

DELINEAMENTOS D -ÓTIMOS EXATOS SOB A POSSIBILIDADE DE TRANSFORMAÇÃO DA VARIÁVEL RESPOSTA

Cássio Pinho dos REIS¹
Luzia Aparecida TRINCA²

- RESUMO: As mais diversas áreas da ciência evoluem com o conhecimento adquirido por experimentação. O uso de delineamentos experimentais eficientes representa ganhos em termos tanto de custo financeiro e tempo como em precisão nos resultados. Os delineamentos ótimos, que focam na eficiência de propriedades específicas, são alternativas de fundamental importância quando o problema prático envolve alguma complexidade não contemplada diretamente na formulação clássica. A maioria dos métodos pressupõe homogeneidade de variâncias, a qual nem sempre é verificada no conjunto de dados. O objetivo deste trabalho é desenvolver uma metodologia para construção de delineamentos D -ótimos exatos em situações de variância não homogênea. Assume-se que linearidade e homoscedasticidade são obtidas via o uso de transformações da família Box-Cox na variável resposta. Resultados sob vários modelos de primeira e segunda ordem são obtidos e discutidos.
- PALAVRAS-CHAVE: Delineamentos eficientes; heterocedasticidade; transformação de Box-Cox; fatorial.

1 Introdução

Métodos de construção de delineamentos ótimos tem grande aplicação principalmente quando as condições experimentais apresentam limitações e/ou o modelo estatístico pressuposto para a análise apresenta alguma complexidade. A teoria foi formalizada por Kiefer (1959), no contexto assintótico (contínuo) e, com a evolução computacional, tem avançado nas aplicações. Para sua realização o experimento requer a versão exata (discreta) do delineamento, ou seja, números

¹Universidade Federal de Mato Grosso do Sul - UFMS, Instituto de Matemática, CEP: 79074-460, Campo Grande, MS, Brasil. E-mail: *cassio.reis@ufms.br*

²Universidade Estadual Paulista - UNESP, Departamento de Bioestatística, CEP: 18618-689, Botucatu, SP, Brasil. E-mail: *luzia.trinca@unesp.br*

de unidades experimentais e repetições de tratamentos finitos e inteiros. Um delineamento exato ótimo é aquele cujo conjunto de tratamentos e respectivas repetições otimizam alguma propriedade desejada da análise estatística. Por conseguinte, sua busca exige a postulação de um modelo de análise dos dados. A busca por esse ótimo, exceto para situações muito simples, ocorre via aplicação de algum algoritmo computacional. Para maiores detalhes sobre as duas formulações de delineamento (contínuo ou exato), propriedades, critérios e algoritmos utilizados para a construção ver Atkinson et al. (2007).

O modelo mais simples para estudar a relação entre uma variável resposta e diversos fatores experimentais é o linear homocedástico (variâncias dos erros homogêneas). Contudo, nem sempre essa pressuposição é satisfeita, o que é frequente nas análises de dados. Não raro, normalidade e homogeneidade podem ser razoavelmente alcançadas via transformações da família Box-Cox (BOX e COX, 1964) na variável resposta, sendo esta alternativa bastante utilizada na prática, principalmente quando a variável resposta é positiva contínua. Sob o modelo de transformação, a matriz de informação apresenta os mesmos problemas de modelos não lineares, o que leva ao requerimento de métodos específicos de delineamento (CHALONER e LARNTZ, 1989; GOTWALT et al., 2009; GILMOUR e TRINCA, 2012; FERREIRA et al., 2014; OVERSTALL e WOODS, 2017; GOOS e MYLONA, 2018).

O problema de planejamento experimental sob heterocedasticidade foi estudado por Atkinson e Cook (1995, 1997) e Gaviria e Ríos (2014) no contexto de delineamentos *D*-ótimos contínuos para fatores quantitativos. Atkinson e Cook (1995) consideraram o modelo com ambas partes, média e variância, expressas por funções paramétricas dos níveis dos fatores quantitativos enquanto que Atkinson e Cook (1997) exploraram o modelo sob transformação. Gaviria e Ríos (2014) compararam as duas abordagens, com um único fator, via simulação. Do ponto de vista de delineamento, a vantagem da estratégia de transformação está no acréscimo de apenas um parâmetro extra ao modelo, o de transformação, contrastada com a estratégia de modelagem das variâncias. Outra alternativa seria a suposição de uma distribuição de probabilidades que incorpora a heterocedasticidade da variável resposta, e proceder via a teoria de modelos lineares generalizados. Enquanto que para variáveis resposta discretas as distribuições que regem seus comportamentos possam ser elucidadas ainda na fase de planejamento, maiores dificuldades são encontradas para as variáveis contínuas. Por exemplo, a distribuição Gama poderia, em princípio, ser uma candidata ao modelo, porém, na fase de ajuste do modelo, ainda pode-se deparar com problemas devido à relação entre média e variância requerida por esse modelo.

O objetivo desse trabalho é o de estender a metodologia apresentada em Atkinson e Cook (1997) para a construção de delineamentos ótimos e/ou eficientes ao caso exato/discreto quando, na fase de planejamento, o pesquisador suspeita de heterocedasticidade. Assume-se que, após alguma transformação, com parâmetro desconhecido, da variável resposta, o modelo de análise dos dados se simplifica ao modelo linear.

2 O problema de delineamento para o modelo sob transformação

Seja $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{qi})$ a combinação dos níveis dos q fatores que definem o tratamento aplicado na unidade experimental i . Para uma variável resposta contínua, os modelos mais populares assumem a forma

$$Y_i = \eta(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad (1)$$

na qual Y_i é a variável resposta a ser observada na unidade experimental i ($i = 1, 2, \dots, n$), $\boldsymbol{\beta}$ é o vetor de p parâmetros relacionados aos fatores cujos efeitos deseja-se estudar, η é uma função em \mathbf{x}_i e $\boldsymbol{\beta}$ que descreve as formas das relações entre Y e os q fatores e ε_i é o i -ésimo componente aleatório aditivo, usualmente suposto Normal, com média nula, variância constante e não correlacionado a nenhum outro $\varepsilon_{i'}$. Se η for linear em $\boldsymbol{\beta}$ o modelo se reduz, em notação matricial, à

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

na qual $\mathbf{Y}_{n \times 1}$ é o vetor da variável aleatória, $\mathbf{X}_{n \times p}$ é a matriz do modelo, $\boldsymbol{\beta}_{p \times 1}$ é o vetor de parâmetros de regressão, $\boldsymbol{\varepsilon}_{n \times 1}$ o vetor de erros, com $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. É bem sabido que a matriz de variâncias e covariâncias do estimador de mínimos quadrados de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, é $V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$ e como σ^2 é uma constante que não depende de \mathbf{X} , delineamentos ótimos são obtidos otimizando-se alguma propriedade de $\mathbf{M} = (\mathbf{X}^T \mathbf{X})$, a matriz de informação para $\boldsymbol{\beta}$ sob o modelo linear homocedástico.

Na prática, uma ou mais violações das suposições associadas ao modelo em (1) podem ocorrer e, ocasionalmente, uma transformação em Y pode tornar o uso do método de mínimos quadrados aceitável. Embora nem sempre todas as violações possam ser corrigidas com uma única transformação (SAKIA, 1992), diversos autores reconhecem o potencial da família Box-Cox (BOX e COX, 1964) na estabilização de variâncias quando essas aumentam com $\mathbb{E}(Y)$ (ATKINSON e COOK, 1995, 1997, GAVIRIA e RÍOS, 2014 e DAVISON, 2003, para citar alguns). Existem algumas variações ou extensões da função de transformações Box-Cox sendo que a originalmente apresentada por Box e Cox (1964), assumindo $y > 0$, é dada por

$$Y^{(\lambda)} = \begin{cases} \frac{(Y^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0, \end{cases} \quad (3)$$

na qual λ é o parâmetro de transformação. A expressão (3) abrange várias transformações usualmente utilizadas na prática como a logarítmica ($\lambda = 0$), raiz quadrada ($\lambda = 1/2$), raiz cúbica ($\lambda = 1/3$), recíproca ($\lambda = -1$), e assim por diante. Após estimar λ , assume-se o modelo linear na escala transformada, com $\varepsilon_i \sim N(0, 1)$, dado por

$$Y_i^{(\lambda)} = f^T(\mathbf{x}_i) \boldsymbol{\beta} + \sigma \varepsilon_i, \quad (4)$$

no qual $f(\mathbf{x})_{p \times 1}$ é uma função aplicada ao vetor \mathbf{x}_i de forma a acomodar os efeitos

em $\boldsymbol{\beta}$, $i = 1, 2, \dots, n$. Para fins de construção do delineamento que leva em conta a informação para estimação de todos os $p^\dagger = (p + 2)$ parâmetros do modelo, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma^2)^\top$, é necessário a função de densidade de Y_i dada por (COOK e WEISBERG, 1982)

$$k\left(y_i^{(\lambda)} | (\boldsymbol{\theta}, \mathbf{x}_i)\right) = \frac{1}{\sigma\sqrt{2\pi}} \left\| \frac{dy_i^{(\lambda)}}{dy_i} \right\| \exp \left\{ -\frac{\left[y_i^{(\lambda)} - f^\top(\mathbf{x}_i)\boldsymbol{\beta} \right]^2}{2\sigma^2} \right\}, \quad (5)$$

resultante da normalidade assumida para a variável transformada.

2.1 Matriz de informação e o critério D

A não linearidade em (5) devido ao parâmetro λ (desconhecido) leva a uma expressão da matriz de informação bem mais complexa do que aquela para o modelo em (2). Por definição (CASELA e BERGER, 2001), para cada observação i , a informação de Fisher é dada por

$$\mathbf{M}_\lambda^i(\boldsymbol{\theta}, \mathbf{x}_i) = -\mathbb{E} \left\{ \frac{\partial^2 \log \left[k\left(y_i^{(\lambda)} | (\boldsymbol{\theta}, \mathbf{x}_i)\right) \right]}{\partial \boldsymbol{\theta}^2} \right\}. \quad (6)$$

Como, em geral, não existe solução explícita para a esperança em (6), aproximações por expansão em séries de Taylor são utilizadas (ATKINSON e COOK, 1992; ATKINSON et al., 2007) tal que que, para a observação i , tem-se a matriz simétrica

$$\mathbf{M}_\lambda^i(\boldsymbol{\theta}, \mathbf{x}_i) = \begin{pmatrix} \frac{f(\mathbf{x}_i)f^\top(\mathbf{x}_i)}{\sigma^2} & 0 & -\frac{f(\mathbf{x}_i)c_\theta(\mathbf{x}_i)}{\sigma^2} \\ \cdot & \frac{1}{2\sigma^4} & -\frac{\log \mu(\boldsymbol{\theta}, \mathbf{x}_i)}{\lambda\sigma^2} \\ \cdot & \cdot & 2\frac{\log^2 \mu(\boldsymbol{\theta}, \mathbf{x}_i)}{\lambda^2} + \frac{(c_\theta(\mathbf{x}_i))^2}{\sigma^2} \end{pmatrix}_{(p+2) \times (p+2)} \quad (7)$$

para $\lambda \neq 0$, na qual $\mu(\boldsymbol{\theta}, \mathbf{x}_i) = \mathbb{E}(Y_i^\lambda) = \lambda f^\top(\mathbf{x}_i)\boldsymbol{\beta} + 1$ e $c_\theta = (\mu(\boldsymbol{\theta}, \mathbf{x}_i) \log \mu(\boldsymbol{\theta}, \mathbf{x}_i) - \mu(\boldsymbol{\theta}, \mathbf{x}_i) + 1)/\lambda^2$.

Para $\lambda = 0$ a expressão em (7) se simplifica em

$$\mathbf{M}_\lambda^i(\boldsymbol{\theta}, \mathbf{x}_i) = \begin{pmatrix} \frac{f(\mathbf{x}_i)f^T(\mathbf{x}_i)}{\sigma^2} & 0 & -\frac{f(\mathbf{x}_i)(f^T(\mathbf{x}_i)\boldsymbol{\beta})^2}{2\sigma^2} \\ \cdot & \frac{1}{2\sigma^4} & -\frac{f^T(\mathbf{x}_i)\boldsymbol{\beta}}{\sigma^2} \\ \cdot & \cdot & 2(f^T(\mathbf{x}_i)\boldsymbol{\beta})^2 + \frac{(f^T(\mathbf{x}_i)\boldsymbol{\beta})^4}{4\sigma^2} \end{pmatrix}_{(p+2) \times (p+2)} \quad (8)$$

Seja \mathbf{X} a matriz cuja i -ésima linha é $f^T(\mathbf{x}_i)$ que, para simplicidade, representará o delineamento. Então, para as n observações, a matriz de informação total aproximada é

$$\mathbf{M}_\lambda(\boldsymbol{\theta}, \mathbf{X}) = \sum_{i=1}^n \mathbf{M}_\lambda^i(\boldsymbol{\theta}, \mathbf{x}_i), \quad (9)$$

na qual o índice λ em \mathbf{M} tem a finalidade de explicitar a matriz de informação baseada no modelo com transformação.

Para n fixo, a construção do delineamento ótimo exato consiste na busca dos tratamentos \mathbf{x} 's, que irão formar as linhas da matriz \mathbf{X} , otimizando alguma propriedade de $\mathbf{M}_\lambda(\boldsymbol{\theta}, \mathbf{X})$. Existem diversas propriedades (critérios de otimalidade) consideradas de potencial interesse, em geral, relacionadas com a precisão dos estimadores dos parâmetros. Uma lista bastante completa pode ser consultada em Atkinson et al. (2007). Nesse trabalho será destacada a mais popular delas, dada pela variância generalizada dos estimadores dos parâmetros, conhecida por critério D .

O delineamento D -ótimo é, portanto, aquele que maximiza o determinante da matriz de informação, ou seja, maximiza a função critério

$$|\mathbf{M}_\lambda(\boldsymbol{\theta}, \mathbf{X})|^{\frac{1}{p^\dagger}}, \quad (10)$$

ou equivalente,

$$\frac{1}{p^\dagger} \log |\mathbf{M}_\lambda(\boldsymbol{\theta}, \mathbf{X})|, \quad (11)$$

esta última preferida a fim de se evitar problemas numéricos (valores muito altos). Note que como \mathbf{M}_λ depende dos valores dos parâmetros, a construção do delineamento requer a incorporação de valores *a priori* para os mesmos.

2.2 Informação *a priori*

A matriz de informação cujo modelo envolve alguma não linearidade nos parâmetros é sempre dependente dos valores desses parâmetros, tornando a busca de delineamentos mais difícil. A alternativa é o uso de informações obtidas *a priori*,

seja de experimentos prévios ou de conhecimento do especialista da área. A maneira mais fácil de resolver o problema é por meio do uso de valores pontuais, o que produzirá um delineamento apenas localmente ótimo. Essa alternativa, embora simples, é útil na prática, já que, após obtenção do delineamento este pode ser avaliado sob outros valores dos parâmetros e sua robustez estudada. Esse tipo de estudo pode ser realizado pela eficiência definida na Seção 2.3.

Outras possibilidades existem, como a incorporação de distribuições de probabilidade *a priori* aos parâmetros e otimização da versão integrada da função critério, por exemplo, o critério D pseudo-Bayesiano dado por

$$\int_{\boldsymbol{\theta} \in \Theta} \frac{1}{p^\dagger} \log |\mathbf{M}_\lambda(\boldsymbol{\theta}, \mathbf{X})| p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (12)$$

no qual $p(\boldsymbol{\theta})$ é a distribuição *a priori*, multivariada, para o vetor de parâmetros. É comum o uso da suposição de independência entre os parâmetros para simplificar, porém, o processo ainda é computacionalmente caro, já que, via de regra, a integral múltipla envolvida não tem solução analítica.

2.3 Eficiência

Como o delineamento ótimo pode ser sensível à informação *a priori* utilizada, recomenda-se um estudo de sua robustez comparando seu desempenho ao ótimo obtido sob outros valores de parâmetros. Essa comparação é feita por uma medida de eficiência que, no caso do critério D , é definida por

$$\text{Eff} = \left(\frac{|\mathbf{M}_\lambda(\boldsymbol{\theta}_0, \mathbf{X})|}{|\mathbf{M}_\lambda(\boldsymbol{\theta}_0, \mathbf{X}^*)|} \right)^{1/p^\dagger}, \quad (13)$$

na qual \mathbf{X} representa o delineamento sendo investigado e \mathbf{X}^* o delineamento ótimo encontrado assumindo $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Se o valor da eficiência for menor que 1, significa que o delineamento \mathbf{X} é menos eficiente que o delineamento \mathbf{X}^* , sob o critério D .

2.4 Algoritmos para a busca de delineamentos exatos

Geralmente, heurísticas são utilizadas para a busca de delineamentos ótimos exatos. O método mais comum é o *point-exchange* (FEDOROV, 1972) e uma variação é o *coordinate-exchange* (MEYER e NACHTSHEIM, 1995). Ambos têm como ponto de partida a seleção de um delineamento inicial aleatório que forma a matriz \mathbf{X} . No *point-exchange* cada tratamento (ou linha da matriz \mathbf{X}) é, sistematicamente, trocado por um outro tratamento que pertence a um conjunto externo de tratamentos (ou pontos) candidatos. Trocas que melhoram o delineamento em termos do valor da função critério são aceitas. O procedimento para quando completa um ciclo sem nenhuma troca aceita. Para minimizar a chance do processo parar em ótimo local, o que é possível em qualquer heurística, o processo todo é repetido para vários delineamentos iniciais distintos. O número de repetições

depende da dificuldade do problema. Um detalhe do *point-exchange* é a construção do conjunto de tratamentos candidatos. Para modelos lineares costuma-se usar os tratamentos formados pelo fatorial completo com o número de níveis compatível com a ordem do modelo que se deseja ajustar. Na presença de não linearidade, a escolha do conjunto candidato é menos óbvia. Quando o número de fatores não é excessivo, o uso do fatorial completo com os níveis em espaçamento bem mais refinado ainda é viável. Na versão *coordinate-exchange*, ao invés do tratamento, quem sofre trocas é cada coordenada do ponto \mathbf{x} .

3 Metodologia

3.1 Delineamentos exatos localmente D -ótimos

Para ilustrar a obtenção de delineamentos exatos localmente D -ótimos sob o modelo em (4) fixou-se o número de fatores em dois ($q = 2$) e $n = 20$. Duas classes de modelos foram exploradas, uma com apenas efeitos lineares dos fatores e outra incluindo o termo de interação. Três valores para o parâmetro de transformação foram considerados, $\lambda = 0, \frac{1}{2}, 1$, e dois valores para o desvio padrão do erro, $\sigma = 0,1$ e $\sigma = 3$, representando variabilidade aleatória bem baixa e alta, respectivamente. Para os valores de β , considerou-se uma variedade de valores, iniciando-se com os mesmos usados por Atkinson e Cook (1997), o que permite contrastar os resultados com aqueles obtidos para delineamentos contínuos, além de outros. Para apresentação na Seção 4, apenas alguns delineamentos foram selecionados já que resumem bem o comportamento segundo o critério D . Estes são modelos simétricos ($\beta_1 = \beta_2 = 4,95$) e assimétricos ($\beta_1 = 3,30; \beta_2 = 6,60$ e $\beta_1 = -4,95; \beta_2 = 4,95$), obedecendo a restrição de que y deve ser positivo, ou seja, $\mathbb{E}(Y^{(\lambda)}) > -\frac{1}{\lambda}$ para $\lambda > 0$, conforme apontado por Atkinson e Cook (1997). Nos modelos com interação usou-se $\beta_{12} = 4,95$. O valor do intercepto foi $\beta_0 = 15$ em todos os modelos.

Para a otimização dos delineamentos utilizou-se o algoritmo *point-exchange* com conjunto de pontos candidatos dado pelo fatorial 21^2 que, numa escala codificada no intervalo $[-1; +1]$, implica em pontos equidistantes com espaçamento de 0,1 para cada fator. O processo de busca do delineamento ótimo foi repetido para 10.000 delineamentos iniciais. O algoritmo foi implementado computacionalmente no programa R (R CORE TEAM, 2019).

3.2 Delineamentos exatos pseudo-Bayesianos

Após investigação dos resultados para os delineamentos localmente D -ótimos explorou-se o método para delineamentos exatos pseudo-Bayesianos otimizando-se então a função em (12), porém com distribuição de probabilidade *a priori* apenas para os parâmetros β já que os delineamentos localmente D -ótimos se mostraram pouco sensíveis aos valores de λ . Nesse caso, atribuiu-se $\beta \sim N(\beta_0, 4\mathbf{I})$ com valores de β_0 fixados nos mesmos valores pontuais descritos na Seção 3.1. A integral em (12) foi aproximada numericamente pelo método de quadratura Gauss-Hermitiana utilizando o pacote *gaussquad* no R (NOVOMESTKY, 2013).

4 Resultados e discussões

Foram construídos delineamentos para diversas combinações de valores dos parâmetros, mas apenas alguns deles serão apresentados, já que resumem bem o comportamento quanto ao critério D . Delineamentos localmente ótimos para modelos sem interação e com interação são apresentados nas Figuras 1 e 2, respectivamente, enquanto que na Figura 3, dois delineamentos são comparados, um localmente ótimo e o outro pseudo-Bayesiano para o mesmo modelo com interação.

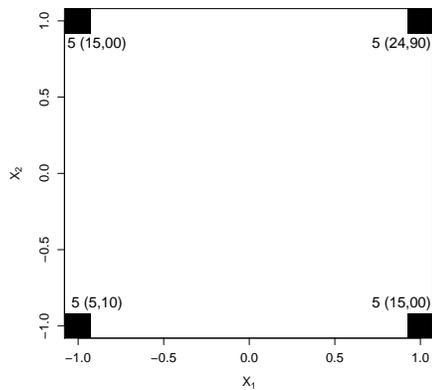
Um primeiro ponto a se destacar sobre os delineamentos é que, no caso de ocorrer desbalanceamento, os resultados mostram que mais repetições são alocadas para os tratamentos que resultam em maior dispersão de $\mathbb{E}(Y^{(\lambda)})$ em relação ao valor médio geral.

As eficiências dos delineamentos localmente ótimos são apresentadas na Tabela 1, mostrando que o valor de λ pouco influencia no delineamento, resultado este também apontado por Atkinson e Cook (1997) para o caso contínuo. Apenas em dois casos, para o modelo sem interação, os delineamentos não foram equivalentes quando variou-se apenas o valor de λ . É oportuno esclarecer que esses resultados não significam que o delineamento é o mesmo sob os dois modelos (com ou sem transformação) mas sim que o valor especificado para esse parâmetro, na fase de planejamento, não influencia muito no delineamento resultante. Se a possibilidade de transformação não for considerada, ou seja, a estimação de λ não for incluída no processo, o delineamento D -ótimo, tanto para o modelo sem, como para o modelo com interação, é o 2^2 balanceado. Sob os modelos com transformação, os delineamentos D -ótimos variam entre o 2^2 balanceado, 2^2 desbalanceado e outras configurações desbalanceadas apresentadas nas Figuras de 1 a 3, apontando a relevância de se considerar a possibilidade de transformação na fase de delineamento.

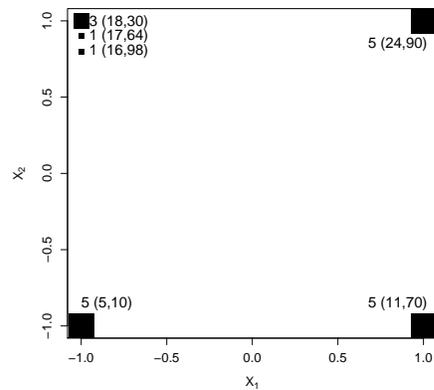
Examinando-se os resultados em maiores detalhes, observa-se que, para todos os modelos sem interação considerados, os delineamentos se assemelham ao fatorial 2^2 , balanceado no caso de simetria dos efeitos e variância pequena (Delineamento 1, Figura 1(a)) e não balanceado para variância grande (Delineamento 3, Figura 1(c)). O delineamento contínuo para três fatores apresentados por Atkinson et al. (2007) também é um fatorial 2^3 desbalanceado, construído supondo-se o modelo sem interação e $\sigma = 0,1856$. Voltando à Figura 1, alterações nos valores de β_1 e β_2 provocam algumas mudanças nos delineamentos, principalmente quando a variância é pequena, com mais tratamentos distintos aparecendo em um dos vértices da região experimental (Delineamento 2, Figura 1(b)). Para $\sigma = 3$, se $\beta_1 < \beta_2$ mas ambos positivos, o delineamento se mantém como na Figura 1(c). Porém, se o efeitos tem sinais opostos, maiores números de repetições são alocados aos tratamentos na diagonal oposta (Delineamento 4, Figura 1(d)).

Avaliando-se os valores das eficiências entre os delineamentos para modelos sem interação, verifica-se que, em geral, a perda de eficiência é pequena ($< 2\%$), embora, em alguns casos, possa chegar a $\approx 5\%$ sob erro na especificação dos valores dos parâmetros.

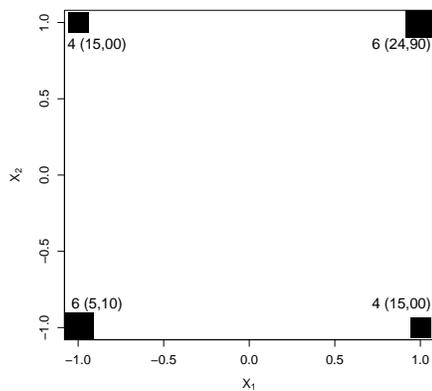
A inclusão do termo de interação no modelo tem efeito importante na



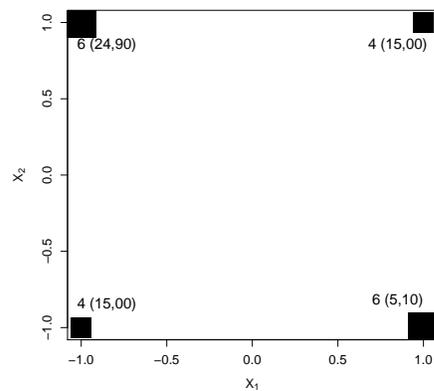
(a) Delineamento 1 ($\beta_1 = \beta_2$; $\sigma = 0,1$)



(b) Delineamento 2 ($\beta_1 < \beta_2$; $\sigma = 0,1$)

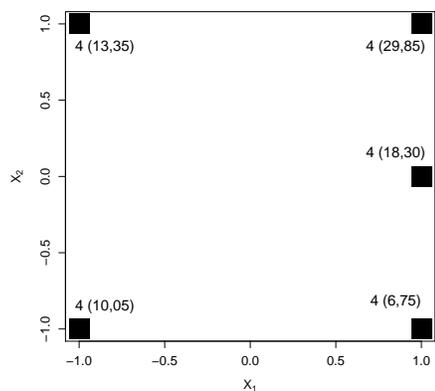


(c) Delineamento 3 ($\beta_1 = \beta_2$; $\sigma = 3,0$)

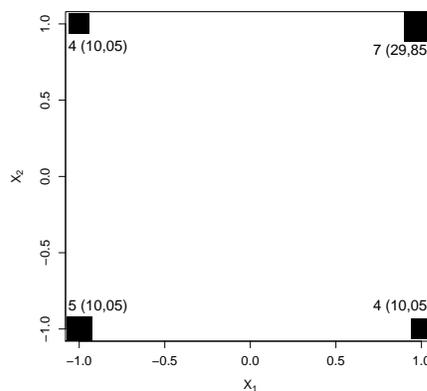


(d) Delineamento 4 ($\beta_1 = -\beta_2$; $\sigma = 3,0$)

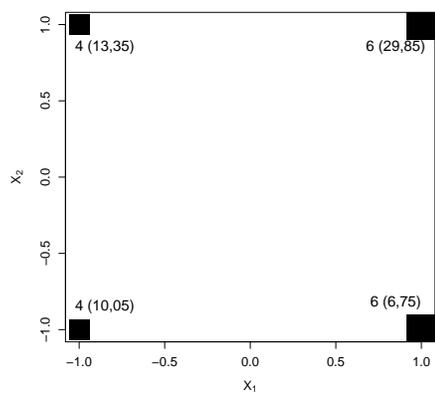
Figura 1 - Delineamentos exatos ($n = 20$), localmente D -ótimos para modelos com dois fatores, sem interação, com o número de repetições e $\mathbb{E}(Y^{(\lambda)})$ (entre parênteses) para cada tratamento.



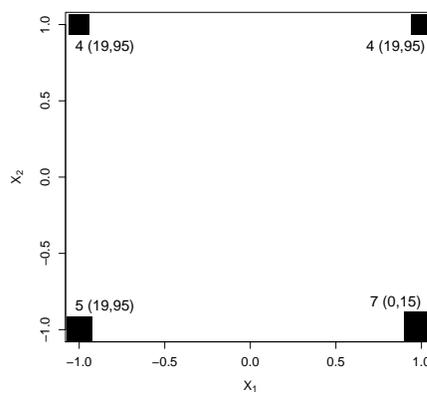
(a) Delineamento 6 ($\beta_1 < \beta_2$; $\sigma = 0,1$)



(b) Delineamento 7 ($\beta_1 = \beta_2$; $\sigma = 3,0$)



(c) Delineamento 8 ($\beta_1 < \beta_2$; $\sigma = 3,0$)



(d) Delineamento 9 ($\beta_1 = -\beta_2$; $\sigma = 3,0$)

Figura 2 - Delineamentos exatos ($n = 20$), localmente D -ótimos para modelos com dois fatores, com interação ($\beta_{12} = 4,95$), com o número de repetições e $\mathbb{E}(Y^{(\lambda)})$ (entre parênteses) para cada tratamento.

Tabela 1 - Eficiências dos delineamentos exatos ($n = 20$) localmente D -ótimos sob diversos modelos

Modelo		Delineamentos ótimos obtidos sob									
		modelos sem interação				modelos com interação					
$\beta; \sigma$	λ	1	2	3	4	5	6	7	8	9	
$\beta_1 = \beta_2 = 4,95$	0	100,00	99,38	98,39	98,38	92,55	91,46	97,84	98,38	97,83	
	$\beta_{12} = 0$	$\frac{1}{2}$	100,00	99,51	98,39	95,25	92,10	91,12	97,83	98,38	96,05
	$\sigma = 0,1$	1	100,00	99,51	98,39	98,37	92,08	91,11	97,83	98,38	97,83
$\beta_1 < \beta_2$	0	100,00	99,95	98,39	98,37	93,57	93,55	97,84	98,38	97,83	
	$\beta_{12} = 0$	$\frac{1}{2}$	99,97	100,00	98,36	98,34	92,84	92,52	97,81	98,35	97,80
	$\sigma = 0,1$	1	99,97	100,00	98,39	98,34	92,79	92,45	97,80	98,34	97,79
$\beta_1 = \beta_2 = 4,95$	0	98,87	98,30	100,00	95,14	91,66	91,02	97,37	97,50	96,43	
	$\beta_{12} = 0$	$\frac{1}{2}$	98,80	98,23	100,00	98,37	91,63	91,00	98,81	96,70	97,83
	$\sigma = 3,0$	1	98,79	98,22	100,00	95,12	91,61	90,99	98,77	96,63	96,45
$\beta_2 = -\beta_1$	0	99,01	98,48	100,00	95,71	92,22	91,23	98,38	97,84	96,20	
	$\beta_{12} = 0$	$\frac{1}{2}$	95,80	97,75	95,14	100,00	91,87	90,68	94,73	98,22	98,93
	$\sigma = 3,0$	1	98,79	97,76	95,12	100,00	91,84	90,63	94,71	98,27	99,00
$\beta_1 = \beta_2 = 4,95$	0	43,31	42,90	43,54	41,61	100,00	92,32	43,92	43,54	41,42	
	$\beta_{12} = 4,95$	$\frac{1}{2}$	43,63	43,22	43,87	41,92	100,00	92,92	44,25	43,87	41,73
	$\sigma = 0,1$	1	43,68	43,27	43,92	41,97	100,00	92,95	44,30	43,97	41,77
$\beta_1 < \beta_2$	0	40,36	40,03	40,44	38,96	94,75	100,00	40,77	40,81	39,06	
	$\beta_{12} = 4,95$	$\frac{1}{2}$	40,53	40,22	40,36	39,42	94,10	100,00	40,66	41,08	39,59
	$\sigma = 0,1$	1	40,60	40,29	40,40	39,52	94,05	100,00	40,70	41,16	39,69
$\beta_1 = \beta_2 = 4,95$	0	98,62	97,62	99,14	94,75	92,93	92,32	100,00	99,14	94,30	
	$\beta_{12} = 4,95$	$\frac{1}{2}$	98,62	97,62	99,14	94,75	93,51	92,92	100,00	99,14	94,30
	$\sigma = 3,0$	1	98,62	97,62	99,14	94,75	93,53	92,95	100,00	99,14	94,30
$\beta_1 < \beta_2$	0	98,52	97,60	97,50	98,72	93,31	94,15	99,51	100,00	95,35	
	$\beta_{12} = 4,95$	$\frac{1}{2}$	98,67	97,73	97,52	98,24	93,76	94,57	98,98	100,00	96,37
	$\sigma = 3,0$	1	98,66	97,71	97,49	98,17	93,73	94,52	98,89	100,00	96,43
$\beta_2 = -\beta_1$	0	98,62	97,68	94,75	99,14	91,87	92,32	94,30	99,14	100,00	
	$\beta_{12} = 4,95$	$\frac{1}{2}$	98,62	97,68	94,75	99,14	90,49	90,37	94,30	99,14	100,00
	$\sigma = 3,0$	1	98,62	97,68	94,75	99,14	90,20	89,95	94,30	99,14	100,00
Mínimo		40,36	40,03	40,36	38,96	90,20	89,95	40,66	40,81	39,06	
Média		87,54	87,01	86,75	86,08	93,31	92,77	86,57	87,18	85,82	
Máximo		99,97	99,95	99,14	99,14	94,75	94,57	98,98	99,62	99,90	

*Máximo excluindo a situação na qual o delineamento em questão é o ótimo.

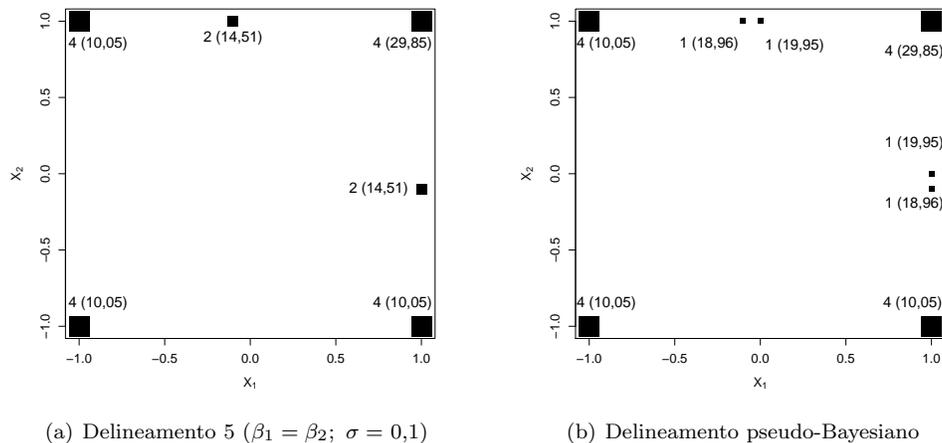


Figura 3 - Delineamentos exatos ($n = 20$), localmente e pseudo-Bayesiano D -ótimos para modelos com dois fatores, com interação, com o número de repetições e $\mathbb{E}(Y^{(\lambda)})$ (entre parênteses) para cada tratamento.

construção do delineamento D -ótimo, principalmente quando σ é pequeno. O Delineamento 5 (Figura 3(a)), construído sob o modelo com todos os efeitos com o mesmo valor, apresenta pontos em dois lados do quadrado, além dos vértices, sendo as posições simétricas em relação à diagonal crescente da região experimental. Esse comportamento também foi verificado em Atkinson e Cook (1997) nos delineamentos contínuos. As posições dos dois pontos internos podem variar levemente de acordo com o valor de λ (delineamentos não apresentados), mas não ocorre grande perda de eficiência em relação a este apresentado. O número de repetições desses tratamentos é a metade dos obtidos para os tratamentos nos vértices. Quando os valores de β_1 e β_2 são diferentes, mas com mesmo sinal, e $\sigma = 0,1$, apenas o ponto da lateral do fator com maior efeito se mantém no delineamento (Delineamento 6, Figura 2(a)). A perda de eficiência entre esses dois delineamentos é de aproximadamente de 6% a 8%, sendo preferível o uso do delineamento com mais pontos (Delineamento 5). Estudos sobre outros valores para o efeito de interação, resultam em delineamentos intermediários (não apresentados) entre o Delineamento 5 e o Delineamento 1, com as posições dos pontos internos migrando para os vértices, conforme o valor da interação diminui.

Analisando as eficiências para os delineamentos apenas sob modelos com interação, nota-se perdas bem maiores, principalmente quando sob erro de especificação de σ . Nota-se também o desempenho superior dos delineamentos que apresentam os tratamentos internos (5 e 6).

Comparações entre todos os delineamentos também levam ao destaque dos

delineamentos 5 e 6 como melhores em termos de desempenho no geral. A menor eficiência do delineamento 5 é 90,20% e do delineamento 6 é 89,95%. Outra vantagem do Delineamento 5 é a possibilidade de verificação de falta de ajuste do modelo com efeitos lineares e interação, já que ele fornece 5 graus de liberdade para o modelo. Em virtude deste fato, é recomendado considerar o modelo com uma possível interação, e com σ baixo, na fase de delineamento como precaução.

Note que ao se utilizar o delineamento balanceado (Delineamento 1) quando existe interação, perde-se mais eficiência (o delineamento chega a ser apenas $\approx 40\%$ eficiente) quando comparado ao se utilizar o delineamento com pontos internos (Delineamento 5) sob os modelos sem interação, cujas eficiências mínimas são altas, $\approx 91 - 92\%$.

A incorporação de densidades *a priori* para o vetor de parâmetros dos efeitos mostrou que os delineamentos pseudo-Bayesianos são muito similares, em alguns casos, idênticos aos localmente ótimos já discutidos. Para ilustração, na Figura 3(b) apresenta-se o delineamento ótimo para um dos casos apenas, o qual mostra a inclusão de dois tratamentos distintos em dois dos lados do quadrado, ao invés, de um tratamento com duas repetições na versão local. Embora com menos tratamentos, a eficiência do Delineamento 5, comparado ao pseudo-Bayesiano é de 99,99%. Em geral, as eficiências dos delineamentos localmente *D*-ótimos avaliados sob o critério pseudo-Bayesiano não foram muito diferentes daquelas apresentadas na Tabela 1 indicando que o custo computacional envolvido na obtenção de delineamentos pseudo-Bayesianos pode não valer a pena do ponto de vista prático.

Para finalizar a discussão, vale destacar que o uso de transformações na análise de dados tem recebido críticas em favor de ajustes de modelos lineares generalizados e/ou que incorporam a modelagem da função de variâncias (ver, por exemplo, Hattab, 2014; Feng et al., 2014 e Agresti, 2015). Nosso argumento é que se o pesquisador dispõe de informações sobre qual é o modelo razoavelmente apropriado aos dados que serão coletados, este deve ser utilizado para delinear o experimento. No entanto, delinear experimentos para modelos lineares generalizados, assunto que vem recebendo bastante atenção (ver, por exemplo, Woods et al. (2006); Dror e Steinberg (2006); Stufken e Yand, (2012); Dette et al. (2013), Atkinson e Woods (2015) e Wong et al. (2019), entre outros), além de mais custoso computacionalmente, exige informação sobre os diversos componentes do modelo (componente aleatório, função de ligação, forma do preditor linear e valores dos parâmetros).

A maioria dos estudos considera respostas binárias com função de ligação logística e, frequentemente, apenas um fator. Para variáveis resposta contínuas, a especificação da função de ligação pode não ser óbvia. Para o modelo Gama, por exemplo, em geral, não se sabe a função de ligação correta, sendo que a própria família Box-Cox pode ser utilizada como ligação (ATKINSON e WOODS, 2015). O delineamento via o modelo de transformação inclui apenas um parâmetro extra (λ), o que pode representar uma simplificação importante no processo de construção do delineamento. Atkinson e Cook (1995) investigaram o problema sob o ponto de vista de modelagem da função de variância em modelos para variável resposta contínua

e, subsequentemente no artigo de 1997, mostraram que a matriz de informação em (7) é equivalente àquela originada de um modelo particular de heterogeneidade de variâncias.

Gaviria e Ríos (2014) também construíram delineamentos ótimos contínuos usando a abordagem de transformação e função de variância e encontraram delineamentos equivalentes, embora tenham explorado apenas o caso de uma única variável explanatória. O uso de transformações também pode ter um apelo prático na modelagem em situações mais complexas, como no caso de modelos mecanísticos, com transformações em ambos os lados (ATKINSON, 2003), e modelos de efeitos mistos. Mesmo com metodologia desenvolvida, ainda podemos nos deparar com dificuldades computacionais para o ajuste e diagnóstico de tais modelos. O uso cauteloso de alguma transformação pode simplificar, permitindo a análise de dados em situações de falhas no ajuste de modelos mais sofisticados.

Conclusões

Neste trabalho foram construídos delineamentos D -ótimos exatos para dois fatores na situação de suspeita de heterocedasticidade, assumindo que o problema possa ser razoavelmente aproximado via transformação da variável resposta. Devido a presença de não linearidade, para a construção dos delineamentos foi necessário a especificação de valores a priori aos parâmetros. A robustez dos delineamentos aos valores especificados foi estudada via uma medida de eficiência.

Os resultados indicam pouca sensibilidade do critério D às alterações dos valores de λ , embora em alguns casos, o delineamento tenha sofrido modificações, principalmente em relação ao parâmetro σ . É clara a tendência do delineamento ótimo colocar mais repetições nos tratamentos cujas médias são as mais discrepantes, priorizando as partes da região experimental que espera-se produzirem respostas com variância muito pequena e aquelas com variância muito grande.

Outra característica que pode ser observada é que os delineamentos D -ótimos construídos sob modelos com o termo de interação e σ pequeno são mais robustos já que perdem menos eficiência quando avaliados em diversos cenários. A robustez é obtida pela inclusão de tratamentos extras quando comparado ao fatorial usual (2^2).

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, via o Programa de Pós-graduação em Biometria, Unesp, Botucatu, SP. O primeiro autor agradece ao Campus de Parauapebas da Universidade Federal Rural da Amazônia, pelo apoio e concessão de afastamento integral para a realização do curso de doutorado, do qual este trabalho é um dos

resultados. Ambos os autores agradecem os revisores e editores pelas sugestões e comentários que possibilitaram uma melhor apresentação do artigo.

REIS, C. P.; TRINCA, L. A. Exact D -optimal design under response variable transformation. *Rev. Bras. Biom.*, Lavras, v.38, n.2, p.226-242, 2020.

- **ABSTRACT:** *In experimental sciences, optimal designs, which focus on the efficiency of specific properties, are alternatives of fundamental importance when the practical problem involves some complexity not directly contemplated in the classical design formulation. Most of the methods assumes homogeneity of variances, which is not always verified in the data set. The aim of this work is to develop a methodology to construct exact optimal or efficient designs in situations of nonhomogeneous variances. It is assumed that application of a transformation from the Box-Cox family accomplishes both linearity and homocedasticity. Results for several examples assuming first and second order models are presented and discussed.*
- **KEYWORDS:** *Heteroscedasticity; Box-Cox transformation; factorial; efficient designs.*

Referências

AGRESTI, A. *Foundations of linear and generalized linear models*. Hoboken: John Wiley & Sons, 2015.

ATKINSON, A. C. Horwitz's rule, transforming both sides and the design of experiments for mechanistic models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v.52, p.261-278, 2003.

ATKINSON, A. C.; COOK, R. D. Designing for a response transformation parameter. *Journal of the Royal Statistical Society, Series B*, v.59, p.111-124, 1997.

ATKINSON, A. C.; DONEV, A. N.; TOBIAS, R. D. *Optimum experimental designs, with SAS*. Oxford: Oxford University Press, 2007.

ATKINSON, A. C.; WOODS, D. C. Designs for generalized linear models. *arXiv e-prints*, p.arXiv:1510.05253, 2015.

BOX, G. E. P.; COX, D. R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, v.26, p.211-246, 1964.

CASELLA, G.; BERGER, R. *Statistical inference*. Boston: Cengage Learning, 2001.

CHALONER, K.; LARNTZ, K. Optimal Bayesian design applied to logistic-regression experiments. *Journal of Statistical Planning and Inference*, v.21, p.191-208, 1989.

COOK, R. D.; WEISEBERG, S. *Residuals and influence in regression*. London: Chapman and Hall, 1982. 62p.

DAVISON, A. C. *Statistical models*. Cambridge: Cambridge University Press, 2003.

DROR, H. A.; D. M. STEINBERG. Robust experimental design for multivariate generalized linear models. *Technometrics*, v. 48, p.520-529, 2006.

DETTE, H.; HOYDEN, L.; KUHNT, S.; SCHORNING, K. Optimal designs for multi-response generalized linear models with applications in thermal spraying. *arXiv e-prints*, p.arXiv:1312.4472, 2013. Disponível em <http://arxiv.org/pdf/1312.4472.pdf>

FEDOROV, V. V. *Theory of optimal experiments*. New York: Academic Press, 1972.

FENG, C.; WANG, H.; LU, N.; CHEN, T.; HE, H. LU, Y.; TU, X. M. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, v.26, p.105-109, 2014.

FERREIRA, I. E. P.; FERREIRA, C. P.; TRINCA, L. A. Delineamentos experimentais eficientes para estudos de cinética química. *Química Nova*, v.37, p.589-596, 2014.

GAVIRIA, J. A.; RÍOS, V. I. L. Locally D-optimal designs with heteroscedasticity: a comparison between two methodologies. *Revista Colombiana de Estadística*, v.37, n.1, p.95-110, 2014.

GILMOUR, S. G.; TRINCA, L. A. Bayesian L-optimal exact design of experiments for biological kinetic models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v.61, p.237-251, 2012.

GOOS, P.; MYLONA, K. Quadrature methods for Bayesian optimal design of experiments with non-normal prior distributions. *Journal of Computational and Graphical Statistics*, v.27, p.179-194, 2018.

GOTWALT, C. M., JONES, B. A.; STEINBERG, D. M. Fast computation of designs robust to parameter uncertainty for nonlinear settings. *Technometrics*, v.51, p.88-95, 2009.

HATTAB, M. W. On the use of data transformation in response surface methodology. *Qual Reliab Engng Int.*, v.34, p.1185-1194, 2018.

KIEFER, J. Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, v.21, p.272-319, 1959.

MEYER, R. K.; NACHTSHEIM, C. J. The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, v.37, p.60-69, 1995.

NOVOMESTKY, F. *gaussquad: Collection of functions for Gaussian quadrature*, 2013. R package version 1.0-2. *<https://CRAN.R-project.org/package=gaussquad>.

OVERSTALL, A. M.; WOODS, D. C. Bayesian design of experiments using approximate coordinate exchange. *Technometrics*, v.59, p.458-470, 2017.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. *<https://www.R-project.org/>, 2019.

SAKIA, R. M. The Box-Cox transformation technique: A review. *Journal of the Royal Statistical Society, Series D*, v.41, p.169-178, 1992.

STUFKEN, J.; M. YANG. Optimal designs for generalized linear models. In K. Hinkelmann (Ed.), *Design and Analysis of Experiments, Special Designs and Applications*, Chapter 4, New York: Wiley, p.137-135, 2012.

WOODS, D. C.; LEWIS, S. M.; ECCLESTON, J. A.; RUSSEL, G. Designs for generalized linear models with several Variables and model uncertainty. *Technometrics*, v.48, p.284-292, 2006.

WONG, W. K.; YIN, Y.; ZHOU, J. Optimal designs for multi-response nonlinear regression models with several factors via semidefinite programming. *Journal of Computational and Graphical Statistics*, v.28, p.61-73, 2019.

Recebido em 04.07.2019.

Aprovado após revisão em 22.10.2019.