UNIVERSIDADE DE SÃO PAULO

BAYESIAN COMPUTATION FOR DICHOTOMOUS VARIABLES WITH CLASSIFICATION ERRORS

JOSEMAR RODRIGUES
VANDA DONIZETTI R. SILVEIRA

Nº 55

NOTAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2577

BAYESIAN COMPUTATION FOR DICHOTOMOUS VARIABLES WITH CLASSIFICATION ERRORS

JOSEMAR RODRIGUES
VANDA DONIZETTI R. SILVEIRA

Nº 55

NOTAS DO ICMC
Série Estatística

São Carlos
Dez./1998

Análise Bayesiana para dado binários com erros de classificação

Inferência Bayesiana para dados binários incluindo erros de classificação é estudado como uma mistura de duas distribuições de Bernoulli. Como a análise Bayesiana geralmente implica em cálculos complexos, o método de Monte Carlos com dados ampliados é desenvolvido para obter os resumos marginais a posteriori. Variáveis latentes foram introduzidas para indicar qual componente da mistura gerou a informação com erro de classificação. Também, um procedimento Bayesiano baseado no conceito de "p-value' e na distância de variação total foi introduzido para medir o efeito do erro na distribuição marginal a posteriori. Uma ilustração com dados simulados é considerada.

Bayesian computation for dichotomous variables with classification errors

Josemar Rodrigues and Vanda Donizetti R. Silveira
ICMC-USP-C.P. 668,13560-970,São Carlos, SP, Brazil

Key Words: Calibrated measure; Bayes p-value; Gibbs sampling; Latent variables; Mixture distributions; Variational distance.

Abstract

Bayesian inference for dichotomous data including errors of classification is studied as a mixture of two Bernoulli distributions. Since a formal Bayesian analysis leads to intractable calculations, a Markov Chain Monte Carlo method with data augmentation is developed to compute the summary and marginal posteriors . We introduce a latent variable that indicates which component of the mixture gives rise the observation with errors. Also, a Bayesian procedure based on the Beyes p-value and the total variation distance to measure the effect of the errors on the data is formulated. An illustration with simulated data is considered.

1 Introduction

In recent years much of the research on sampling practice has been devoted to formulate procedures to analyse data with errors. These errors can produce serious effect on the estimation of the population parameters. In this paper we are particulary interested in dichotomous data with possible response errors. For example, suppose we are interested in the proportion of consumers who have seen a certain advertisement. Two types of errors are possible: an individual who has not seen the ad might report having seen it, and someone who has seen the ad might report not having seen it. If you do not take into account this kind of errors the sample proportion will be very unrealistic and the tendency is to overestimate the proportion of consumers who have seen the ad. Gaba and Winkler (1992) introduced an exact Bayesian approach for this problem considering conjugated priors. The purpose of this paper is to look at this problem in a different way, that is, to see the data rising from a mixture of two Bernoulli distributions. Since the formal Bayesian analysis is intractable, we introduce a latent variable (Tanner and Wong, 1987) to indicate which component of the mixture the observation is coming and based on the data and these latent variables the marginal posteriors are obtained via Gibbs sampling. The paper is organized as follows: Section 2 the mixture model is motivated and some important relations are introduced. The Bayesian procedure via Gibbs sampling and a divergence measure to evaluate the effect of the errors on the data are presented in section 3 and an illustrative example with simulated data is studied in section 4. The paper ends with a conclusion in section 5.

2 The dichotomuous data with errors via the mixture model.

The proportion of individuals satisfying a certain condition is of interest. For various reasons the data is reporting with errors, for example, intentional lying, mistakes, etc. are included in the data with serious implications for the usual procedures (see, Gaba and Winkley, 1992). The problem is how to use this data to make inference about the proportion of interest in presence of these possible errors.

Let p the proportion of consumers who purchase a given product and suppose that n consumers of a given product have been interviewed whether purchased it or not. For each consumer we define the following dichotomous variable with errors:

$$X_i = \begin{cases} 1 & \text{the ith-consumer purchases the product} \\ 0 & \text{the ith- consumer does not purchase the product} \end{cases} \tag{1}$$

Let $q = P[X_i = 1], i = 1, \ldots, n$ and $p = P[Y_i = 1]$, where Y_i is the dichotomous variable without error. The parameter p is the real proportion of consumers who purchase a given product. Our purpose is to get information about p, based on the dichotomous data with errors, $\{X_1, \ldots, X_n\}$. To motivate our mixture distribution for this data we introduce the following notations:

$$\theta_1 = P[X_i = 0 \mid Y_i = 1] :$$

probability that a purchase is erroneously recorded as nonpurchaser.

$$\theta_2 = P[X_i = 1 \mid Y_i = 0] :$$

the probability that the ith-purchaser is mistakenly recorded as a purchaser. From these notations the following relation will be useful for our purposes:

$$q = p(1 - \theta_1) + (1 - p)\theta_2,$$

From the above notations we can easily find the joint ditribution of (X_i, Y_i) which is showed in the following way:

Table 1.: The joint distribution of (X, Y).

-	$X_i = 0$	$X_i = 1$
$Y_i = 0$	$(1-p)(1-\theta_1)$	$(1-p)\theta_1$
$Y_i = 1$	$p\theta_2$	$p(1-\theta_2)$

Motivated by the joint distribution presented in Table 1., it is easy to see that the observed dichotomous data with errors, $X_i, \ldots, X_n$ is generated by the following mixture distribution:

$$f(x \mid \theta_1, \theta_2, p) = pBin(x \mid 1 - \theta_1) + (1 - p)Bin(x \mid \theta_2) \qquad (2)$$

where $1 - \theta_1 > \theta_2$. The aim of this restriction is to avoid identification problems as it happens in the usual likelihood approach (see, Gaba and Hinkley, 1992). The notation $Bin(. \mid .)$ means the Bernoulli distribution. This kind of formulation is totally different from that one given by Gaba and Winkley (1992).

3 Gibbs Sampling

In this section we develop the conditional distributions used in Gibbs sampling algorithm. Gibbs sampling is a MCMC technique(see, Gelfand and Smith, 1990). The transition distribution of the Markov chain is the product of several condicional densities. The stationary distribution of the chain is the posterior distribution we desire for p and q.

Since the formal Bayesian analysis for mixture model are in general intractable (see Diebolt and Robert, 1994), we introduce latent variables denoted by $I_i, i = 1, \ldots, n$, so, we can obtain the likelihood function and the posterior distributions of p, q, θ_1 and θ_2, by iteratively sampling I_i from $f(I_i \mid p, \theta_1, \theta_2, X)$, where $X = (X_1, \ldots, X_n)$, and sampling (p, θ_1, θ_2) from $f(p, \theta_1, \theta_2 \mid X, I)$. Let $I = (I_1. \ldots, I_n)$ the vector of the latent variables, where $I_i = 1$ if the ith observation, X_i, is generated by the first component of (3) and $I_i = 0$ otherwise. Let us to consider the conditional density of I_i, given X, p, θ_1, θ_2, to be Bernoulli(p_i) with

$$p_i = \frac{pf(x_i \mid \theta = 1 - \theta_1)}{pf(x_i \mid \theta = 1 - \theta_1) + (1 - p)f(x_i \mid \theta = \theta_2)}, \tag{3}$$

where

$$f(x_i \mid \theta) = \theta^{x_i}(1 - \theta)^{1 - x_i}.$$

The likelihood function based on the augmented data (X, I) is

$$\begin{aligned} L[p, \theta_1, \theta_2 \mid I, X] &= \\ &= p^{\sum_{i=1}^n I_i}(1 - p)^{n - \sum_{i=1}^n I_i}\theta_1^{\sum_{i=1}^n (1 - X_i) I_i}(1 - \theta_1)^{\sum_{i=1}^n X_i I_i}. \\ &. \quad \theta_2^{\sum_{i=1}^n X_i I_i}(1 - \theta_2)^{\sum_{i=1}^n (1 - X_i)(1 - I_i)}. \end{aligned} \tag{4}$$

To develop a Bayesian analysis for (p, θ_1, θ_2) we consider the following priors:

$$\begin{aligned} p &\sim B[a, b] : \quad a, b : \text{known} \\ \theta_i &\sim B[a_i, b_i] : \quad i = 1, 2, \quad \text{known} \end{aligned} \tag{5}$$

where $B[.,.]$ denotes the beta distribution. Also, we assume that p, θ_1 and θ_2 are independent parameters. So, the joint posterior distribution for (p, θ_1, θ_2), given the augmented data (X, I), is

$$\begin{aligned} &\pi[p, \theta_1, \theta_2 \mid X, I] \propto p^{a + \sum_{i=1}^n I_i}(1 - p)^{b + n - \sum_{i=1}^n I_i}\theta_1^{a_1 + \sum_{i=1}^n (1 - X_i) I_i}. \\ &. \quad (1 - \theta_1)^{b_1 + \sum_{i=1}^n X_i I_i}\theta_2^{a_2 + \sum_{i=1}^n X_i (1 - I_i)}(1 - \theta_2)^{b_2 + \sum_{i=1}^n (1 - X_i)(1 - I_i)}. \end{aligned} \tag{6}$$

The Gibbs algorithm is given by the following stages :

- Stage 1: Construct $(I_i, 1 - I_i), \quad i = 1, \ldots, n$, given p, θ_1, θ_2, from the Bernoulli distribution with p_i given by (4).

- Stage 2:

$$p \mid X, I \sim B[a + \sum_{i=1}^{n} I_i; n + b - \sum_{i=1}^{n} I_i]$$

$$\theta_1 \mid X, I \sim B[a_1 + \sum_{i=1}^{n}(1 - X_i)I_i; b_1 + \sum_{i=1}^{n} X_i I_i]$$

$$\theta_2 \mid X, I \sim B[a_2 + \sum_{i=1}^{n} X_i(1 - I_i); b_2 + \sum_{i=1}^{n}(1 - X_i)(1 - I_i)]$$

Consumers may intentionally misreport, may not remember their behavior acurately, or may misunderstand survey questions; and so on. These kind of errors can have a serious effects on inferences, so, we propose in this paper a Bayesian procedure to measure the quality of the data X. The effects of the errors on the data can be measured in the following two stages:

- **Bayes p-value:**
  Let us assume that $\pi(p \mid X) \sim B(a_3, b_3)$ and $\pi(q \mid X) \sim B(a_4, b_4)$. Our purpose in this stage is to find (a_3, b_3) and (a_4, b_4) such that the marginal beta densities of p and q can be well fitted to their Gibbs samples $p^* = \{p^{(l)}, l = 1, \ldots, m\}$ and $q^* = \{q^{(l)}, l = 1, \ldots, m\}$, respectively. As suggested by Gelman et al. (1995), we define the Bayes p-value for the parameter p, B_p, in the following way:

  $$B_p = \frac{\sum_{j=1}^{J} I_{[T_{a_3,b_3}(p^{rep,j}) \geq T_{a_3,b_3}(p^*)]}}{J}, \tag{7}$$

  where the sum is taken over the J replicated data $p^{rep,j} = \{p_1^{rep,j}, \ldots, p_m^{rep,j}\}$ generated by $\pi(p \mid X)$ for a fixed value (a_3, b_3), where

  $$T_{a_3,b_3}(p^\star) = \sum_{l=1}^{m} \frac{(p^{(l)} - \frac{a_3}{a_3+b_3})^2}{\frac{a_3 b_3}{(a_3+b_3)^2(a_3+b_3+1)}}.$$

  We choose the pair (a_3, b_3) which the Bayes p-value is close to 0.5. The same procedure we apply to find (a_4, b_4).

- **The variational distance:(Peng and Dey, 1995)**
  Let us denote the marginal beta distribution for p and q by $\pi_p(\theta) = B(\theta; a_3, b_3)$ and $\pi_q(\theta) = B(\theta; a_4, b_4)$, respectively, where (a_k, b_k), $k = 3, 4$, were obtained in the first stage using the Bayes-p

value. The variational distance between these two marginal densities is defined as:

$$D_V = D[\pi_p(\theta), \pi_q(\theta)] = \frac{1}{2}\int \mid \frac{\pi_p(\theta)}{\pi_q(\theta)} - 1 \mid \pi_q(\theta)d\theta. \qquad (8)$$

This distance was calibrated by Peng and Dey (1995) suggesting the following scale :

$$D_V > 0.25 \Longrightarrow \text{a strong influence of the errors on the data}$$

$$0.1 \leq D_V \leq 0.25 \Longrightarrow \text{a mild influence of the errors on the data}$$

$$D_V < 0.1 \Longrightarrow \text{no influence of the errors on the data}$$

A Monte Carlo procedure to compute D_V can be found in Peng and Dey, 1995.

4 An illustrative example

Some numerical results for Bayesian inference are given in this section for simulated data. We simulate the data, X, with $n = 10$, from the mixture distribution (3) for $p = 0.20, \theta_1 = 0.03$ and $\theta_2 = 0.60$ obtaining the following data:

$$X = [0, 1, 1, 1, 1, 1, 1, 1, 0, 0].$$

We consider the priors for p, θ_1 and θ_2 given in (6) with $a = 10, b = 30, a_1 = 2, b_1 = 48, a_2 = 30$ and $b_2 = 20$. We munitor the convergence of Gibbs sampler using the Gelman and Rubin (1992) method that uses the analysis of variance technique to determine whether further iterations are need. We generated 10.000 observations and "burn-in" 2.000 observations and after that we take each every 10th draw having a final Gibbs sample of size 800. In Table 1, we have obtained the summary posteriors for p, θ_1, θ_2 and q. Figure 1 and 2 give the the histograms and fitted marginal densities of p and q obtained by the Bayes-p value, respectively. The estimated divergence measure between this two marginal densities is, $D_V = 0.50$ indicating a strong influence of the errors on the data.

Table 1: Summaries of posterior inferences

-	2.5%	25%	Median	75%	97.5%
p	0.1056	0.1643	0.1989	0.3204	0.3204
θ_1	0.0047	0.0239	0.0404	0.0626	0.1181
θ_2	0.4891	0.5779	0.6193	0.6658	0.7330
q	0.5736	0.6467	0.6872	0.7285	0.7854

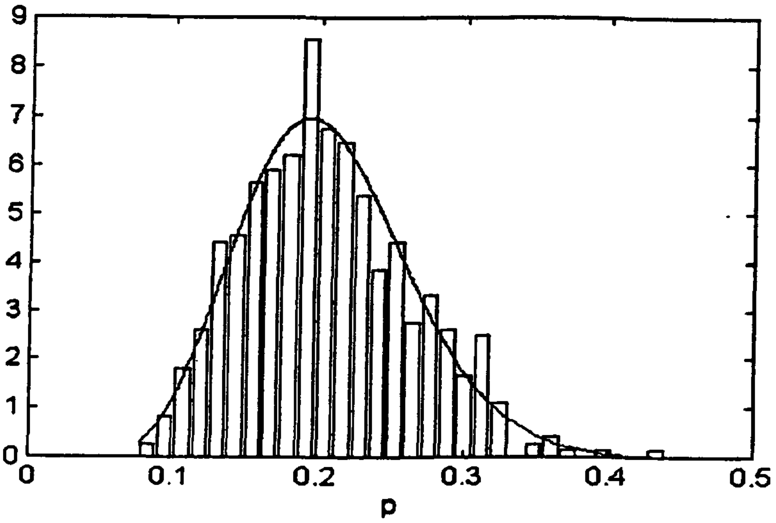

Figure 1. Histogram of 300 draws of posterior marginal of p

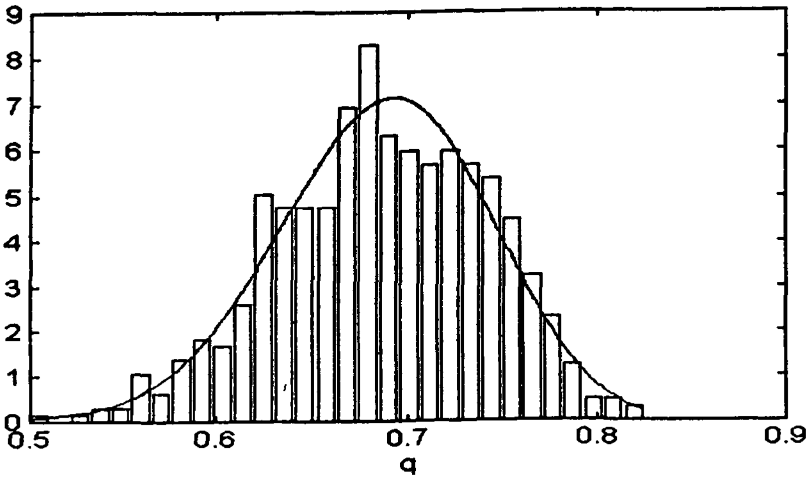

Figure 2. Histogram of 300 draws of posterior marginal of q.

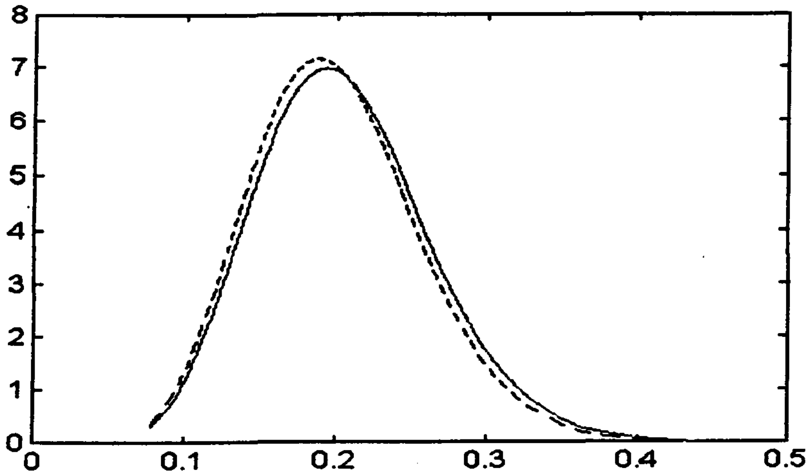

Figure 3. Dashed line: exact posterior marginal of p;full line: fitted posterior marginal of p.

Figure 3 shows the exact posterior marginal of p obtained by a mixture of beta distributions given by Gaba and Winkley (1992) and the fitted beta posterior marginal of p obtained by Gibbs sampling via Bayes p-value. This figure shows a very nice beta marginal approximation for the exact mixture marginal density given by Gaba and Winkley.

5 Final considerations

The divergence measure proposed in this paper, Figure 1 and 2 and Table 1 of our numerical example show very clearly the effect of the errors on the inference of the parameter of interest. These analysis show that errors can have a signficant impact on inference about the real proportion of consumers. Using standard procedures that ignore such errors can result in misleading inferences. Our results have important implications from the computation point of view and measure in an efficient way the lost of information we have when introducing errors in the data. Also, it was shown that a beta marginal density for p is a good approximation for the exact mixture proposed by Gaba and Winkley (1992).

References

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions throught Bayesian Sampling, J.R.Statist. SOc., v.56, no.2, pp. 363-375

Gaba, A. and Winkley, R. (1992). Implications of errors in survey data: A Bayesian model, Management Science, v.38, no. 7, pp. 913-925

Gelman,A.; Carlin, J.B.; Stern, H.S. and Rubin, D.B. (1995). Bayesian Data Analysis, Chapman&Hall

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling -Based approaches to calculating marginal densities, J. Amer. Statist. Assoc.,85, pp. 398-409.

Gelman,A.E. and Rubin,D. (1992). Inference from iterative simulation using multiple sequences, Statistical Science, 7, pp. 457-472.

Peng, F. and Dey, D.K., 1995. Bayesian analysis of outlier problems using divergence measures, The Canadian J. of Statistics, v. 23, no.2, pp. 199-213.

Tanner, M.and Wong, W. (1987). The calculation of posterior distributions by data augmentation , J.Amer.Statist.Assoc.,82, pp. 528-550.

NOTAS DO ICMC

SÉRIE ESTATÍSTICA

054/98 ACHCAR, J.A.; FORTULAN, V.C. – Relação entre o uso de hormônio e câncer em mulheres: uma aplicação de meta-análise sob um enfoque bayesiano.

053/98 CID, J. E.R.; ACHCAR, J..A. – Bayesian inference for nonhomogeneous poisson processes in software reliability models assuming nonmonotonic intensity functions.

052/98 ANDRADE FILHO, M.G; MIZOI, M.F. – Aplicação de MCMC na estimação de máxima verossimilhança para processos AR(p) e MA(q).

051/98 RODRIGUES, J. – Bayesian analysis for the accelerated life tests with informative prior distributions obtained from fixed stresses.

050/98 ACHCAR, J.A.; ANDRADE, .M.G.; LOIBEL, S. - A bayesian analysis for homogeneous poisson processes with change-points.

049/98 ACHCAR, J.A; PEREIRA, G.A. - Use of mixture of exponential power distributions for interval-censored survival data in presence of covariates.

048/98 CID, J.E.R.; ACHCAR, J.A. - Software reliability considering the super position of non-homogeneous Poisson processes in the presence of a covariate.

047/98 ACHCAR, J.A.; PEREIRA, G.A. - Use of exponential power distributions for mixture models in the presence of covariates.

046/98 ANDRADE, M.G.; HUTTER, C.F.F. - Teste de sazonalidade para função de autocorrelação de processos auto-regressivos periódicos - PAR (pm)

045/98 ACHCAR, J.A.; ANDRADE, M.G.; LOIBEL, S. - Weibull hazard function with a change-point: a bayesian approach using Markov chain Monte carlo methods.