Probabilidades e Estatística

Editor Convidado: Maria Antónia Turkman

Abordagem bayesiana não paramétrica do problema de triagem

Sandra Ramos
Instituto Politécnico do Porto & CEAUL
Rua Dr. António Bernardino de Almeida, 431
4200-072 Porto, Portugal
e-mail: sfr@isep.ipp.pt

Antónia Amaral Turkman, *Marília Antunes*
Faculdade de Ciências da Universidade de Lisboa & CEAUL
Bloco C6 - Piso 4, Campo Grande
1749-016 Lisboa, Portugal
e-mail: antonia.turkman@fc.ul.pt
marilia.antunes@fc.ul.pt

Resumo: O procedimento de triagem envolve a construção de uma região de especificação $C_{\mathbf{X}}$, num espaço $d-$dimensional, de modo a que um indivíduo futuro com um vetor de características em $C_{\mathbf{X}}$ tenha maior probabilidade de ser identificado como um *sucesso* (a resposta Y pertence a uma região conhecida C_Y). Na abordagem preditiva bayesiana a obtenção da região $C_{\mathbf{X}}$ é baseada num critério ótimo assente na maximização de $P(Y \in C_Y | \mathbf{X} \in C_{\mathbf{X}}; D)$, restringida à classe das regiões $C_{\mathbf{X}}$ com probabilidade preditiva de triagem α. Habitualmente, a construção da região $C_{\mathbf{X}}$ baseia-se em modelos paramétricos para $(Y, \mathbf{X})$, mas que nem sempre descrevem adequadamente o processo que gera os dados. Neste trabalho, propõe-se uma abordagem não paramétrica bayesiana que relaxa o pressuposto paramétrico.

Abstract: The screening procedure consists in building a specification region $C_{\mathbf{X}}$ in a $d-$dimensional space such that a future individual with a characteristic vector in $C_{\mathbf{X}}$ has higher probability of being a success that is, a response variable Y of interest belongs to a well defined set C_Y. In the Bayesian predictive approach, $C_{\mathbf{X}}$ is obtained considering an optimality criterion based on the maximization of $P(Y \in C_Y | X \in C_{\mathbf{X}}; \mathcal{D})$ constrained to the class of regions $C_{\mathbf{X}}$ of size α, that is, with fixed predictive probability of screening α. Parametric modeling is a usual way to obtain the predictive distributions required for the formulation of the screening problem. Such modeling often implies the specification of a certain number of assumptions which are difficult to verify in practice. In this work the parametric assumption is relaxed by proposing a Bayesian nonparametric screening methodology.

palavras-chave: Triagem; modelação não paramétrica; métodos MCMC.
keywords: Screening; nonparametric Bayesian models; MCMC methods.

1 Introdução

Os procedimentos de triagem são atualmente usados em vários contextos como a medicina, psicologia, educação, ambiente e controlo de qualidade. Estes procedimentos têm como objetivo a retenção de indivíduos da população de modo a que, para os indivíduos retidos, a probabilidade de uma determinada característica estar presente (a resposta Y pertencer a uma região conhecida C_Y) exceda um determinado valor pré-especificado δ. Os indivíduos que apresentam a característica são rotulados como *sucesso*.

Em várias situações Y é de difícil obtenção direta, devendo ser observada apenas quando o indivíduo tem grande probabilidade de ser classificado como um *sucesso*. A avaliação indireta dessa probabilidade pode ser feita recorrendo à observação de um vetor $d-$dimensional ($d \geq 1$) $\mathbf{X}$ correlacionado com Y e de mais fácil observação, de modo reter indíviduos uma alta probabilidade de serem *sucesso*. Assim, o procedimento de triagem envolve a construção de uma região de especificação $C_{\mathbf{X}}$, no espaço $d-$dimensional, de modo a que um indivíduo futuro com um vetor de características em $C_{\mathbf{X}}$ tenha maior probabilidade de ser identificado como um *sucesso*. No campo preditivo bayesiano a obtenção da região $C_{\mathbf{X}}$ é baseada num critério ótimo assente na maximização de $P(Y \in C_Y | \mathbf{X} \in C_{\mathbf{X}}; \mathcal{D})$, restringida à classe das regiões $C_{\mathbf{X}}$ com probabilidade preditiva de triagem $\alpha = P(\mathbf{X} \in C_{\mathbf{X}} | \mathcal{D})$. Em [3] é apresentado um critério de triagem ótimo de onde resulta a região: $C_{\mathbf{X}} = \left\{ \mathbf{x} \in \mathbb{R}^d : P(Y \in C_Y | \mathbf{x}, \mathcal{D}) \geq k \right\}$ onde k é tal que $P(\mathbf{X} \in C_{\mathbf{X}} | \mathcal{D}) = \alpha$ e $\mathcal{D} = \{(y_1, x_{11}, x_{21}), \ \cdots, (y_n, x_{1n}, x_{2n})\}$ o conjunto de dados.

É prática corrente, tanto em contexto clássico como bayesiano, utilizar modelos paramétricos na construção de $C_{\mathbf{X}}$. Porém, em muitas situações práticas os modelos paramétricos não conseguem descrever o processo que gera as observações, justificando-se assim a necessidade de relaxar o pressuposto paramétrico. Em [1] é apresentada uma abordagem clássica que traz alguma flexibilidade ao problema de triagem, não sendo conhecido trabalho semelhante num quadro bayesiano. Neste trabalho desenvolve-se um método de triagem bayesiano que não especifica qualquer modelo paramétrico.

2 Abordagem não paramétrica bayesiana

Considerando a decomposição $[\mathbf{X}, Y] = [Y|\mathbf{X}][\mathbf{X}]$, a obtenção da solução não paramétrica bayesiana do problema de triagem e das estimativas das CO (probabilidades preditivas que interessam ter em consideração no problema de triagem; veja-se [3] e [2] para uma descrição destas medidas) resume-se à estimação não paramétrica bayesiana de $P(Y \in C_Y | \mathbf{x}; \mathcal{D})$ e $P(X \in C_{\mathbf{X}} | \mathcal{D})$.

Por sua vez, a obtenção dessas estimativas exige a estimação das funções densidade de probabilidade subjacentes por aplicação de métodos não paramétricos de estimação de densidades. Nesse problema, a distribuição desconhecida é vista como um parâmetro aleatório e é considerada uma distribuição *a priori* para esse parâmetro. Este raciocínio requer a introdução do conceito de medidas de probabilidade aleatórias (RPM) que são, genericamente, definidas como medidas de probabilidade sobre uma coleção de funções de distribuição. Há várias RPM descritas na literatura, sendo os processos de Dirichlet (DP) e as árvores de Pólya (PT) as mais estudadas.

2.1 O modelo

Sejam $\{y_i, \mathbf{x}_i\}_{i=1}^n$ os dados, onde $Y \in \mathbb{R}$ e $\mathbf{X} \in \mathbb{R}^d$. Considere-se que $y_i|\mathbf{x}_i \sim f(.|\mathbf{x}_i)$ com $f(.|\mathbf{x}_i)$ desconhecida e para a qual se considerou uma mistura por um processo de Dirichlet dependente (DDP) como distribuição *a priori*:

$$f(.|\mathbf{x}) = \int \phi(.|\mathbf{x}'\beta, \sigma^2) dG(\beta, \sigma^2); G|\alpha, G^* \sim DP(\alpha, G^*),$$

$$G^* = N_d(\beta|\mu_b, s_b)\Gamma(\sigma^2|\tau_1/2, \tau_2/2).$$

Para a completa especificação do modelo consideraram-se a ditribuições: $\alpha|a_0, b_0 \sim \Gamma(a_0, b_0)$, $\tau_2|\tau_{s_1}, \tau_{s_2} \sim \Gamma(\tau_{s_1}/2, \tau_{s_2}/2)$ e $\mu_b|m_0$.

Para a distribuição desconhecida de $\mathbf{X}$ considerou-se uma mistura por uma árvore de Pólya multivariada finita (PT) como distribuição *a priori*:

$$\mathbf{X}|G \sim G; G|c, m, C \sim \mathrm{PT}\left(c, \Pi^{\mathbf{m},\mathbf{C}}\right); p(\mathbf{m}, \mathbf{C}) \propto \mathbf{C}^{-1}; c|a, b \sim \Gamma(a, b).$$

Na simulação *a posteriori* do modelo consideraram-se métodos MCMC.

Como facilmente se reconhece, não é possível obter uma região com expressão em forma fechada, tendo-se implementado uma adaptação do procedimento apresentado em [2] para a aproximar a região ótima.

3 Aplicação

Esta secção apresenta resultados da aplicação do método proposto a 3 conjuntos de dados reais. (X_1, X_2) são níveis de expressão genética de um par de genes. Na Figura 1 encontra-se, para cada estudo, o diagrama de dispersão do par de genes, assim como uma apreciação intuitiva da natureza dos limites da região. A fronteira de natureza quadrática encontra-se também

representada. Na Tabela 1 mostram-se as estimativas das CO das regiões representadas. Analisando os resultados, pode-se concluir que o método proposto apresenta bom desempenho. Os valores das estimativas das CO são satisfatórios. Por exemplo, no conjunto de dados 1 a probabilidade de preditiva de sucesso passa de 0.532 para 0.981 quando $\mathbf{X} \in C_{\mathbf{X}}$ é considerado e a probabilidade de um indivíduo não retido ser um *sucesso* é 0.082.

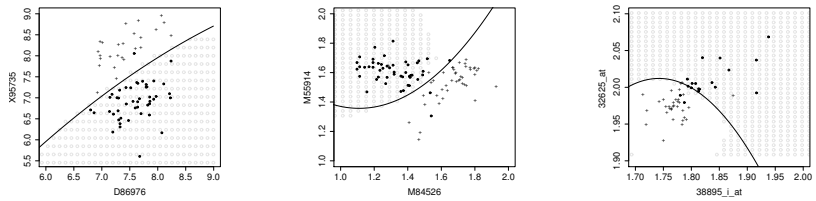

Figura 1: Região de especificação ótima aproximada.

Tabela 1: CO; polinómios de segundo grau na aproximação das fronteiras.

Estudo	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\epsilon}$	$\hat{\beta}$	$\hat{\eta}$
I	0.532	0.501	0.981	0.082	0.924	0.979
II	0.649	0.578	0.985	0.198	0.872	0.975
III	0.352	0.250	0.843	0.181	0.924	0.980

$\gamma{=}P\big(Y{\in}C_Y|\mathcal{D}\big);\alpha{=}P\big(\mathbf{X}{\in}C_{\mathbf{X}}|\mathcal{D}\big);\delta{=}P\big(Y{\in}C_Y|\mathbf{X}{\in}C_{\mathbf{X}};\mathcal{D}\big),\epsilon{=}P\big(Y{\in}C_Y|\mathbf{X}{\notin}C_{\mathbf{X}};\mathcal{D}\big);$
$\beta{=}P\big(\mathbf{X}{\in}C_{\mathbf{X}}|Y{\in}C_Y;\mathcal{D}\big);\eta{=}P\big(\mathbf{X}{\notin}C_{\mathbf{X}}|Y{\notin}C_Y;\mathcal{D}\big).$

4 Conclusões

Neste trabalho apresentou-se uma solução não paramétrica bayesiana para o problema de triagem. O método foi ilustrado em três conjuntos de dados, tendo revelado bom desempenho. Contudo, é necessário proceder a um estudo de simulação de forma a avaliar uma aplicação ampla desta abordagem.

Referências

[1] Boys, R.J., "On a Kernel Approach to a Screening Problem", *J. R. Statist. Soc. B*, Vol.54, No.1 (1992), pp. 157–169.

[2] Ramos, S., Amaral Turkman, M.A., e Antunes, M., "Bayesian classification for bivariate normal gene expression", *Computational Statist. and Data Analysis*, Vol.54, No.8 (2010), pp. 2012–2020.

[3] Turkman, K.F. e Amaral Turkman, M.A., "Optimal screening methods", *J. R. Statist. Soc. B*, Vol.51, No.2 (1989), pp. 287–295.

Agradecimentos: Este trabalho foi parcialmente financiado pela FCT: Projetos $PTDC/MAT/118335/2010$ e $Pest-OE/MAT/UI0006/2011$.

Abordagem bayesiana não paramétrica para o estudo da adequação de modelos

Maria J. Polidoro e *Fernando J. Magalhães*
CEAUL e Instituto Politécnico do Porto
Rua Doutor Roberto Frias
4200-465 Porto, Portugal
e-mail: mjp@estgf.ipp.pt
fjmm@iscap.ipp.p

Maria A. Amaral Turkman
CEAUL e FCUL
Bloco C6, Piso 4 - Campo Grande
1749-016 Lisboa, Portugal
e-mail: maturkman@fc.ul.pt

Resumo: A base de muitas metodologias estatísticas pressupõe que um determinado modelo probabilístico paramétrico se ajusta a um conjunto de dados observados. Se esta suposição falha, a qualidade das inferências realizadas é posta em causa. Uma das soluções proposta pela abordagem bayesiana, para o estudo da adequabilidade de um modelo, consiste em definir um modelo bayesiano não paramétrico alternativo que incorpore o modelo paramétrico em estudo. Seguidamente, a averiguação da adequabilidade do modelo é feita através de métodos de comparação de modelos, destacando-se o factor de Bayes como método de eleição para a comparação.

Neste trabalho, propõe-se um teste de ajustamento bayesiano não paramétrico para o estudo da adequabilidade do modelo exponencial, que considera um modelo bayesiano alternativo baseado em mistura de árvores de Pólya. São ainda referidos os resultados de um estudo de simulação, sobre o desempenho do teste de ajustamento bayesiano com alguns dos testes de ajustamento clássicos.

Abstract The basis for several statistical methodologies assumes that a specified parametric probabilistic model fits a observed data set. If this assumption does not hold, the quality of the inferences is doubtful. One of the solutions proposed by the Bayesian approach, to study the adequacy of a model, is to define a Bayesian nonparametric alternative model that embed the parametric model under study. Next, to study the adequacy of the model, we use measures of comparison between the two models. The Bayes factor is one of the most relevant of such measures.

Encontro Nacional da SPM 2014, Probabilidades e Estatística, pp. 117-120

In this work, we propose a nonparametric Bayesian test of fit to study the adequacy of the exponential model, using as Bayesian alternative model a mixture of Pólya trees. It is also referred some practical examples and the performance of the Bayesian test of fit is compared, through a simulation study, with some of the classic tests.

palavras-chave: teste de ajustamento bayesiano não paramétrico; factor de Bayes; mistura finita de árvores de Pólya; estudo de simulação.

keywords: nonparametric Bayesian test of fit; Bayes factor; finite mixture of Pólya trees; simulation study.

1 Introdução

A distribuição exponencial é uma das mais simples e importantes distribuições utilizadas na modelação de dados que representam o tempo até à ocorrência de um determinado acontecimento de interesse. O estudo da adequabilidade da distribuição exponencial é fundamental para que as inferências realizadas sejam válidas.

A abordagem clássica para o estudo da adequabilidade tem sido um tema bastante debatido. Henze e Meintains [1] fizeram um estudo de simulação Monte Carlo, onde compararam vinte e uma estatísticas de teste para o estudo da adequabilidade da distribuição exponencial contra dezoito distribuições alternativas. O estudo exaustivo dos referidos autores dá indicações que algumas das dezoito estatísticas de teste, estão entre as mais potentes e simples de calcular.

A abordagem bayesiana sobre métodos para estudar a adequabilidade de um modelo contínuo, ao contrário da literatura clássica, é ainda muito reduzida e focada no estudo da adequabilidade da distribuição gaussiana (ver [2] e [3]).

Neste trabalho propõe-se um teste bayesiano não paramétricos para o estudo da adequabilidade da distribuição exponencial considerando como modelo bayesiano não paramétrico alternativo (H_1), o modelo de mistura de árvores de Pólya (ver [4] e [5]). A averiguação da adequabilidade do modelo proposto na hipótese nula (H_0) é feita utilizando o factor de Bayes.

Na secção 2 apresenta-se muito resumidamente a abordagem bayesiana ao problema (um estudo pormenorizado encontra-se em [6] e [7]) e na secção 3 apresentam-se as conclusões obtidas, através de um estudo de simulação, sobre o desempenho do novo teste com alguns dos testes clássicos considerados como os mais potentes.

2 Abordagem Bayesiana

O teste de ajustamento bayesiano não paramétrico pressupõe a comparação de dois modelos. O modelo bayesiano paramétrico (H_0) é dado por

$$X_i|\lambda \overset{\text{iid}}{\sim} \text{Exp}(\lambda), \text{ para } i = 1, 2, \ldots, n,$$
$$\lambda \sim \text{Ga}(a, b)$$

e o modelo bayesiano não paramétrico ou alternativo (H_1) é dado por

$$X_1, X_2, \ldots, X_n|G \overset{\text{iid}}{\sim} G$$
$$G|\Pi, \mathcal{A}_\lambda \sim \text{MPT}_M(\Pi, \mathcal{A}_\lambda),$$
$$\lambda \sim \text{Ga}(a, b)$$

onde $\text{MPT}_M(\Pi, \mathcal{A}_\lambda)$ define uma distribuição *a priori* mistura finita de árvores de Pólya, com parâmetros $(\Pi, \mathcal{A}_\lambda)$ e M níveis pré-especificados.

A averiguação da adequabilidade do modelo exponencial é feita utilizando o factor de Bayes a favor de H_0 e contra H_1, dado por

$$\text{BF}_{01}(x) = \frac{p_0(x)}{p_1(x)}.$$

onde $p_0(x)$ e $p_1(x)$ é, respetivamente, a distribuição preditiva *a priori* de cada modelo.

Com o objectivo de comparar o desempenho do teste bayesiano não paramétrico proposto com alguns dos testes clássicos mais potentes, realizou-se um estudo de simulação. Foram utilizados 6 testes clássicos e várias distribuições alternativas de entre as distribuições frequentemente consideradas em outros estudos e com diferentes taxas de falha; a distribuição Gama, a distribuição Weibull, a distribuição Log-Normal, a distribuição Half-Normal, a distribuição do Qui-Quadrado e a distribuição Half-Cauchy. Os parâmetros destas distribuições alternativas foram escolhidos de modo a que as formas das distribuições fossem diferindo da forma de uma distribuição exponencial padrão. Foram simuladas amostras com diferentes dimensões: $n = 25$, 50 e 100. Pormenores e resultados deste estudo podem ser encontrados em [6] e [7].

3 Conclusões

Para as distribuições alternativas com taxa de falha crescente, notou-se que a potência empírica do teste de ajustamento bayesiano é quase sempre

superior à dos testes clássicos. Por outro lado, quando as amostras simuladas são obtidas a partir de distribuições alternativas com função taxa de falha decrescente, o teste de ajustamento bayesiano é, pelo menos, tão potente quanto os clássicos. Saliente-se, ainda, o facto de que quando as amostras são de pequena dimensão, o teste de ajustamento bayesiano é o que apresenta melhor desempenho. Assim, pode afirmar-se que o estudo de simulação efetuado, não sendo exaustivo, na medida que se restringiu o trabalho a distribuições alternativas usualmente consideradas em outros estudos, permite concluir que o teste bayesiano não paramétrico proposto para o estudo da adequabilidade da distribuição exponencial tem, de uma forma geral, um excelente desempenho.

Agradecimentos: Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia, no âmbito do projecto PEst-OE/MAT/UI0006/2014.

Referências

[1] N. Henze e S. Meintanis, "Recent and classical tests for exponentiality: A partial review with comparisons". *Metrika*, Vol. 61, (2005), pp.29-45

[2] J. O. Berger e A. Guglielmi, "Bayesian Testing of a Parametric Model versus Nonparametric Alternatives", *Journal of the American Statistical Association*, Vol. 96, (2001), pp. 174-184.

[3] S. T. Tokdar e R. Martin, (2011). "Bayesian test of normality versus a Dirichlet process mixture alternative", Tech. Rep., 2011.

[4] N. L. Hjort, C. Holmes, P. Muller e S. G. Walker, *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[5] M. Lavine, "Some aspects of Polya tree distributions for statistical modeling", *The Annals of Statistics*, Vol. 20, (1992), pp. 1222-1235.

[6] M. J. Polidoro, F. J. Magalhães e M. A. Amaral Turkman, "Classical and Bayesian Goodness-of-fit Tests for the Exponential Model: A Comparative Study", *14th International Conference on Computacional Science and its Applications*, B. Murgante et al. (Eds.): ICCSA 2014, Part III, LNCS 8581, pp. 483–497.

[7] M. J. Polidoro, "Metodologia bayesiana e adequação de modelos", Tese de Doutoramento, Universidade de Lisboa, Portugal, 2014.