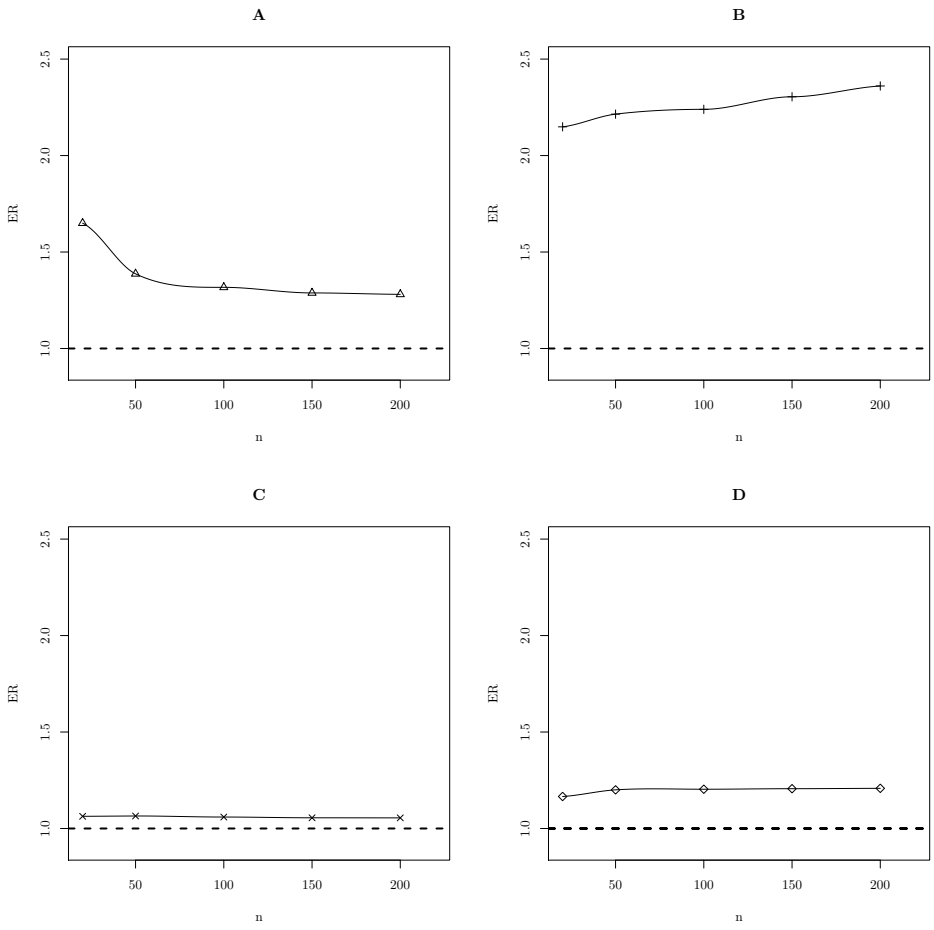

Figura 7 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ (A e B), em comparação com a eficiência relativa $\boldsymbol{ER_T}$ (C e D) em função de diferentes tamanhos amostrais n, considerando $p = 5$ variáveis, taxa de mistura fixada em $\delta = 0{,}40$ das distribuições *t-student* e *log-normal*, utilizando o algoritmo *MVE*

discrepantes, simulada em $\delta = 0{,}40$ e $p = 5$, ao se comparar, via método *MVE*, os resultados referentes às eficiências $\boldsymbol{ER_D}$ (Figuras 7A e 7B) e $\boldsymbol{ER_T}$ (Figuras 7C e 7D) notou-se o mesmo efeito. Os resultados apresentados para a $\boldsymbol{ER_T}$ nas Figuras 6 e 7 confirmaram que a estimativas da matrizes de covariâncias dos métodos *MCD* e *MVE* foram mais eficientes, quando essas observações apresentaram maiores desvios de simetria.

4.1 Exemplo com dados reais relacionados à aplicação das medidas de eficiência na análise sensorial da qualidade do café

Os resultados apresentados na seção anterior avaliaram, empiricamente, as medidas de eficiência, em se tratando da aplicabilidade na análise sensorial da qualidade do café e tendo como suporte o conjunto de dados, cujo tamanho amostral é $n = 42$, apresentado na seção 3.4 e no Anexo A.

Preliminarmente à identificação das observações discrepantes, tendo por base o modelo de regressão multivariada especificado na seção 2.4.1, obteve-se as estimativas dos parâmetros de regressão e os valores médios preditos considerando os métodos de mínimos quadrados (MMQ), *MCD* e *MVE*. Tais resultados encontram-se apresentados nas Tabelas 3 e 4

Observa-se na Tabela 3 que para as variáveis teores finais de trigonelina (X_1) e cafeína (X_3), as estimativas de mínimos quadrados são bem diferentes daquelas obtidas pelos métodos robustos, em especial para o atributo sensorial sabor (Y_2). Os resultados apresentados em Figueiredo (2010), indicaram que os níveis de trigonelina são maiores para os cafés com melhor qualidade de bebida, entretanto não encontrou nenhuma relação entre o teor de cafeína e a qualidade do café.

Tabela 3 Resumo das estimativas para os parâmetros dos modelos de regressão ajustados para a análise sensorial da qualidade do café.

Método	Coeficientes	Estimativas dos coeficientes de regressão**				
MMQ	β_0	7,7861	7,5941	7,3816	7,6731	6,9385
	β_1	-0,2282	-0,4047	0,0012	-0,3631	0,0224
	β_2	0,0224	-0,0638	-0,0120	0,0306	-0,1206
	β_3	-0,2553	0,3332	-0,0413	-0,2324	0,6530
MCD	β_0	7,4880	6,2767	7,1931	7,0121	6,7473
	β_1	-0,0865	-0,6924	0,0122	-0,6548	0,1331
	β_2	-0,0146	-0,0231	-0,0410	0,0579	-0,1031
	β_3	0,0525	1,4815	0,2488	0,4443	0,6301
MVE	β_0	7,1247	6,5749	7,3610	7,2030	6,8329
	β_1	0,3549	-0,3546	-0,0127	-0,6242	0,4823
	β_2	0,0507	-0,0070	0,0018	0,0560	-0,1052
	β_3	-0,3620	0,8466	-0,0830	0,2602	0,2622

** cada coluna corresponde, respectivamente, às estimativas dos coeficientes relacionados com as variáveis (Y_1), (Y_2), (Y_3), (Y_4) e (Y_5).

Os resultados descritos na Tabela 4, evidenciaram que os métodos robustos de estimação *MCD* e *MVE* apresentaram o mesmo desempenho em relação ao de mínimos quadrados. Convém ressaltar que tais métodos, em se tratando da eficiência relativa em relação ao MMQ (Tabela 5), apresentaram valores próximos. Dessa forma, é razoável supor que os modelos de regressão ajustados tenham apresentado as mesmas propriedades relacionadas à predição e qualidade de ajuste.

Tabela 4 Valores médios preditos dos modelos de regressão ajustados para a análise sensorial da qualidade do café.

Variáveis	Valores médios preditos		
	MMQ	*MCD*	*MVE*
Y_1	7,4118	7,3987	7,3758
Y_2	7,2274	7,2215	7,2024
Y_3	7,2785	7,2710	7,2865
Y_4	7,2251	7,2360	7,2285
Y_5	7,0337	7,0305	7,0181

Pode-se notar, pelos resultados descritos na Tabela 5, que a utilização do determinante no cálculo da eficiência relativa, tanto no método *MCD* quanto no método *MVE*, apresentam ligeiras diferenças com relação aos valores de eficiência relativa. Entretanto, não há evidências de que o método *MVE* tenha se mostrado mais eficiente.

Tabela 5 Eficiência relativa dos métodos *MCD* e *MVE* com relação ao método de mínimos quadrados, dados reais.

técnica	método	eficiência relativa
Determinante	*MCD*	1,2189
	MVE	1,3829
Traço	*MCD*	1,0254
	MVE	1,0234

Rousseeuw e Zomeren () perceberam, usando dados reais, que a porcentagem de cobertura do elipsoide de volume mínimo é menor do que o esperado, segundo a distribuição χ^2_k. Os autores realizaram ainda, estudos de simulação para diversos valores de k e diversos tamanhos de amostra, chegando à conclusão que a proporção de cobertura melhora, quando o tamanho de amostra cresce. Logo, a aplicação do método *MVE* poderá produzir melhores resultados, no caso em que se considerar conjuntos com tamanho amostral maior do que o analisado nesse trabalho.

Em se tratando da identificação de dados discrepantes, os autores Rousseeuw e Zomeren (1990) afirmam que a utilização de métodos robustos proporciona uma melhor discriminação de observações discrepantes. Procedeu-se, então, com a análise gráfica para identificação de tais observações utilizando os resultados da distância generalizada de Mahalanobis, para a construção dos gráficos obtidos, com os valores de corte dados por $\sqrt{\chi^2_{3;0,975}}$ (seção 2.7). Os valores da distância foram comparados com o ponto de corte $3{,}0574$. Assim, uma observação, cuja

distância d_i fosse maior que o ponto crítico, seria considerada um possível ponto de alavanca.

Pela investigação dos valores das distâncias clássicas, poderíamos tentar distinguir os pontos de alavanca, mas essas distâncias são afetadas pelas observações discrepantes existentes na amostra. De acordo com Rousseeuw e Zomeren (1990), isso se deve ao fato de que são calculadas com base na matriz de covariâncias amostrais $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})$ e no vetor de médias amostrais $\widehat{\boldsymbol{\mu}}(\boldsymbol{X})$, que possuem ponto de ruptura zero.

Pela análise dos dados da Figura 8, o ajuste pelo método de mínimos quadrados, para se obter as estimativas de $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})$ e $\widehat{\boldsymbol{\mu}}(\boldsymbol{X})$, não revela observações como sendo discrepantes, identificando a observação 2 como suspeita, porém, sem exceder o ponto de corte fixado.

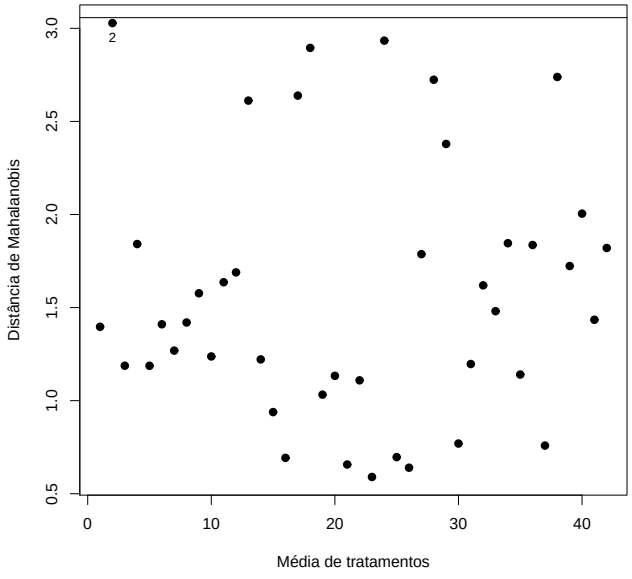

Figura 8 Representação gráfica da distância generalizada de Mahalanobis, pelo método de mínimos quadrados

O ajuste robusto pelo método *MCD* revela claramente a discrepância da observação 2. Esse fato pode ser visualizado na Figura 9. Além disso, utilizando-se o método, identificaram-se as observações 17, 18, 24, 28 e 38 como sendo discrepantes. Ressalta-se que as observações 18, 24 e 28 pertencem ao ambiente São Sebastião da Grama, indicado pelos resultados de Figueiredo (2010) como o que apresentou maior teor de trigonelina e de cafeína, além de apresentar maior aptidão para a produção de cafés especiais.

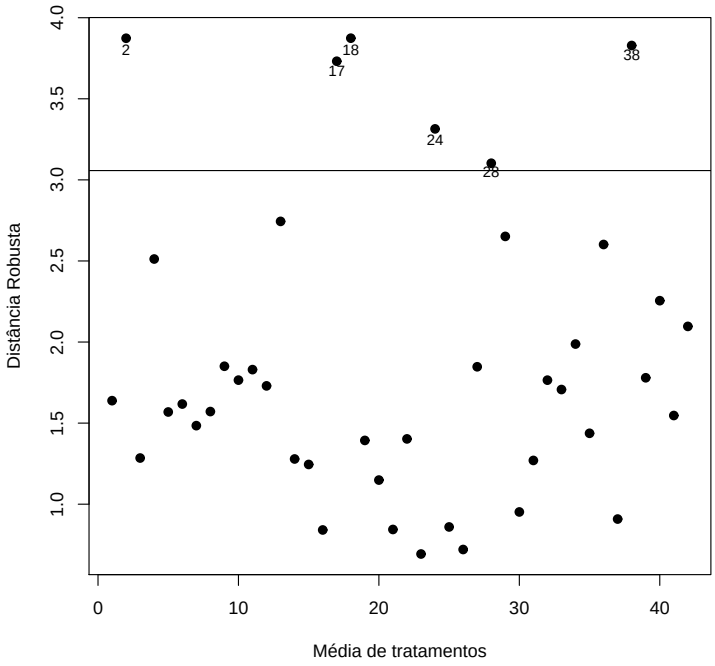

Figura 9 Representação gráfica da distância robusta, pelo método *MCD*

O ajuste, obtido pelo método *MVE* (Figura 10), mostrou-se consistente em relação aos resultados e notou-se que, além daquelas observações reveladas pelo método *MCD*, as observações 4, 12, 36 e 40 foram identificadas como discrepantes

e que a observação 28 se localizou próxima à linha de corte.

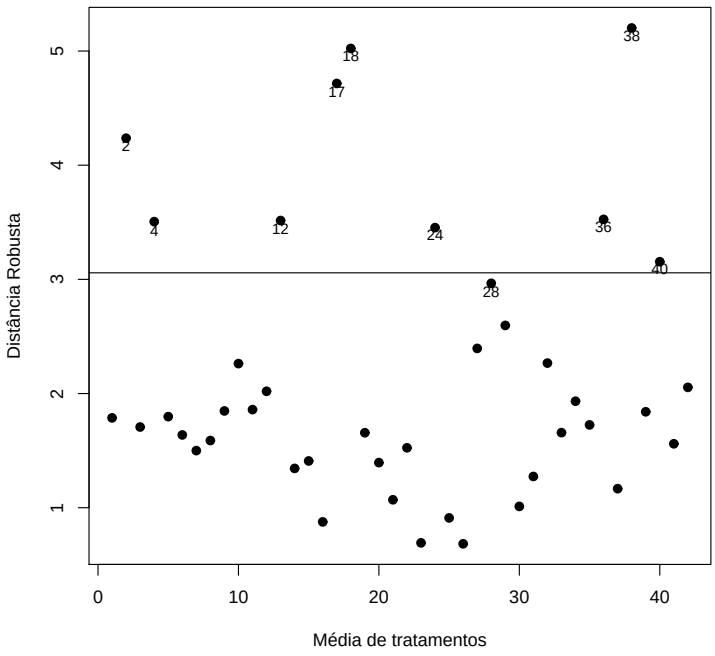

Figura 10 Representação gráfica da distância robusta, pelo método *MVE*

Os resultados gráficos (*D-D plot*) apresentados nas Figuras 11 e 12, respectivamente, por meio dos métodos *MCD* e *MVE*, serviram de base para confirmar a informação sobre os pontos de alavanca. Verificou-se que ao se utilizar o método *MCD*, as observações 2, 17, 18, 24, 28 e 38 foram identificadas como ponto de alavanca (região superior). Enquanto que as observações 2, 4, 12, 17, 18, 24, 28, 36 e 40 foram identificadas pelo método *MVE*.

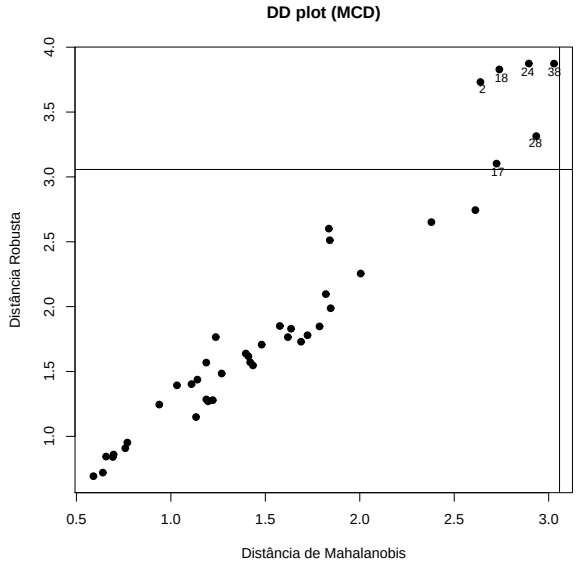

Figura 11 Representação gráfica das distâncias robustas versus distâncias clássicas, pelo método *MCD*

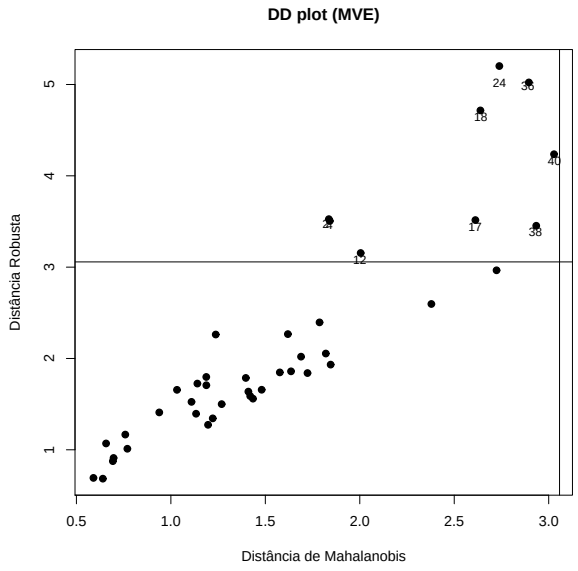

Figura 12 Representação gráfica das distâncias robustas *versus* distâncias clássicas, pelo método *MVE*

5 CONCLUSÕES

Na situação em que foi simulada uma porcentagem mínima de observações discrepantes ($\gamma = 0{,}05$) e considerando-se o efeito dessas observações provenientes de distribuições com desvios de curtose, observou-se que a eficiência relativa não foi afetada para $n > 50$. Evidenciando que a estimativa da matriz de covariância usual poderá ser utilizada, sugerindo ao pesquisador não haver necessidade da utilização de métodos robustos.

O estimador *MVE* foi mais eficiente, quando se considerou o efeito das observações provenientes de distribuições com desvio de simetria e excesso de curtose. A eficiência relativa não foi afetada com o aumento no número de variáveis.

Os resultados proporcionados, pelo uso do traço no cálculo da eficiência relativa, foram promissores, considerando-se o efeito das observações provenientes de distribuições com desvios de curtose e desvios de simetria para todos os tamanhos amostrais (n).

Os modelos de regressão, ajustados pelos métodos *MCD* e *MVE*, foram mais adequados para o estudo das variáveis físico-químicas e sensoriais por terem identificado como discrepantes, observações pertencentes ao ambiente São Sebastião da Grama, sendo condizente com os resultados experimentais obtidos em regiões de maiores altitudes.

REFERÊNCIAS

ATKINSON, A. C. Fast very robust methods for the detection of multiple outliers . **Journal of the American Statistical Association,** New York, v. 89, n. 428, p. 1329–1339, Dec. 1994.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. 3th. ed. Chichester: J. Wiley, 1994. 604p.

COCHRAN, W. G.; COX, G. M. **Experimental designs**. 2th. ed. Canada: J. Wiley, 1992. 611 p.

CROUX, C.; HAESBROECK, G. Influence functions and efficiency of the minimum covariance determinant scatter matrix estimator. **Journal of Multivariate Analysis,** New York, v. 71, n. 2, p. 161–190, Nov. 1999.

CUNHA, U. S. da; MACHADO, S. do A.; FIGUEIREDO FILHO, A. Uso de análise exploratória de dados e de regressão robusta na avaliação do crescimento de espécies comerciais na terra firme da amazônia. **Revista Árvore,** Viçosa, MG, v. 26, n. 4, p. 391–402, Jul/Ago. 2002.

DAMIÃO, J. E. F. **Comparação de carteiras otimizadas segundo o critério média-variância através de estimativas robustas de risco e retorno.** 2007. 36 p. Dissertação (Mestrado em Economia) — Faculdade Ibmec São Paulo, São Paulo, SP, 2007.

DJAUHARI, M. A. Outlier detection: some challenging problem for future research. In: INTERNATIONAL CONFERENCE OF RESEARCH AND EDUCATION IN MATHEMATICS, 2, 2005, Kuala Lumpor. **Proceeding...** Malaysia: Universiti Putra Malaysia, 2005. 1 CD–ROM.

DONOHO, D. L.; HUBER, P. J. The notion of breakdown point. In: BICKEL, P. J. et al. **A Festschrift for Erich L. Lehmann in honor of his sixty-fifth birthday**. Belmont: Wadsworth, 1983. p. 1157–1184.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 3th. ed. New York: John Wiley, 1998. 706 p.

FIGUEIREDO, L. P. **Perfil sensorial e químico de genótipos de cafeeiro Bourbon de diferentes origens geográficas.** 2010. 81 p. Dissertação (Mestrado em ciências dos Alimentos) — UFLA, Lavras, MG, 2010.

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

HAMPEL, F. R. A general qualitative definition of robustness. **Annals of Mathematical Statistics,** Ann Arbor, v. 42, n. 6, p. 1887–1896, Nov. 1971.

HAMPEL, F. R.; RONCHETTI, E. M.; ROUSSEEUW, P. J.; STAHEL, W. A. **The Approach Based on Influence Functions**. New York: John Wiley, 1986.

HERWINDIATI, D. E.; DJAUHARI, M. A.; MASHURI, M. Robust multivariate outlier labeling. **Communications in Statistics-Part B:** Simulation and Computation, New York, v. 36, n. 4, p. 1287–1294, April. 2007.

HERWINDIATI, D. E.; ISA, S. M. The robust distance for similarity measure of content based image retrieval In: WORLD CONGRESS ON ENGINEERING, 2., 2009, London. **Proceedings of the...** London: WCE, U. K., 2009.

HUBERT, M.; ROUSSEEUW, P. J.; AELST, S. V. High-breakdown robust multivariate methods. **Statistical Science,** Ann Arbor, v. 23, n. 1, p. 92–119, Aug. 2008.

JOHNSON, M. E. **Multivariate Statistical Simulation**. New York: J. Wiley, 1987.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th. ed. New Jersey: Prentice Hall, 2007. 793 p.

KOENKER, R. **Quantile Regression**. 1th. ed. New York: Cambridge University Press, 2005. 349 p.

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

MACHADO, H. C. **Detecção de Dados Atípicos e Métodos de Regressão com Alto Ponto de Ruptura.**1997. 153 p. Dissertação (Mestrado em Estatística) — Instituto de Matemática, Estatítica e Computação Científica da Universidade Estadual de Campinas, Campinas, SP, 1997.

MARONNA, R. A. Robust m-estimators of multivariate location and scatter. **The Annals of Statistics,** Hayward, v. 22, n. 4, p. 51–56, Sep. 1976.

MENDES, B. V. M.; LEAL, R. P. C. Robust multivariate modeling in finance. **International Journal of Managerial Finance**, v. 1, n. 2, p. 95–107, June 2005.

NOGUEIRA, F. E. **Modelos de regressão multivariada.** 2007. 96 p. Dissertação (Mestrado em Ciências) — Instituto de Matemática e Estatítica da Universidade de São Paulo., São Paulo, SP, 2007.

OLIVE, D. J. **Applied Robust Statistics**. Carbondale: Southern Illinois University: [s.n.], 2008. 571 p. Disponível em: <http://www.math.siu.edu/olive/run.pdf>. Acesso em: junho. 2011.

PAULA, E. R. **Análise Condicionada da demanda de energia elétrica: Aplicação a um caso real.** 2006. 86 p. Dissertação (Mestrado em Engenharia Elétrica) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.

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

PISON, G.; AELST, V.; WILLEMS, G. Small sample corrections for *LTS* and *MCD*, New York. **Metrika**, v. 55, n. 1/2, p. 111–123, Apr. 2002.

R DEVELOPMENT CORE TEAM. **R:** A Language and Environment for Statistical Computing. Vienna, 2010. Disponível em: <http://www.R-project.org>. Acesso em: 18 de junho. 2011.

ROCKE, D. M.; WOODRUFF, D. L. Identification of outliers in multivariate data. **Journal of the American Statistical Association,** New York, v. 91, p. 1047–1061, Sep. 1996.

ROUSSEEUW, P. J. Multivariate Estimation with High Breakdown Point. In:GROSSMAN, W. et al. **Matematical Statistics and Applications**, Dordrecht: Reidel, 1985. p. 283–297.

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

ROUSSEEUW, P. J.; AELST, S. V.; DRIESSEN, K. V.; AGULLÓ, J. Robust multivariate regression. **Technometrics,** Alexandria, v. 46, n. 3, p. 293–305, Aug. 2004.

ROUSSEEUW, P. J.; DRIESSEN, K. V. Fast algorithm for the minimum covariance determinant estimator. **Technometrics,** Alexandria, v. 41, n. 3, p. 212–223, Aug. 1999.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust Regression and Outlier Detection**. New York: John Wiley, 1987. 343 p.

ROUSSEEUW, P. J.; ZOMEREN, B. C. Robust Distances: Simulations and Cutoff Values. In: STAHEL, W.; WEISBERG, S. (Ed.) **Directions In Robust Statistics and Diagnostics:** part II. New York: Springer-Verlag, 1991. p. 195–203.

ROUSSEEUW, P. J.; ZOMEREN, B. C. Unmasking multivariate outliers and leverage points. **Journal of the American Statistical Association,** New York, v. 85, n. 411, p. 633–651, Sept. 1990.

THEIL, H. A rank-invariant method of linear and polynomial regression analysis. **Verhandelingen der Koninklijke Nederlandsche Akademie van Wetenschappenes**, v. 53, pt. 1-3, p. 386–392, p. 521–525, p. 1397–1412, 1950.

TODOROV, V.; FILZMOSER, P. An object-oriented framework for robust multivariate analysis. **Journal of Statistical Software,** New York, v. 32, n. 3, p. 1–47, Oct. 2009.

ANEXOS

ANEXOS **Páginas**

ANEXO A

ANEXO A - Tabela dos valores médios dos teores de trigonelina, ácidos clorogênicos, cafeína ($\%$ m.s) e dos atributos sensoriais fragrância, sabor, acidez, corpo e equilíbrio, dos 14 genótipos de café referentes às diferentes altitudes.

Tabela 6 Valores médios dos teores de trigonelina (X_1), ácidos clorogênicos (X_2) e cafeína (X_3) ($\%$ m.s); dos atributos sensoriais fragrância (Y_1), sabor (Y_2), acidez (Y_3), corpo (Y_4) e equilíbrio (Y_5), dos 14 genótipos de café referentes às diferentes altitudes.

$Altitude*$	$trat.$	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4	Y_5
950	1	0,96	6,21	1,10	7,54	7,38	7,25	7,33	6,83
	2	0,91	5,16	1,10	7,38	7,17	7,33	7,25	7,08
	3	0,93	5,80	1,19	7,21	7,25	7,17	7,21	6,83
	4	1,00	5,86	1,29	7,25	7,25	7,25	7,29	7,00
	5	0,93	5,90	1,10	7,71	7,33	7,38	7,21	7,04
	6	0,93	6,14	1,14	7,50	7,25	7,21	7,13	7,04
	7	0,95	5,58	1,16	7,71	7,54	7,29	7,33	7,13
	8	0,89	5,84	1,12	7,58	7,04	7,04	7,38	7,00
	9	0,90	6,05	1,09	7,42	7,29	7,38	7,46	7,00
	10	0,96	5,88	1,10	7,50	7,29	7,50	7,29	7,04
	11	0,89	6,02	1,12	7,25	7,17	7,33	7,29	7,00
	12	0,88	5,67	1,16	7,50	7,38	7,25	7,19	7,13
	13	0,83	5,58	1,18	7,17	7,13	7,25	7,17	6,92
	14	0,91	5,72	1,15	7,04	6,92	7,08	7,29	6,88
1300	1	1,05	6,14	1,23	7,13	7,29	7,29	7,08	7,13
	2	1,05	5,92	1,23	7,42	7,17	7,25	7,17	7,00
	3	1,11	6,06	1,37	7,25	7,04	7,13	7,13	6,96
	4	1,08	5,70	1,39	7,13	6,94	7,13	7,00	7,13
	5	1,06	5,96	1,26	7,63	7,44	7,50	7,50	7,38
	6	1,08	6,16	1,19	7,13	6,94	7,06	7,06	6,75
	7	1,02	5,86	1,23	7,33	7,17	7,25	7,17	7,04
	8	1,08	5,98	1,26	7,50	7,29	7,13	7,38	7,13
	9	1,04	5,89	1,22	7,63	7,58	7,46	7,42	7,33
	10	1,19	5,72	1,28	7,33	7,21	7,38	7,08	7,17
	11	1,00	5,76	1,18	7,38	7,42	7,25	7,21	6,96

$Altitude*$	$trat.$	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4	Y_5
1300	12	1,01	5,77	1,20	7,42	7,21	7,33	7,13	7,00
	13	0,92	5,53	1,22	7,83	7,67	7,50	7,17	7,42
	14	1,13	5,47	1,31	7,58	7,50	7,46	7,21	7,25
1050	1	1,09	6,66	1,20	7,46	7,21	7,33	7,13	6,88
	2	1,06	6,04	1,23	7,08	7,00	7,13	7,04	6,83
	3	1,07	6,28	1,20	7,42	7,21	7,17	7,21	7,00
	4	1,11	6,08	1,19	7,00	6,71	6,92	7,04	6,83
	5	0,99	6,34	1,16	7,50	7,08	7,29	7,17	6,83
	6	1,02	6,49	1,14	7,42	7,08	7,29	7,42	7,08
	7	1,04	5,78	1,18	7,17	6,96	7,17	7,13	6,96
	8	0,98	5,93	1,07	7,17	6,83	7,04	7,08	6,79
	9	1,00	6,09	1,14	7,58	7,33	7,21	7,21	7,00
	10	1,02	5,79	1,06	7,33	7,25	7,21	7,29	6,88
	11	1,06	6,46	1,16	7,75	7,29	7,54	7,25	7,13
	12	1,08	6,30	1,12	7,33	7,08	7,33	7,21	7,04
	13	0,91	5,58	1,15	7,29	7,00	7,25	7,17	6,92
	14	1,03	6,48	1,18	7,33	7,25	7,33	7,17	7,00

*Altitudes de 950 m (Lavras), 1.300 m (São Sebastião da Grama) e 1.050 m (Santo Antônio do Amparo).

ANEXO B

ANEXO B - Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ em função de diferentes tamanhos amostrais n, considerando $p = 5$ e $p = 10$ variáveis, taxas de mistura fixadas em δ igual a $0{,}05; 0{,}10; 0{,}20;\ 0{,}30$ das distribuições *t-student* e *log-normal*, utilizando os algoritmos *MCD* (A e B) e *MVE* (C e D).

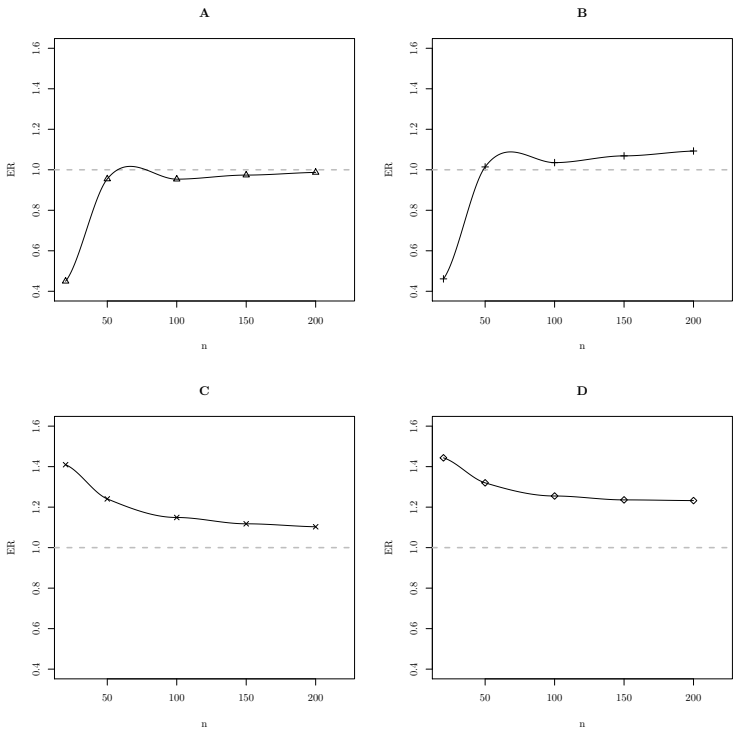

Figura 13 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 10$ variáveis, taxa de mistura $\delta = 0{,}05$

.

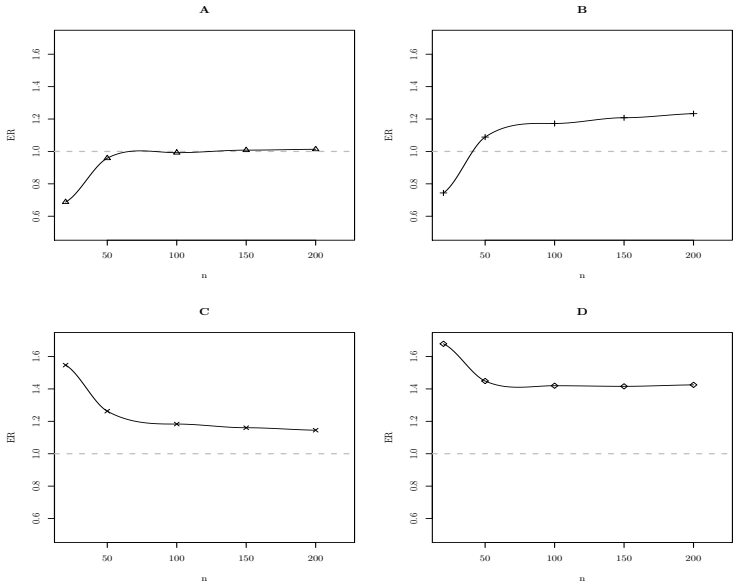

Figura 14 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 5$ variáveis, taxa de mistura $\delta = 0{,}10$

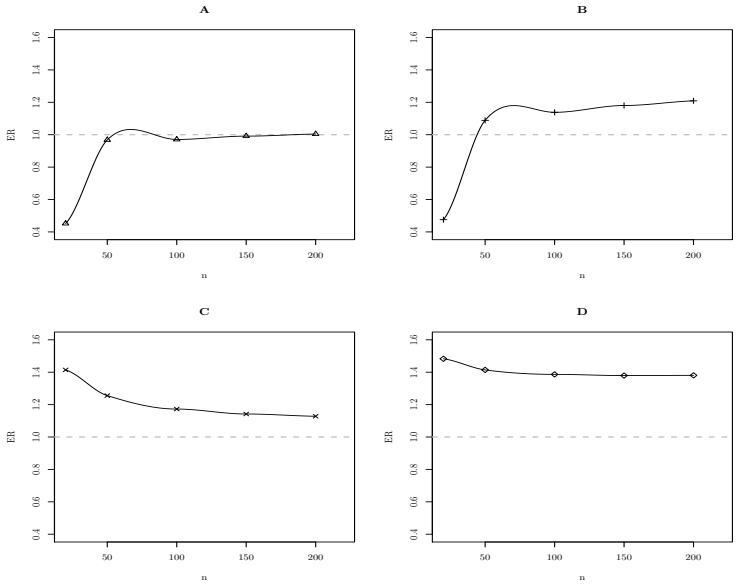

Figura 15 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 10$ variáveis, taxa de mistura $\delta = 0{,}10$

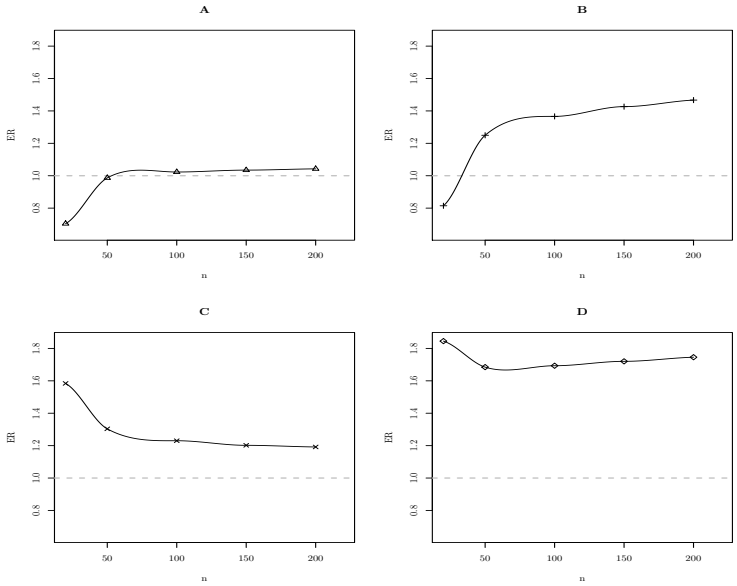

Figura 16 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 5$ variáveis, taxa de mistura $\delta = 0{,}20$

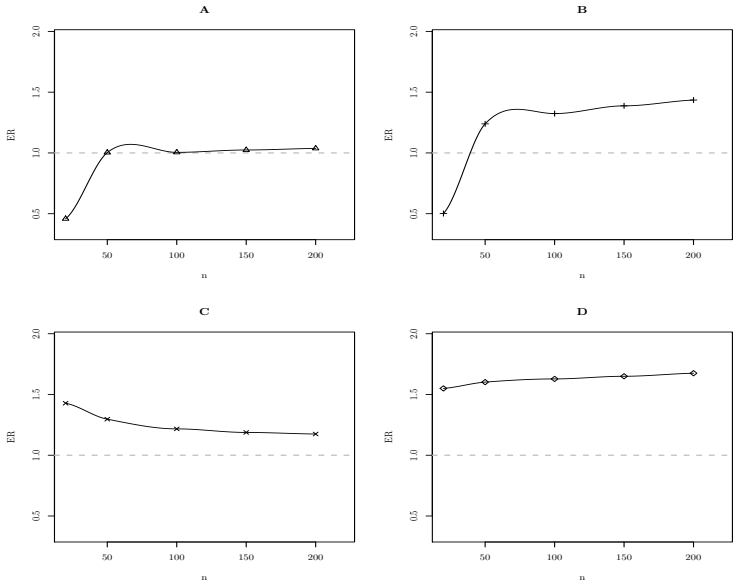

Figura 17 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 10$ variáveis, taxa de mistura $\delta = 0{,}20$

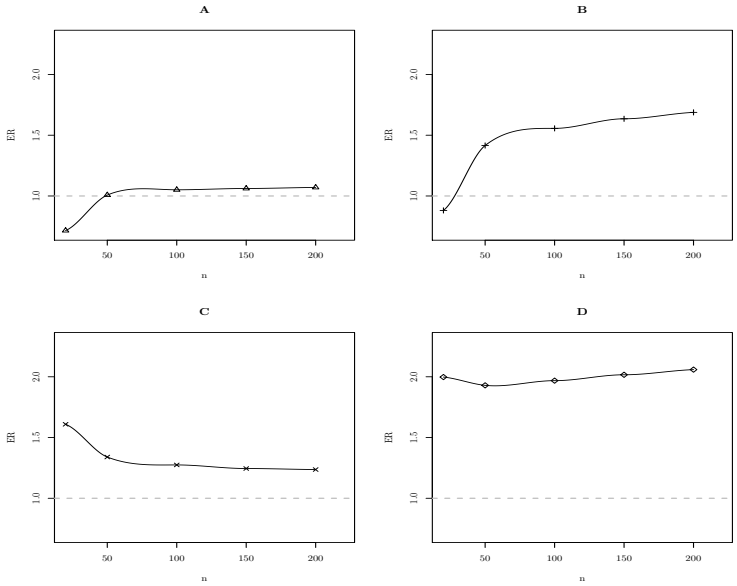

Figura 18 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 5$ variáveis, taxa de mistura $\delta = 0{,}30$

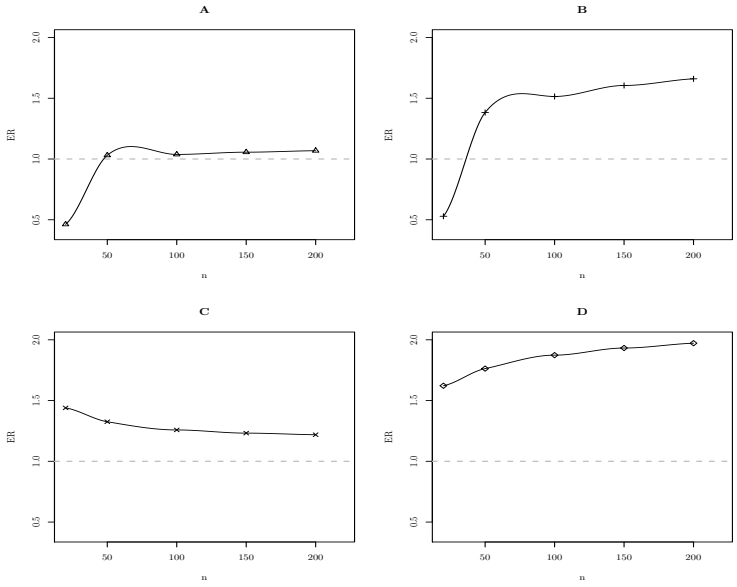

Figura 19 Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$, $p = 10$ variáveis, taxa de mistura $\delta = 0{,}30$

ANEXO C

ANEXO C - Representação gráfica do comportamento da eficiência relativa $\boldsymbol{ER_D}$ (A e B), em comparação com a eficiência relativa $\boldsymbol{ER_T}$ (C e D) em função de diferentes tamanhos amostrais n, considerando $p = 5$ e $p = 10$ variáveis, taxas de mistura fixadas em δ igual a $0{,}05; 0{,}10; 0{,}20; 0{,}30$ das distribuições *t-student* e *log-normal*, utilizando o algoritmos *MCD* e *MVE*

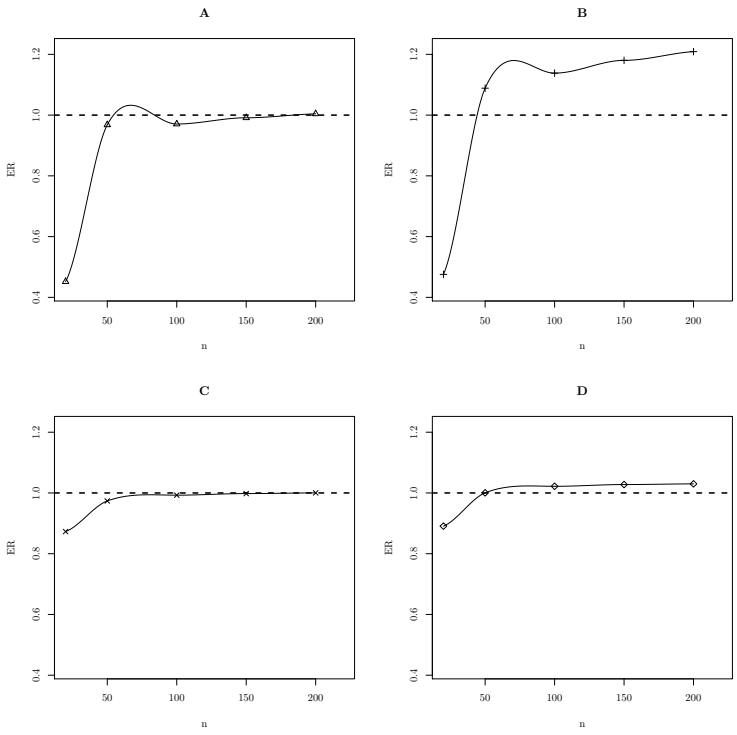

Figura 20 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}10$

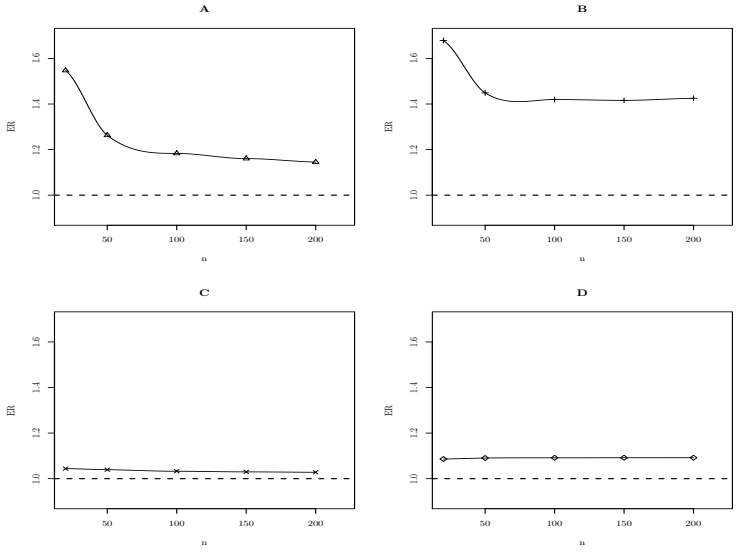

Figura 21 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}10$

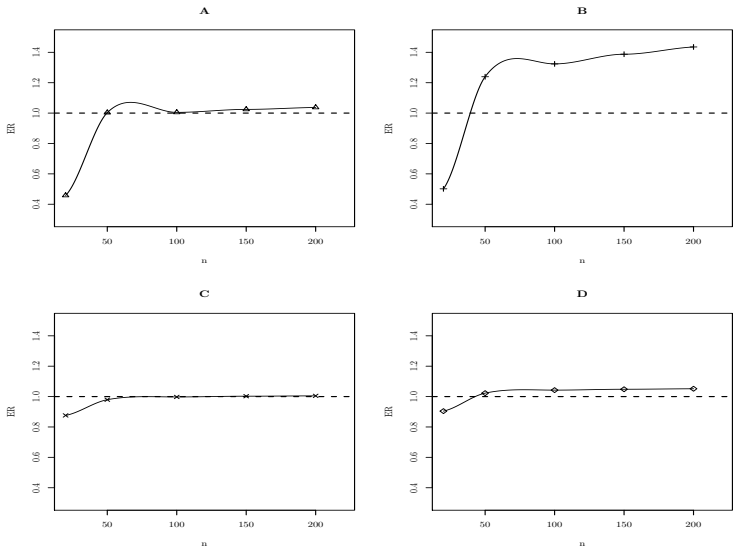

Figura 22 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}20$

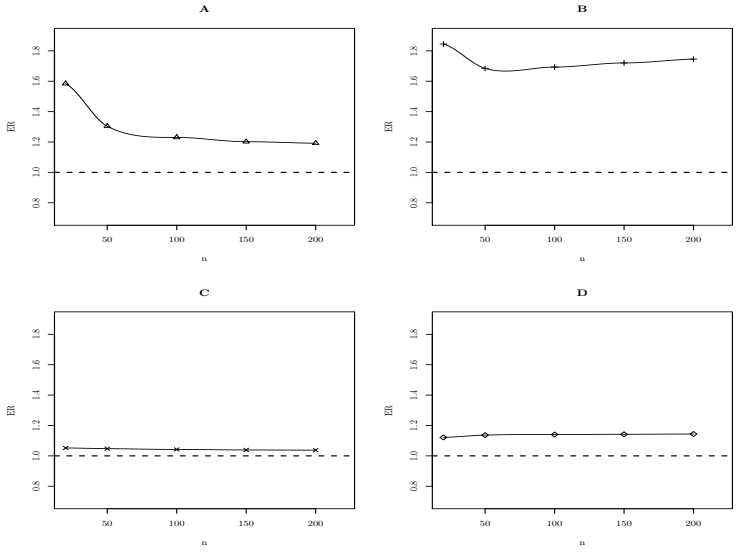

Figura 23 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}20$

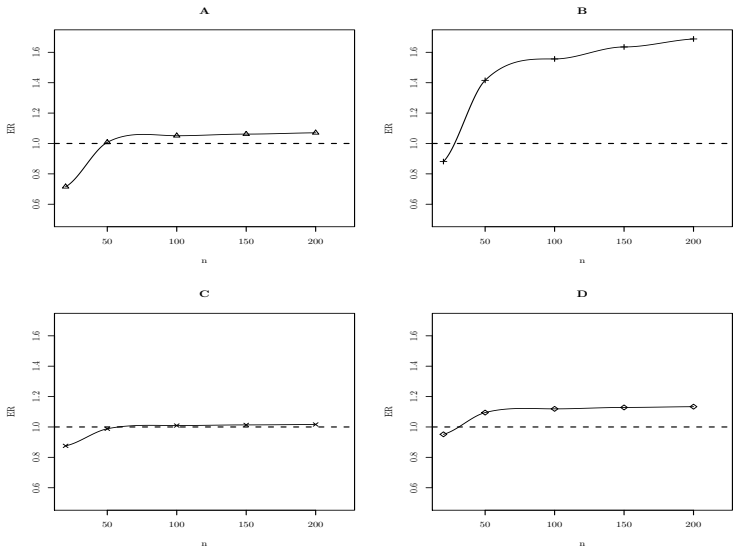

Figura 24 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}30$

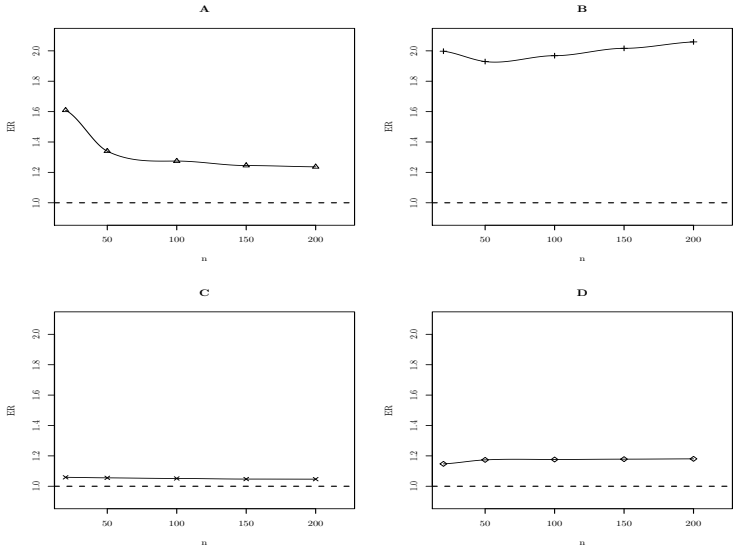

Figura 25 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 5$ variáveis, taxa de mistura $\delta = 0{,}30$

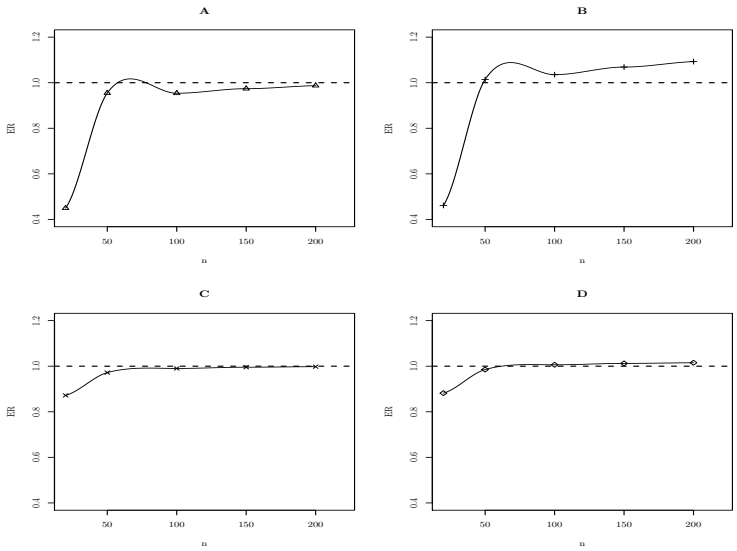

Figura 26 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}05$

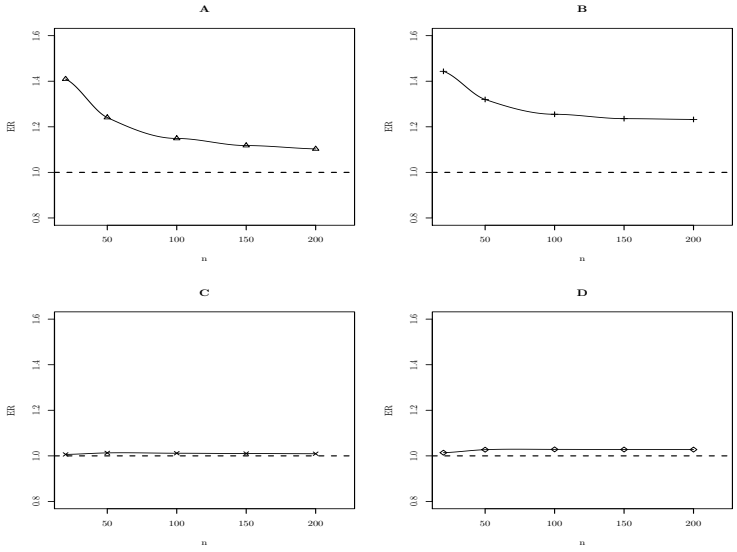

Figura 27 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}05$

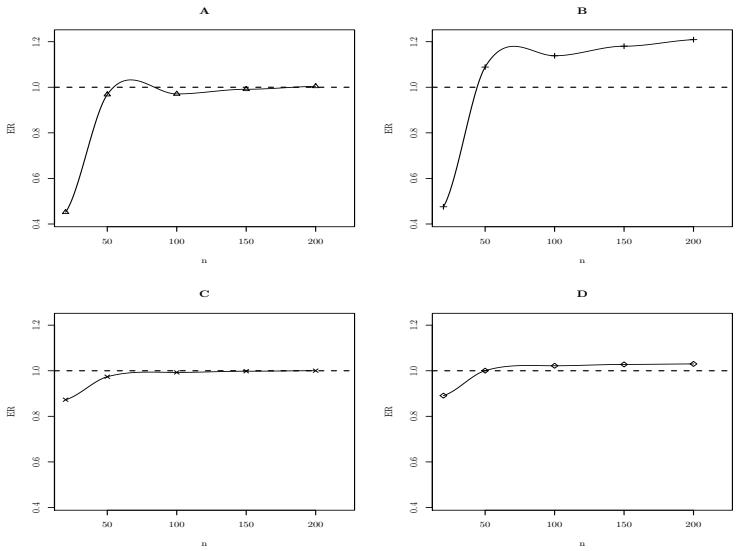

Figura 28 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}10$

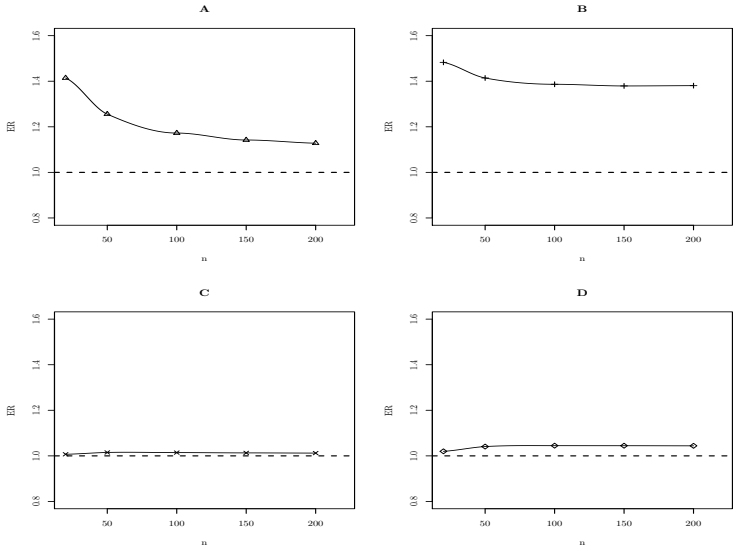

Figura 29 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}10$

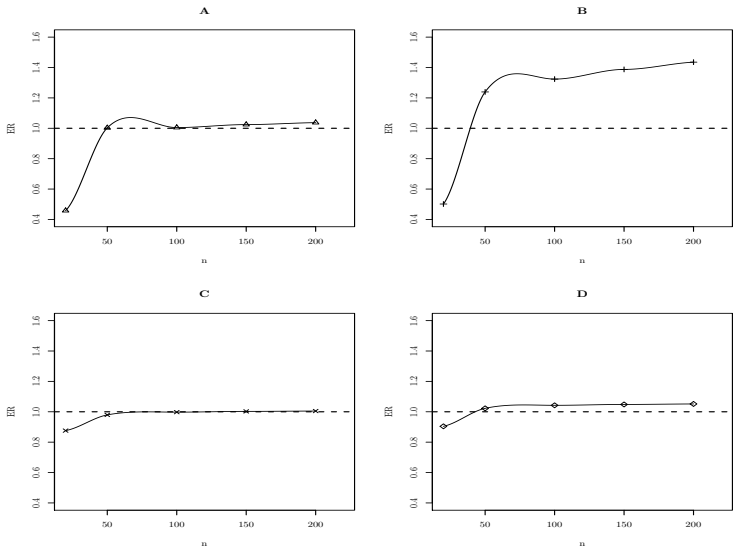

Figura 30 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}20$

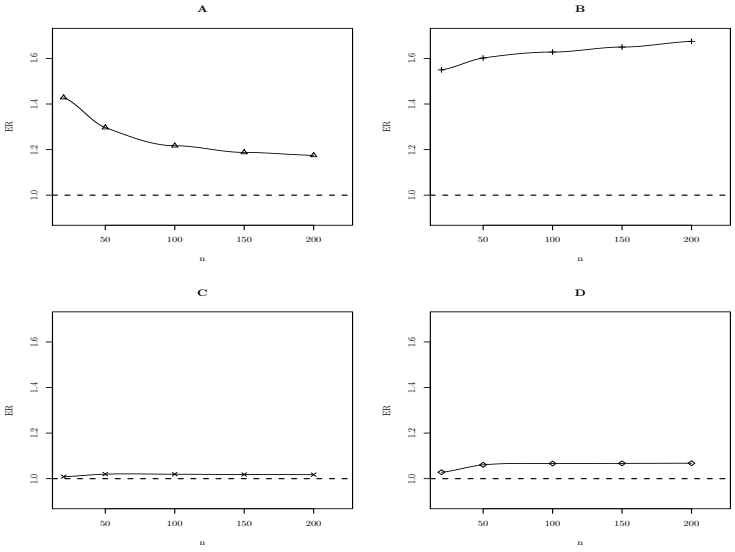

Figura 31 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}20$

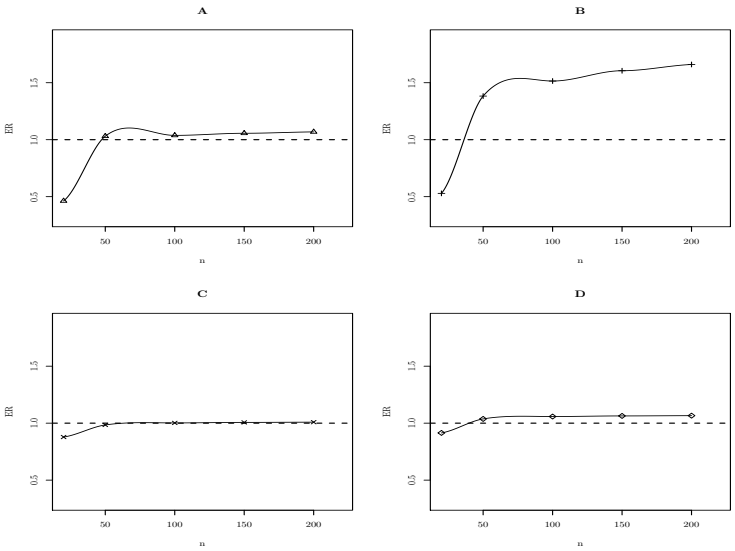

Figura 32 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}30$

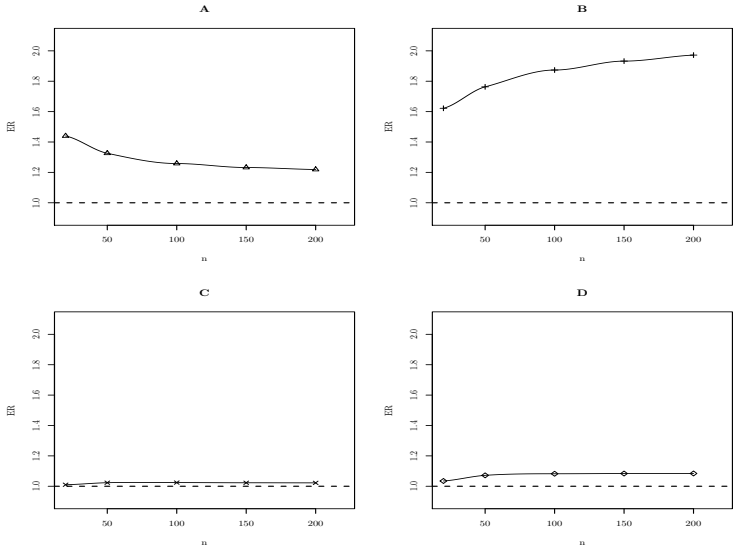

Figura 33 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}30$

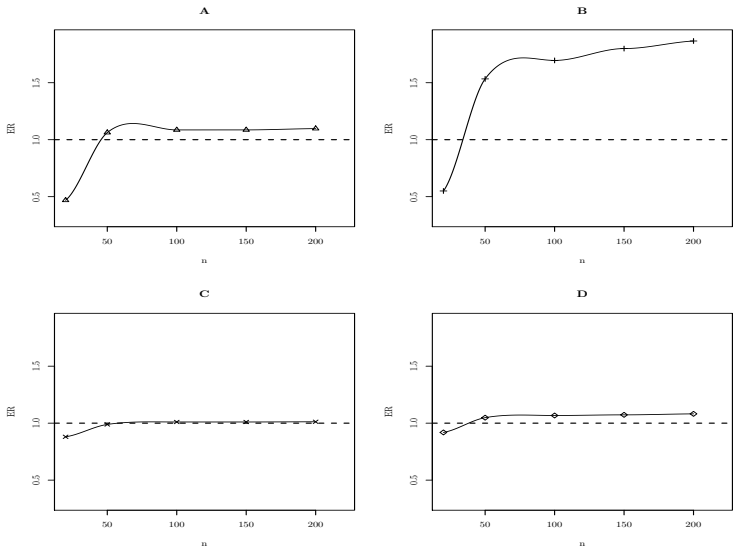

Figura 34 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}40$

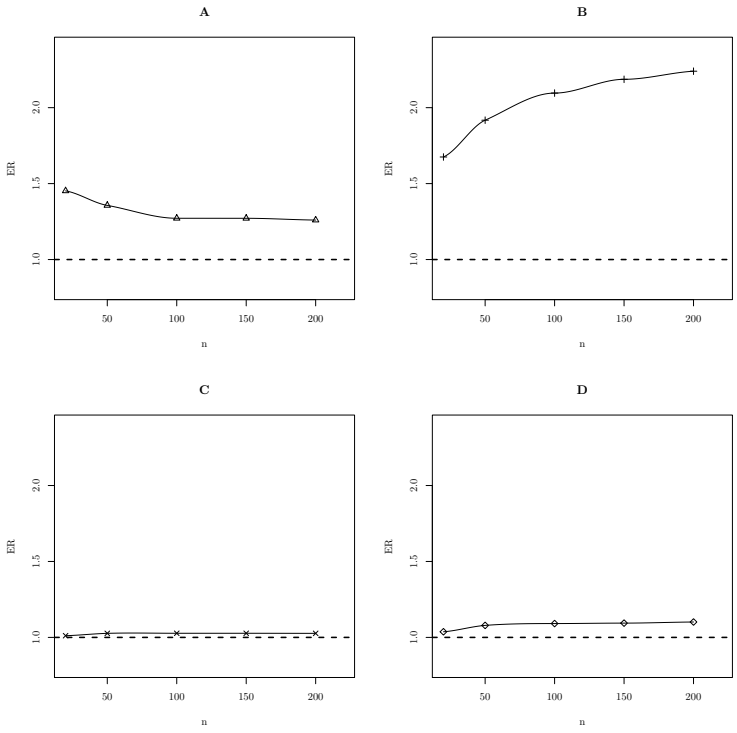

Figura 35 Representação gráfica do comportamento da eficiência relativa $\mathbf{ER_D}$ (A e B), em comparação com a eficiência relativa $\mathbf{ER_T}$ (C e D), $p = 10$ variáveis, taxa de mistura $\delta = 0{,}40$

ANEXO D

ANEXO D - Programa R de simulação utilizado para obtenção dos resultados

---

```
#### pacotes requeridos ####
library(mvtnorm)
library(robustbase)
library(rrcov)
library(car)

#### parâmetros do modelo ####

p<-5                ######      número de variáveis dependentes
k<-3                ######      numero de variáveis independentes
n<-20               ######      tamanho amostral
nsim<-10000         ######      número de simulação
delta<- 0.05        ######      prob. da mistura
rho<-0.5            ######      grau da correlação

#### Definiçõa das funções ####

X<-matrix(c(rep(1,n),runif(n), runif(n), runif(n)),ncol=k+1)
Beta<-matrix(c(c(1,0.5,0.5,1),c(-1,-0.5,-0.5,-1), c(-0.5,-0.5,1,1)
               , c(1,-0.5,0.5,-1), c(-1,-1,0.5,-0.5)), ncol=p)

AR<-matrix(0,p,p)    ;      eco<-matrix(0,p,p)  ;
mi<-c(rep(0,p))     ;       mresul<-matrix(0,nsim,4)

#### Cálculo da eficiência relativa ####

 for (i in 1:p)

 {
      for (j in 1:p)

 {
         if (i==j)  AR[i,j]=1
         if (i!=j)  AR[i,j]=rho^(abs(i-j))

         if (i==j)  eco[i,i]=1
         if (i!=j)  eco[i,j]=rho
  }

}

    covp1 <- AR  ; covp2<-eco
```

```
for (i in 1:nsim)

{

     res<-matrix(0,1,p)
     for (r in 1:n)

  {

        u<-runif(1)
        obs<- rmvnorm(1,mean=mi,sigma=covp1)
        obst<- rmvt(1, sigma=covp2, df=5)

        if (u<=delta)

          {
             # Erro <-exp(obs)    ###  observações discrepantes
                                      geradas pela log-normal
               Erro <-obst        ###  observações discrepantes
                                      geradas pela t-student
          }

        if (u>delta)     Erro <- obs

        res<-rbind(Erro,res)

   }

       res<-res[1:nrow(res)-1,]
       Y<-X%*%Beta + res

       mcd <- covMcd(Y,alpha = 0.50,nsamp=10000,use.correction=T)
       mve <- cov.mve(Y,nsamp=10000)

         S_mcd<-mcd$cov  ;   S <-cov(Y) ; S_mve<-mve$cov

         i_mcd_t<-(sum(diag(S_mcd)) / sum(diag(S)))^(1/p)
         i_mcd_d<-(det(S_mcd) / det(S))^(1/p)
         i_mve_t<-(sum(diag(S_mve)) / sum(diag(S)))^(1/p)
         i_mve_d<-(det(S_mve) / det(S))^(1/p)

         mresul[i,1]<-i_mcd_t  ;  mresul[i,2]<-i_mcd_d ;
         mresul[i,3]<-i_mve_t  ;  mresul[i,4]<-i_mve_d

}
    summary(mresul)
```

ANEXO E

ANEXO E - Script para o cálculo das estimativas dos modelos de regressão ajustados

---

```
###### Cálculo das estimativas dos modelos de regressão
###### ajustados para a análise sensorial da qualidade do café

#### pacotes requeridos
require(MASS)
require(stats)
require(robustbase)
require(robust)
dados_cafe <- read.table("dados_final.txt",h=T)
dados <- dados_cafe[,3:10]


############## Método dos mínimos quadrados (MMQ) ###############

S <- ccov(dados)

S1 <- S$cov
SXX <- S1[1:3,1:3]
SXY <- S1[1:3,4:8]
SYY <- S1[4:8,4:8]

mu <- S$center
muX <- mu[1:3]
muY <- mu[4:8]

#### estimativas dos coeficientes de regressão obtidas
#### pelo método MMQ

B <- solve(SXX) %*% SXY

B0 <- muY - t(B) %*% muX

Se <- SYY - t(B) %*% SXX %*% B
e <-  as.matrix(c(mean(Se[,1]), mean(Se[,2]), mean(Se[,3]),
      mean(Se[,4]), mean(Se[,5])))


#### valores médios preditos obtidos pelo método MMQ

Y <- t(B) %*% muX + B0 + e
```

```
####### Método covariância de determinante mínimo (MCD) #######

mcd <- cov.rob(dados)

Sr <- mcd$cov
SXX_r <- Sr[1:3,1:3]
SXY_r <- Sr[1:3,4:8]
SYY_r <- Sr[4:8,4:8]

mu1 <- mcd$center
muX_r <- mu1[1:3]
muY_r <- mu1[4:8]

#### estimativas dos coeficientes de regressão obtidas
#### pelo método MCD

B_r <- solve(SXX_r) %*% SXY_r

B0_r <- muY_r - t(B_r) %*% muX_r

Se_r <- SYY_r - t(B_r) %*% SXX_r %*% B_r
e_r <- as.matrix(c(mean(Se_r[,1]), mean(Se_r[,2]),
       mean(Se_r[,3]), mean(Se_r[,4]), mean(Se_r[,5])))

#### valores médios preditos obtidos pelo método MCD

Y_r <- t(B_r) %*% muX  + B0_r + e_r


####### Método elipsoide de volume mínimo (MVE) ########

mve <- cov.rob(dados)

Sr1 <- mve$cov

SXX_r1 <- Sr1[1:3,1:3]
SXY_r1 <- Sr1[1:3,4:8]
SYY_r1 <- Sr1[4:8,4:8]

mu2 <- mve$center
muX_r1 <- mu2[1:3]
muY_r1 <- mu2[4:8]

#### estimativas dos coeficientes de regressão obtidas
#### pelo método MVE

B_r1 <- solve(SXX_r1) %*% SXY_r1
```

```
B0_r1 <- muY_r1 - t(B_r1) %*% muX_r1

Se_r1 <- SYY_r1 - t(B_r1) %*% SXX_r1 %*% B_r1
e_r1 <- as.matrix(c(mean(Se_r1[,1]), mean(Se_r1[,2]),
        mean(Se_r1[,3]), mean(Se_r1[,4]), mean(Se_r1[,5])))

#### valores médios preditos obtidos pelo método MVE

Y_r1 <- t(B_r1) %*% muX + B0_r1 + e_r1

########## método MMQ####################
S
B
B0
Y

########## método MCD ###################
mcd
B_r
B0_r
Y_r

########## método MVE ###################
mve
B_r1
B0_r1
Y_r1
```