Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

ESTIMAÇÃO BAYESIANA NO MODELO MULTINOMIAL COM ERROS DE CLASSIFICAÇÃO E CLASSIFICAÇÕES REPETIDAS

Magda Carvalho Pires[1], Letícia Silva Nunes[2]

Resumo: *A tarefa de classificar indivíduos segundo alguma característica está presente na maioria das áreas de conhecimento. Entretanto, o processo de classificação pode estar sujeito a erros, o que quer dizer que, em um caso dicotômico, um sucesso pode ser erroneamente classificado como fracasso ou vice-versa. Ignorando esses erros de classificação, são produzidas estimativas viciadas das quantidades de interesse [Bross, (1953)]. Diante do problema, muitos métodos foram sugeridos. Pires (2006) e Quinino et al. (2010), por exemplo, propõem utilizar repetidas classificações dos elementos amostrais e incorporar no modelo a classificação mais frequente. Em uma situação mais complexa, consideremos que a classificação dos elementos amostrais em mais de duas categorias pode estar sujeita a erros, ou seja, um elemento amostral pode estar alocado a uma categoria que não corresponde ao seu estado verdadeiro. Nesse contexto, apresentamos uma proposta para estender a metodologia de classificações repetidas à análise de dados multinomiais utilizando dados aumentados e uma abordagem bayesiana. Estudos de Simulação Monte Carlo demonstraram bom desempenho do modelo proposto em relação ao modelo que não considera apenas uma classificação, no sentido de produzir estimativas a posteriori menos viciadas e com menor variabilidade.*

Palavras-chave: *Distribuição Multinomial, Erros de Classificação, Classificações Repetidas*.

Abstract: *When sampling units are classified into three or more categories and this classification may be erroneous, we propose to perform repeated and independent ratings. Results using augmented data and Bayesian approach were satisfactory (posteriors estimatives with smaller relative bias and variability) compared with the model that considers only one rating.*
Keywords: *Multinomial Distribution, Misclassification, Repeated Measures.*

1 Introdução

A tarefa de classificar indivíduos segundo alguma característica está presente na maioria das áreas de conhecimento. Uma situação mais simples é aquela em que os elementos amostrais são classificados em apenas duas categorias, como doente / não doente, defeituosa / não defeituosa, ou genericamente, sucesso / fracasso. Alguns processos de classificação podem, entretanto, estar sujeitos a erros, o que quer dizer que um sucesso pode ser erroneamente classificado como fracasso ou vice-versa. Bross (1954) foi o pioneiro em

[1] Departamento de Estatística/Universidade Federal de Minas Gerais. E-mail: magda@est.ufmg.br
[2] Graduanda do Curso de Estatística da Universidade Federal de Minas Gerais.
Agradecimentos à FAPEMIG e à PRPq/UFMG pelo apoio financeiro.

analisar o impacto de tais erros de classificação no processo inferencial, demonstrando que, ao ignorá-los, são produzidas estimativas viciadas da proporção de interesse. Diante do problema, alguns métodos clássicos foram sugeridos e uma revisão pode ser encontrada em Johnson et al. (1991). Em uma ótica bayesiana, Gaba e Winkler (1992) consideraram uma abordagem que requer a utilização de uma distribuição *a priori* informativa. No intuito de minimizar o impacto dos erros de classificação nas estimativas, Pires (2006) e Quinino et al. (2010) propõem utilizar classificações repetidas e independentes dos elementos amostrais e incorporá-las no modelo que considera os erros de classificações.

Em uma situação mais complexa a classificação dos elementos amostrais pode ser realizada em mais de duas categorias. Podemos, por exemplo, classificar as pessoas de acordo com a raça (branca, negra, parda, amarela) ou a gravidade de uma doença (leve, moderada, grave), peças produzidas por uma indústria de acordo com sua qualidade (perfeitas, reparáveis, refugo) etc. Nesses casos, estamos interessados em estimar a probabilidade de um indivíduo pertencer a cada uma das categorias. Novamente, alguns processos de classificação podem estar sujeitos a erros, ou seja, um elemento amostral pode ser alocado a uma categoria que não corresponde ao seu estado verdadeiro. Alguns autores já analisaram esse problema em abordagem clássica (Cheng [1989]) e bayesiana (Perez et al. [2007]).

Utilizando uma abordagem bayesiana, o objetivo desse trabalho é estender o uso de classificações repetidas à análise de dados multinomiais sujeitos a erros de classificação, visando reduzir o vício das estimativas das probabilidades de um indivíduo pertencer a cada uma das categorias.

2 Materiais e métodos

Suponha que os indivíduos de uma população sejam divididos em m categorias disjuntas denotadas por A_1, A_2,..., A_m. Uma amostra de tamanho n dos indivíduos é retirada dessa população, mas suponha que os indivíduos podem estar classificados erroneamente nas categorias de acordo com a matriz de probabilidade de erro $\Lambda(mxm)$, em que cada elemento λ_{ij} denota a probabilidade de que um indivíduo de A_i seja classificado em A_j. Considere θ_1, θ_2,...,θ_m as probabilidades de pertencer às categorias A_1, A_2,..., A_m, respectivamente. Seja X, a variável aleatória que representa a categoria verdadeira do indivíduo e $\mathbf{Y}$ uma matriz com a frequência das classificações dos indivíduos. Assim, a função de verossimilhança é (Perez et. al.[2007]):

$$f(y \mid \theta, \Lambda) = \frac{n!}{y_1! \ldots y_m!} \prod_{j=1}^{m} \left(\sum_{i=1}^{m} \theta_i \lambda_{ij} \right)^{y_j}$$

Suponha agora que cada elemento da amostra é avaliado k vezes. Seja y_{pq} ($p = 1, 2, ..., n;$ $q = 1,2,...,k$) o número de classificações obtidas pelo p-ésimo elemento na q-ésima categoria. Assim $y_{23}=2$ significa que o segundo elemento foi classificado duas vezes na terceira categoria. Além disso, $\sum_{j=1}^{m} y_{pj} = k$ para todo p.

Como as classificações são independentes, temos então que, condicionado ao seu estado verdadeiro, o vetor aleatório $\mathbf{Y_p}$ tem distribuição Multinomial, ou seja,

$$P(\mathbf{Y_p} = \mathbf{y_p} \mid X_p = A_i) \propto \prod_{j=1}^{m} (\lambda_{ij})^{y_{pj}} .$$

Assim,

$$P(\mathbf{Y}_p = \mathbf{y}_p) = \sum_{j=1}^{m} P(\mathbf{Y}_p = \mathbf{y}_p \mid X_p = A_j) P(X_p = A_j)$$

$$= \sum_{i=1}^{m} \theta_i \left[\frac{k!}{y_{p1}! \ldots y_{pm}!} \prod_{j=1}^{m} \left(\lambda_{ij}\right)^{y_{pj}} \right].$$

Portanto, a função de verossimilhança pode ser escrita como:

$$L(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \propto \prod_{p=1}^{n} \left\{ \sum_{i=1}^{m} \theta_i \prod_{j=1}^{m} (\lambda_{ij})^{y_{pj}} \right\}. \tag{1}$$

Como Perez *et al.* (2007), utilizamos a distribuição *a priori* Dirichlet com hiperparâmetros ($\gamma_1, \gamma_2,...,\gamma_m$) e ($\alpha_{i1}, \alpha_{i2},...,\alpha_{im}$) para θ e λ_i. Temos, então, as seguintes expressões para as distribuições *a priori*:

$$\pi(\boldsymbol{\theta}) \propto \prod_{i=1}^{m} \theta_i^{\gamma_{i-1}} \tag{2}$$

$$\pi(\boldsymbol{\lambda}_i) \propto \prod_{j=1}^{m} (\lambda_{ij})^{\alpha_{ij-1}}. \tag{3}$$

Assumindo a independência entre os λ_i, podemos reescrever (3) da seguinte forma:

$$\pi(\boldsymbol{\lambda}) \propto \prod_{i=1}^{m} \prod_{j=1}^{m} (\lambda_{ij})^{\alpha_{ij-1}}. \tag{4}$$

A partir de (1), (2) e (4), a distribuição *a posteriori* conjunta de $\boldsymbol{\theta}$ e $\boldsymbol{\lambda}$ é:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{y}) \propto \left\{ \prod_{p=1}^{n} \left(\sum_{i=1}^{m} \theta_i \prod_{j=1}^{m} (\lambda_{ij})^{y_{pj}} \right) \right\} \times \prod_{i=1}^{m} \theta_i^{\gamma_{i-1}} \times \prod_{i=1}^{m} \prod_{j=1}^{m} (\lambda_{ij})^{\alpha_{ij-1}} \tag{5}$$

Percebemos que expressão da distribuição *a posteriori* (5) é muito complicada, sendo difícil aplicar métodos numéricos de inferência. Para resolver esse problema, utilizaremos a abordagem de dados aumentados.

Seja d_{pi} uma variável indicadora de que o indivíduo p com vetor de classificações $\mathbf{Y_p}$ pertence, na realidade, à categoria A_i. Então,

$$P(d_{pi}=1|\mathbf{Y_p}=\mathbf{y_p})=\frac{\theta_i\prod_{j=1}^{m}(\lambda_{ij})^{y_{pj}}}{\sum_{j=1}^{m}\theta_j\prod_{q=1}^{m}(\lambda_{jq})^{y_{pq}}}.$$

A função de verossimilhança (1) pode então ser reescrita como:

$$L(\mathbf{d}|\boldsymbol{\theta},\boldsymbol{\lambda})\propto\prod_{i=1}^{m}\theta_i^{\sum_{p=1}^{n}d_{pi}}\times\prod_{i=1}^{m}\prod_{j=1}^{m}(\lambda_{ij})^{\sum_{p=1}^{n}y_{pj}d_{pi}}. \quad (6)$$

Percebemos que a função de verossimilhança (6) pode ser fatorada como $L(\boldsymbol{\theta},\boldsymbol{\lambda}|\mathbf{d})\propto L(\boldsymbol{\theta}|\mathbf{d})\times L(\boldsymbol{\lambda}|\mathbf{d})$ e, por isso, podemos reescrever a distribuição *a posteriori* como $\pi(\boldsymbol{\theta},\boldsymbol{\lambda}|\mathbf{d})\propto L(\boldsymbol{\theta}|\mathbf{d})\times L(\boldsymbol{\lambda}|\mathbf{d})\times\pi(\boldsymbol{\theta})\times\pi(\boldsymbol{\lambda})=\pi(\boldsymbol{\theta}|\mathbf{d})\times\pi(\boldsymbol{\lambda}|\mathbf{d})$, em que:

$$\pi(\boldsymbol{\theta}|\mathbf{d})\propto\prod_{i=1}^{m}(\theta_i)^{\sum_{p=1}^{n}d_{pi}}\times\prod_{i=1}^{m}\theta_i^{\gamma_i-1}=\prod_{i=1}^{m}(\theta_i)^{\sum_{p=1}^{n}d_{pi}+\gamma_i-1}$$

e

$$\pi(\boldsymbol{\lambda}|\mathbf{d})\propto\prod_{i=1}^{m}\prod_{j=1}^{m}(\lambda_{ij})^{\sum_{p=1}^{n}y_{pj}\times d_{pi}}\times\prod_{i=1}^{m}\prod_{j=1}^{m}(\lambda_{ij})^{\alpha_{ij}-1}=\prod_{i=1}^{m}\prod_{j=1}^{m}(\lambda_{ij})^{\sum_{p=1}^{n}(y_{pj}.d_{pi})+\alpha_{ij}-1}.$$

Em um procedimento do tipo GibbsSampler, pode-se então realizar inferências através do algoritmo CDA (Chained Data AugmentationAlgorith) proposto por Tanner (1996).

Para avaliar o desempenho do modelo proposto, implementamos uma rotina no software Ox (versão 5.1) (Doornik, 2007). Considerou-se uma amostra de tamanho n=400 de uma distribuição Multinomial, em que cada elemento amostral foi avaliado k=1, 2, 3, 4, 5 vezes, pois o processo de classificação estava sujeito a erros de classificação. As Tabelas 1 e 2 apresentam, respectivamente, os valores simulados dos parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\Lambda}$, além da média e desvio padrão das distribuições *a priori* Dirichlet para $\boldsymbol{\theta}$ e $\boldsymbol{\Lambda}$ nos dois diferentes conjuntos (Caso 1 e Caso 2) de distribuições *a priori* utilizados. Distribuições mais informativas, ou seja, com menor variância, foram utilizadas no Caso 1.

Tabela 1: Valores simulados, média e desvio padrão das distribuições *a priori* para $\boldsymbol{\theta}$

Parâmetro	Valor simulado	Caso 1		Caso 2	
		Média	DP	Média	DP
θ_1	0,400	0,471	0,162	0,400	0,200
θ_2	0,300	0,353	0,155	0,300	0,187
θ_3	0,200	0,118	0,105	0,200	0,163
θ_4	0,100	0,059	0,076	0,100	0,122

Para cada caso foram realizadas 100 simulações Monte Carlo, ou seja, foram geradas 100 amostras sujeitas a erros de classificação de magnitudes apresentadas na Tabela 2.

Tabela 2: Valores simulados, média e desvio padrão das distribuições *a priori* para $\boldsymbol{\Lambda}$

Parâmetro	Valor simulado	Caso 1		Caso 2	
		Média	DP	Média	DP
λ_{11}	0,800	0,820	0,022	0,800	0,030
λ_{12}	0,060	0,137	0,039	0,060	0,035
λ_{13}	0,050	0,003	0,006	0,050	0,032
λ_{14}	0,090	0,041	0,022	0,090	0,043
λ_{21}	0,025	0,018	0,013	0,018	0,024
λ_{22}	0,900	0,893	0,079	0,686	0,125
λ_{23}	0,035	0,071	0,025	0,267	0,086
λ_{24}	0,040	0,018	0,013	0,030	0,030
λ_{31}	0,045	0,002	0,004	0,046	0,026
λ_{32}	0,020	0,016	0,013	0,018	0,017
λ_{33}	0,850	0,974	0,085	0,849	0,105
λ_{34}	0,085	0,008	0,009	0,087	0,036
λ_{41}	0,015	0,018	0,013	0,016	0,015
λ_{42}	0,010	0,088	0,028	0,092	0,034
λ_{43}	0,025	0,009	0,009	0,027	0,019
λ_{44}	0,950	0,885	0,079	0,865	0,098

3 Resultados e discussões

Procedendo-se a análise bayesiana dos dados simulados segundo metodologia descrita na seção anterior, a convergência das cadeias foi verificada monitorando-se as médias ergódicas, partindo de valores iniciais distintos. Como explica Paulino et al. (2003), quando verifica-se que essas médias convergem para os mesmos valores, procede-se uma amostragem usando uma única cadeia para realizar inferências. A Figura 1 apresenta os gráficos de monitoramento das médias ergódicas de $\boldsymbol{\theta}$ em uma das simulações Monte Carlo, onde percebemos que as cadeias convergem rapidamente para um mesmo valor que se estabiliza de forma satisfatória a partir da iteração 50.000 (período utilizado como burn-in).

Além da convergência da cadeia, é preciso avaliar a presença de autocorrelação entre as observações (Paulino, 2003). A Figura 2 apresenta a função de autocorrelação para o vetor de parâmetros $\boldsymbol{\theta}$ em uma das simulações Monte Carlo, a partir da qual concluímos que um *lag* de tamanho 50 é suficiente para que as amostras sejam independentes.

Resultados semelhantes da avaliação de convergência e autocorrelação foram obtidos para os dois casos, em que vários cenários foram analisados.

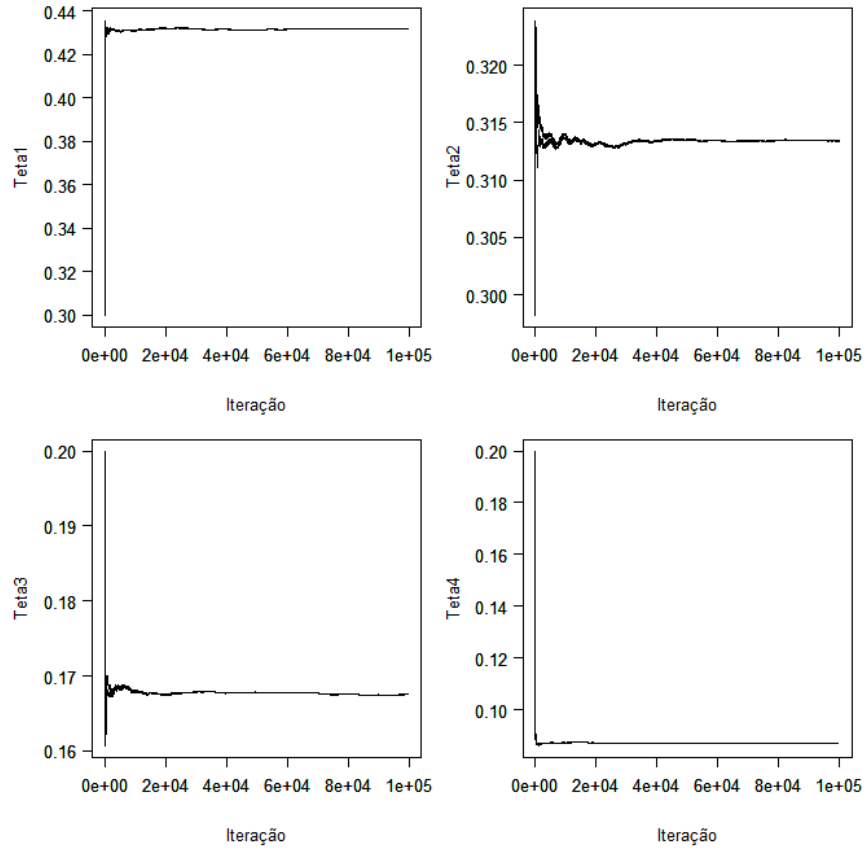

Figura 1: Médias ergódicas das duas cadeias geradas para o vetor de parâmetros **θ**

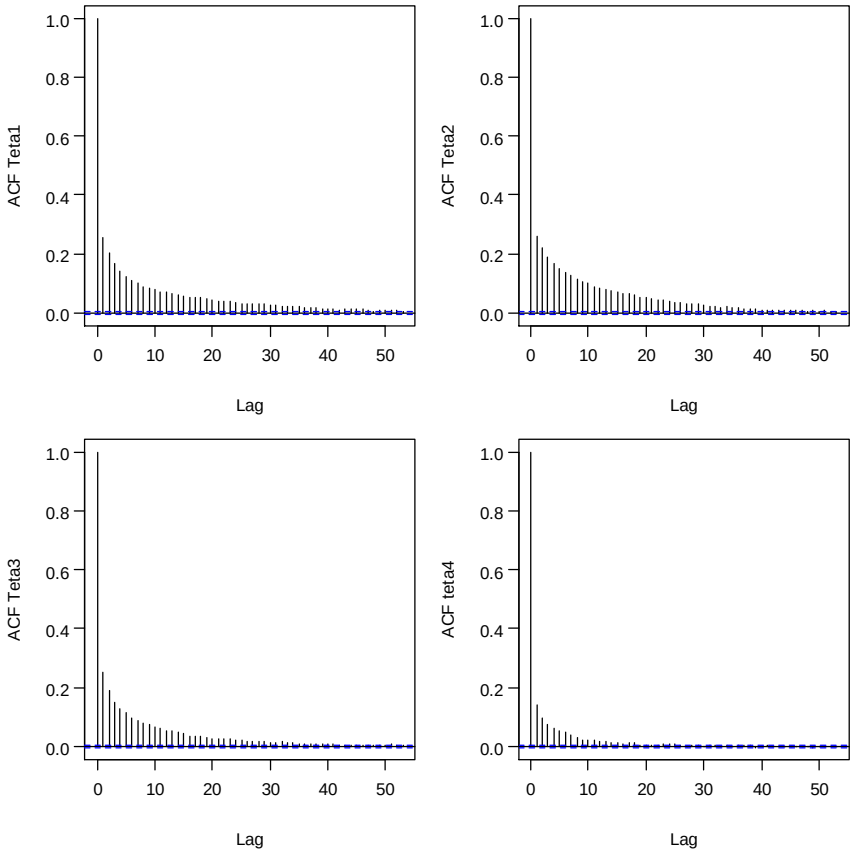

Figura 2 – Função de autocorrelação para o vetor de parâmetros **θ**

Finalmente, estimativas *a posteriori* da média, desvio padrão, percentil 5 e percentil 10 para as probabilidades de ocorrência das categorias obtidas em cada caso são apresentados nas Tabelas 3 a 6. Os valores apresentados representam a média das 100 simulações Monte Carlo.

Tabela 3: Estimativas *a posteriori* das quantidades de interesse para θ_1

k	Caso 1				Caso 2			
	Média	DP	Perc. 5	Perc.95	Média	DP	Perc. 5	Perc.95
1	0,401	0,021	0,370	0,437	0,407	0,023	0,368	0,443
2	0,415	0,026	0,372	0,457	0,399	0,025	0,356	0,439
3	0,409	0,028	0,364	0,454	0,410	0,026	0,365	0,448
4	0,404	0,026	0,365	0,452	0,404	0,027	0,363	0,445
5	0,396	0,024	0,353	0,432	0,399	0,024	0,365	0,446

Tabela 4: Estimativas *a posteriori* das quantidades de interesse para θ_2

k.	Caso 1				Caso 2			
	Média	DP	Perc. 5	Perc.95	Média	DP	Perc. 5	Perc.95
1	0,314	0,021	0,282	0,345	0,347	0,023	0,308	0,390
2	0,273	0,021	0,234	0,304	0,295	0,021	0,257	0,331
3	0,303	0,025	0,264	0,345	0,299	0,023	0,263	0,342
4	0,300	0,022	0,265	0,336	0,298	0,022	0,262	0,329
5	0,304	0,023	0,269	0,338	0,303	0,022	0,266	0,331

Tabela 5: Estimativas *a posteriori* das quantidades de interesse para θ_3

k	Caso 1				Caso 2			
	Média	DP	Perc. 5	Perc.95	Média	DP	Perc. 5	Perc.95
1	0,169	0,018	0,141	0,200	0,140	0,014	0,119	0,164
2	0,173	0,018	0,143	0,205	0,190	0,019	0,161	0,220
3	0,195	0,021	0,161	0,227	0,200	0,024	0,166	0,237
4	0,196	0,020	0,168	0,230	0,201	0,022	0,165	0,232
5	0,201	0,020	0,172	0,240	0,200	0,020	0,172	0,232

Tabela 6: Estimativas *a posteriori* das quantidades de interesse para θ_4

Num.Clas.	Caso 1				Caso 2			
	Média	DP	Perc. 5	Perc.95	Média	DP	Perc. 5	Perc.95
1	0,116	0,015	0,093	0,143	0,106	0,013	0,085	0,126
2	0,139	0,020	0,103	0,175	0,116	0,016	0,090	0,141
3	0,093	0,015	0,070	0,117	0,091	0,014	0,066	0,111
4	0,100	0,015	0,075	0,125	0,097	0,016	0,070	0,123
5	0,099	0,014	0,073	0,119	0,098	0,013	0,077	0,119

Utilizando classificações repetidas, as estimativas encontradas usando a média *a posteriori* de θ_i tiveram um desempenho satisfatório, no sentido de terem, em média, um vício relativo menor em comparação ao cenário em que se utiliza apenas uma classificação dos

indivíduos. O desempenho é melhor na medida em que o número de classificações aumenta, exceto para θ_1. Tanto no caso 1 como no caso 2 as melhores estimativas foram obtidas quando realizamos cinco classificações repetidas, exceto na estimativa de θ_2 no caso 2, em que uma estimativa ligeiramente melhor foi obtida quando usamos três classificações repetidas. Percebe-se também que a disponibilidade menor de informação *a priori*, representada pela maior variância das distribuições do Caso 2, parece não afetar significativamente as estimativas *a posteriori*, já que resultados semelhantes ao Caso 1 foram obtidos. Além disso, na maioria dos casos percebemos que a variância das estimativas diminui quando aumentamos o número de classificações.

Gráficos das médias *a posteriori* obtidas nas 100 simulações podem ser visualizados no Apêndice A. Gráficos da distribuição *a posteriori* de alguns são apresentados no Apêndice C.

4 Conclusões

Nesse trabalho propomos um modelo para incorporar o conceito de classificação repetida na estimação dos parâmetros do modelo Multinomial quando os dados estão sujeitos a erros de classificação.

O modelo se mostrou satisfatório no que se refere ao vício relativo, independente da distribuição *a priori* utilizada, demonstrando que utilizar classificações repetidas produz estimativas melhores que aquelas obtidas quando os elementos amostrais são avaliados apenas uma vez (Perez et al. 2007). Além disso, estimativas satisfatórias (pouco viciadas e com pequena variabilidade) são obtidas com apenas três classificações repetidas, corroborando com a viabilidade do método para obter estimativas menos viciadas dos parâmetros de interesse.

Este texto apresenta apenas os resultados obtidos quando simulamos amostras em que a variável resposta está sujeita a uma única matriz de erros de classificação. Deve-se mencionar então que resultados semelhantes foram obtidos utilizando outras matrizes e diferentes parâmetros para a distribuição Multinomial.

Pretende-se ainda avaliar o comportamento do modelo utilizando diferentes tamanhos amostrais e números distintos de classificações repetidas para cada elemento amostral, além de aplicar a dados reais e estender o modelo ao caso multivariado, em que a classificação dos indivíduos é função de outras variáveis (modelos de regressão para variáveis categóricas).

Referências

[1] BROSS, I. Misclassification in 2×2 tables. **Biometrics**, v. 10, p. 478–486, 1954.

[2] CHENG, K. F., HSUEH, H. M. Estimation of a logistic regression model with mismeasured observations. **Statistica Sinica**, 13, 111-127, 1989.

[3] DOORNIK, J.A. **An Object-Oriented Matrix Language Ox 5**, London: Timberlake Consultants Press, 2007.

[4] GABA, A.; WINKLER, R. L. Implications of errors in survey data: A Bayesian model. **Management Science**, v. 38, n. 7, p. 913–925, 1992.

[5] JOHNSON, N. L.; KOTZ, S.; WU, X. **Inspection Errors for Attributes in QualityControl.** London: Chapman & Hall, 1991.

[6] PAULINO, C. D..; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003, 446 p.

[7] PÉREZ, C. J.; GIRÓN, F. J.; MARTÍN, J.; RUIZ, M.;Rojano, C. Misclassified multinomial data: a Bayesian approach. **Rev. R. Acad. Cien**. Serie A. Mat, p 71-80, 2007.

[8] PIRES, M. C. A**nálise Bayesiana Empírica de Dados Dicotômicos com Erros e Classificações Repetidas**. Dissertação (Mestrado em Estatística), Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

[9] QUININO, R. C.; PIRES, M. C.; SUYAMA, E.; Ho, L. L. Estimation of the Conformance Fraction in a Presence of Misclassification Errors: a Bayesian Analysis in an Absence of Expert´s Knowledge. **Brazilian Journal of Operations and Production Management**, v. 7, p. 181-193, 2010.

[10] TANNER, M. A. **Tools for Statistical Inference**. 3ª edição. New York: Springer. 1996.

Apêndice A

As médias *a posteriori* das probabilidades de ocorrência de cada categoria nas 100 simulações Monte Carlo de cada cenário podem ser observadas nas Figuras 3 a 6.

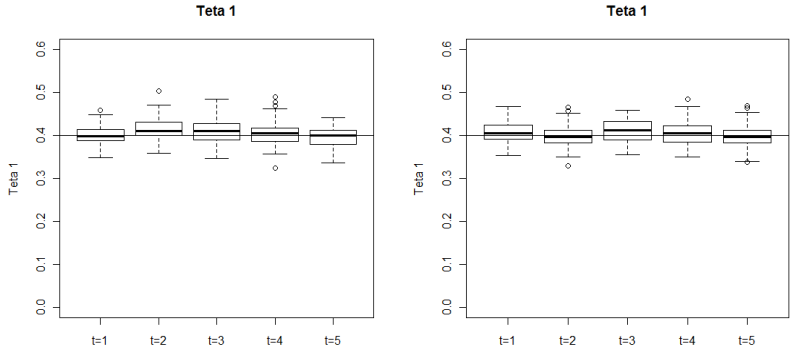

Figura 3: Médias *a posteriori* de θ_1 nos casos 1 e 2, respectivamente

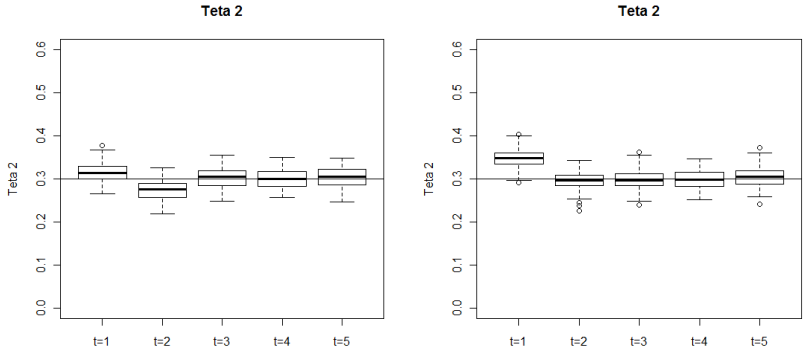

Figura 4: Médias *a posteriori* de θ_2 nos casos 1 e 2 respectivamente

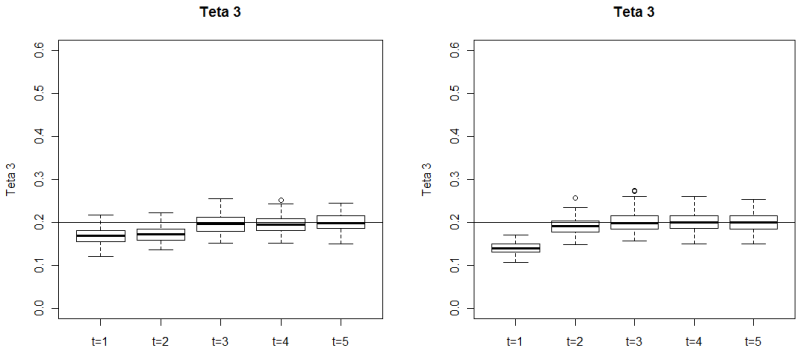

Figura 5: Médias *a posteriori* de θ_3 nos casos 1 e 2 respectivamente

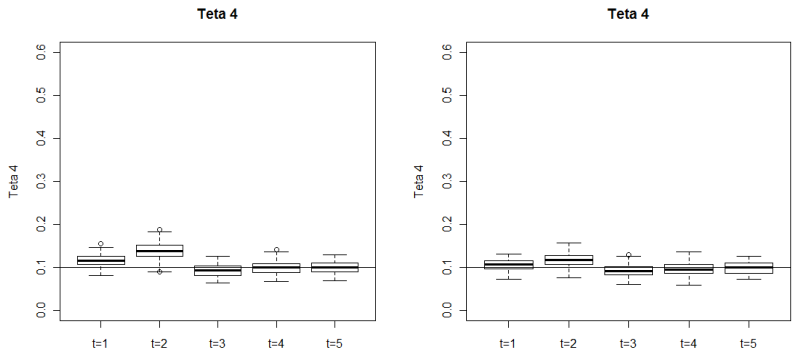

Figura 6: Médias *a posteriori* de θ_4 nos casos 1 e 2, respectivamente.

Apêndice B

A Figura 7 apresenta a distribuição *a posteriori* para as probabilidades de cada categoria obtida em um dos cenários simulados.

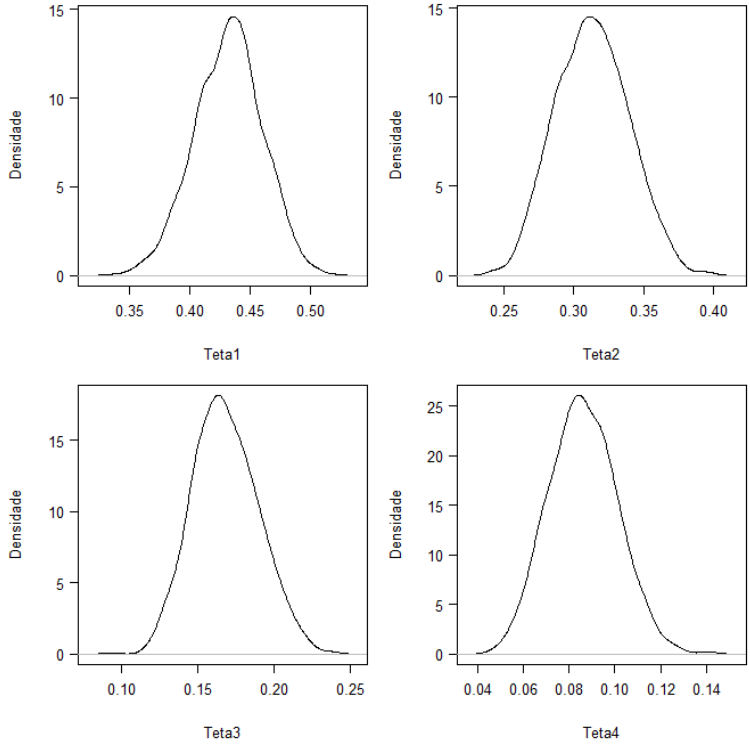

Figura 7: Distribuição *a posteriori* para as probabilidades de cada categoria