A New Multinomial Model and a Zero Variance Estimation

Luciana Dalla Valle, Fabrizio Leisen

▶ **To cite this version:**

Luciana Dalla Valle, Fabrizio Leisen. A New Multinomial Model and a Zero Variance Estimation. Communications in Statistics - Simulation and Computation, 2010, 39 (04), pp.846-859. 10.1080/03610911003650375 . hal-00583562

HAL Id: hal-00583562
https://hal.science/hal-00583562

Submitted on 6 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Multinomial Model and a Zero Variance Estimation

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2009-0187.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	22-Jan-2010
Complete List of Authors:	Dalla Valle, Luciana; University of Milan Leisen, Fabrizio; Universidad de Navarra
Keywords:	Multinomial Dirichlet Model, Multinomial Beta, Zero Variance Principle
Abstract:	The analysis of categorical response data through the Multinomial model is very frequent in many statistical, econometric and biometric applications. However, one of the main problems is the precise estimation of the model parameters when the number of observations is very low. We propose a new Bayesian estimation approach where the prior distribution is constructed through the Multivariate Beta of Olkin and Liu (2003). Moreover, the application of the Zero-Variance principle allows us to estimate moments in Monte Carlo simulations with a dramatic reduction of their variances. We show the advantages of our approach through applications to some toy examples.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
DallaValleLeisenR2.tex communications-simulation.zip	

scholarONE™
Manuscript Central

For Peer Review Only

A New Multinomial Model and a Zero Variance Estimation

Luciana Dalla Valle*, Fabrizio Leisen†

January 22, 2010

Abstract

The analysis of categorical response data through the Multinomial model is very frequent in many statistical, econometric and biometric applications. However, one of the main problems is the precise estimation of the model parameters when the number of observations is very low. We propose a new Bayesian estimation approach where the prior distribution is constructed through the transformation of the Multivariate Beta of Olkin and Liu (2003). Moreover, the application of the Zero-Variance principle allows us to estimate moments in Monte Carlo simulations with a dramatic reduction of their variances. We show the advantages of our approach through applications to some toy examples, where we get efficient parameter estimates.

1 Introduction

Our aim is to introduce a new model for independent and identically distributed trials with more than two possible outcomes and to find out more efficient parameter estimates for this model.

The usefulness of such a model is proved by the number of applications in different fields, such as economics, statistics and biometrics. Examples include vote choice in multi-party elections, choice of product in consumer choice models, choice of location for foreign direct investments by companies and the effect of risk factors for patients affected by a certain disease. For example, in the paper by Albert and Chib (1993),

*Department of Economics, Business and Statistics, Faculty of Political Science, University of Milan, Italy. luciana.dallavalle@unimi.it

†Faculty of Economics, University of Navarra, Campus Universitario, edificio de biblioteca (entrada este), 31008, Pamplona, Spain. fabrizio.leisen@gmail.com

For Peer Review Only

the authors predict the Carter/Ford vote in the 1976 US Presidential election using socioeconomic and regional variables. Moreover, the work of Kamakura et al. (1996) finds out consumer segments of households on the basis of their preferences toward different brands of peanut butter. Then, in order to determine the preferred foreign direct investments entry strategies in China, Wei et al. (2005) analyze data about more than 10000 foreign investment projects. At last, in the paper by Fine et al. (2004) the risk factors (as smoking and drinking) that mainly contribute to chronic disease prevalence are identified.

As it is well known, the general model for polychotomous data is the *Multinomial*, a distribution used to describe a wide variety of phenomena. From a Bayesian perspective, the most popular prior for the Multinomial is the *Dirichlet* distribution, which is the conjugate prior.

We may cite a number of works applying this type of model. Gelman et al. (2004), for instance, apply the Multinomial-Dirichlet model to a sample survey question with three possible responses about the preferences in the US presidential elections of 1988. Rannala and Mountain (1997) employed the model for a population genetics problem, where genotypes were used to identify individuals who are immigrants, or have recent immigrant ancestry. Moreover, applications to sampling theory can be found, for example, in Brier (1980), who used the Multinomial-Dirichlet distribution for cluster sampling to take into account the similarity of responses of members of the same cluster. The Multinomial-Dirichlet model was also implemented in marketing applications, as in Goodhardt et al. (1984), where the aim was to model buyers behavior and their choice among different product brands.

Moreover, the Dirichlet distribution was commonly used as a prior for samples generated by the Multinomial density with unknown number of categories and unknown probabilities, as illustrated in Boender and Rinnooy Kan (1987). The model is also a useful tool in case of "extra variation" or "overdispersion", as pointed out by Morel and Nagaraj (1993). More recently, Seaman and Richardson (2001 and 2004) analyzed case-control studies with a retrospective likelihood and a Dirichlet prior for the exposure probabilities in the control group.

Efforts to propose alternative priors to the Dirichlet distribution are not very frequent in the literature. For example, the logistic-normal distribution may be seen as a substitute for the Dirichlet prior to model contingency table data as well as for the analysis of compositional and probabilistic data, as advised by Aitchison and Shen (1980).

In this paper we propose an alternative prior for the Multinomial Model. This prior is constructed with the multivariate Beta distribution of Olkin and Liu (2003), with a construction method that resembles the stick breaking construction of the Dirichlet Process introduced by Ferguson (1973). Indeed, the Dirichlet Process could be defined through a random probability measure $\tilde{p} = \sum_{i=1}^{\infty} p_i \delta_{x_i}$ where the weights p_i are equal

For Peer Review Only

to $V_i \prod_{j=1}^{i}(1-V_j)$ with $V_1, ..., V_n, ...$ i.i.d. Beta(1,α), $\alpha > 0$. For an account on the so called *stick breaking priors* see Ishwaran and James (2001).

The resulting posterior distribution generated by this proposal prior has some attractive properties compared to the traditional model and it is particularly adequate to some applications we illustrate in the following sections, where we study the characteristics and the behavior of the new Multinomial model. The estimation of the parameters could be done with an independent Metropolis-Hastings algorithm introduced by Metropolis et al. (1953) and improved by Hastings (1970). However, with very few data, the problem of a precise estimation of the parameters is of particular concern, and a simulation with the Metropolis-Hastings algorithm could be not enough for having a good estimation of the parameters of the new multinomial model. An improvement for the estimation could be achieved through the so called Variance Reduction techniques. In the Markov Chain Monte Carlo setting a typical way for doing this is to choose in the simulation an updating law that gives rise to a sample with the lowest asymptotic variance. In the paper of Peskun (1973) a simple rule on the transition matrices is given in the discrete setting for choosing a transition rule with a low asymptotic variance. More recently, Hobert and Marchev (2007) shown that the transition rule of a new class of data augmentation algorithms has a lower asymptotic variance than the usual data augmentation algorithm.
Another way for achieving a better variance is by changing the estimator of the parameter but not a lot of papers in the statistical literature are devoted to this reduction strategy. In Casella and Robert (1996) a "Rao-Blackwellizzation" is used for obtaining better estimates and Delmas and Jourdain (2006) used the "Randomness Recicle" idea for improving the variance. In this paper this strategy is followed and in particular we use the *Zero Variance Principle* introduced by Assaraf and Caffarel (1999) for estimating the parameters when we have few data. Their machinery is a quite general way of constructing estimators with better variance and here it is applied to samples of the new multinomial model generated with the independent Metropolis Hasting Algorithm.

The paper is organized as follows. Section 2 introduces the Multinomial-Dirichlet model. Section 3 illustrates in detail the methodology we propose, analyzing the construction of the prior and the derivation of the posterior distribution. Section 4 describes the simulation methods we adopted to estimate the model parameters and outlines the application of our model to some toy examples, showing the simulation results. In Section 5 the definition of the Zero Variance principle is given, focusing on the continuous case. In section 6 the application of the Zero Variance principle is illustrated, with a comparison between the Multinomial-Dirichlet and the proposed model. Finally, remarks and conclusions are given in Section 7.

For Peer Review Only

2 The Multinomial-Dirichlet Model

Suppose to have categorical dependent variables with more than two response categories and that the outcomes fall into one of these categories. Let Ω be the sample space with exaustive and mutually exclusive elements $\{\omega_1, \omega_2, \ldots, \omega_n\}$, where $n \geq 2$, and suppose that each observation falls into one category ω_i (for $i = 1, \ldots, n$). Suppose to have N independent trials from Ω with identical probability distribution $P(\omega_i) = \theta_i$, for $i = 1, \ldots, n$ where $\theta_i \geq 0$ and $\sum_{i=1}^{n} \theta_i = 1$ (Walley (1996)). Then let the random variables X_i denote the number of observations of category ω_i over the N trials. The vector $\underline{X} = (X_1, \ldots, X_n)$ follows a *multinomial* distribution with parameters N and $(\theta_1, \ldots, \theta_n)$ (see Balakrishnan et al. (1997)). The likelihood function of the Multinomial distribution is the following:

$$L(x_1, \ldots, x_n | \theta_1, ..., \theta_n) = \binom{n}{x_1 \ldots x_n} \prod_{i=1}^{n} \theta_i^{x_i},$$

where X_i is a non negative integer $(i = 1, \ldots, n)$ and $\sum_{i=1}^{n} X_i = N$.

From a Bayesian perspective we need to elicitate the prior distribution for the vector $(\Theta_1, \ldots, \Theta_n)$ (Berger (1985)). Assuming that this distribution is denoted by $h(\theta_1, \ldots, \theta_n)$, according to Bayes' theorem the posterior density function for the model is, up to a normalizing constant:

$$\pi(\theta_1, \ldots, \theta_n | \underline{x}) \propto \binom{n}{x_1 \ldots x_n} \prod_{i=1}^{n} \theta_i^{x_i} h(\theta_1, \ldots, \theta_n). \quad (1)$$

The most popular choice for the joint density of the vector $(\Theta_1, \ldots, \Theta_n)$ is the Dirichlet distribution, which is the conjugate prior. If $(\Theta_1, \ldots, \Theta_n) \sim Dirichlet(\alpha_1, \ldots, \alpha_n)$, then the probability density function is:

$$h(\theta_1, \ldots, \theta_n) \propto \prod_{i=1}^{n} \theta_i^{\alpha_i - 1}.$$

As it is well known, this choice of prior distribution for $h(\theta_1, \ldots, \theta_n)$ gives us a posterior which is again a Dirichlet with parameters $(\alpha_1 + x_1, \ldots, \alpha_n + x_n)$, taking the form:

$$\pi(\theta_1, \ldots, \theta_n | \underline{x}) \propto \prod_{i=1}^{n} \theta_i^{\alpha_i + x_i - 1}.$$

This is called the *Multinomial-Dirichlet* model (Ghosh et al. (2006)). This distribution should not be confused with Dirichlet-Multinomial originally proposed by Mosimann (1962) as Compound Multinomial distribution.

For Peer Review Only

3 The new Multinomial model

In order to define a *new Multinomial* model, we propose a new prior distribution for $\Theta = (\Theta_1, \ldots, \Theta_n)$.

The vector $\Theta = (\Theta_1, \ldots, \Theta_n)$ is a vector of probabilities with the usual linear constraint $\Theta_1 + \ldots + \Theta_n = 1$. Without loss of information the dimensionality of Θ can be reduced by removing its last component Θ_n. Note that $(\Theta_1, \ldots, \Theta_{n-1})$ takes values in this set

$$\left\{ (y_1, \ldots, y_{n-1}) : \sum_{i=1}^{n-1} y_i \leq 1 \right\}.$$

Let $g(\theta_1, \ldots, \theta_{n-1})$ be the density of the random vector $(\Theta_1, \ldots, \Theta_{n-1})$. Since $\theta_n = 1 - \sum_{i=1}^{n-1} \theta_i$, we can rewrite equation (1), the posterior distribution, as:

$$\pi(\theta_1, \ldots, \theta_{n-1} | \underline{x}) \propto \binom{n}{x_1 \ldots x_n} \left(1 - \sum_{i=1}^{n-1} \theta_i\right)^{x_n} \prod_{i=1}^{n-1} \theta_i^{x_i} \cdot g(\theta_1, \ldots, \theta_{n-1}). \quad (2)$$

Therefore, to calculate the posterior $\pi(\theta_1, \ldots, \theta_{n-1} | \underline{x})$, we have to determine the prior distribution $g(\theta_1, \ldots, \theta_{n-1})$. In order to do that, we introduce the random vector $(V_1, \ldots, V_{n-1})$, which is distributed according to the *Multivariate Beta of Olkin* (see Olkin and Liu (2003)). The density of $(V_1, \ldots, V_{n-1})$ is the following:

$$f(v_1, \ldots, v_{n-1}) \propto \frac{\prod_{i=1}^{n-1} v_i^{a_i - 1}}{\prod_{i=1}^{n-1} (1 - v_i)^{a_i + 1}} \left[1 + \sum_{i=1}^{n-1} \frac{v_i}{1 - v_i}\right]^{-a} \quad (3)$$

where $(v_1, \ldots, v_{n-1}) \in [0, 1]^{n-1}$ and $(a_1, \ldots, a_{n-1})$ are the parameters of the Multivariate Beta distribution, with $a_1, \ldots, a_{n-1} > 0$ and $a = \sum_{i=1}^{n-1} a_i$.

Then we define the vector $(\Theta_1, \ldots, \Theta_{n-1})$ according to the *Stick Breaking transformation*, as in the following:

$$\Theta_1 = V_1$$

$$\Theta_2 = V_2(1 - V_1)$$

$$\ldots = \ldots$$

$$\Theta_{n-1} = V_{n-1} \prod_{i=1}^{n-2} (1 - V_i)$$

and finally

$$\Theta_n = 1 - \sum_{i=1}^{n-1} \Theta_i.$$

For Peer Review Only

Note that $\Theta_n = \prod_{i=1}^{n-1}(1 - V_i)$. Since $V_i \in [0, 1]$, this implies that $\sum_{i=1}^{n-1} \Theta_i \leq 1$. About this transformation see also Kotz, Balakrishnan and Johnson (2000). After some algebra, the prior $g(\theta_1, \ldots, \theta_{n-1})$ is:

$$g(\theta_1, \ldots, \theta_{n-1}) \propto \prod_{i=1}^{n-1} \left\{ \frac{\theta_i^{a_i - 1}}{\left(1 - \sum_{j=1}^{i} \theta_j\right)^{a_i + 1}} \left(1 - \sum_{j=1}^{i-1} \theta_j\right) \right\} \cdot \left[1 + \sum_{i=1}^{n-1} \frac{\theta_i}{1 - \sum_{j=1}^{i} \theta_j}\right]^{-a}.$$

At last we determine the form of the posterior distribution $\pi(\theta_1, \ldots, \theta_{n-1}|\underline{x})$ generated by the prior defined above, according to equation (2):

$$\pi(\theta_1, \ldots, \theta_{n-1}|\underline{x}) \propto \left(1 - \sum_{i=1}^{n-1} \theta_i\right)^{x_n} \left[1 + \sum_{i=1}^{n-1} \frac{\theta_i}{1 - \sum_{j=1}^{i} \theta_j}\right]^{-a} \cdot \prod_{i=1}^{n-1} \left\{ \frac{\theta_i^{a_i + x_i - 1}}{\left(1 - \sum_{j=1}^{i} \theta_j\right)^{a_i + 1}} \left(1 - \sum_{j=1}^{i-1} \theta_j\right) \right\}. \quad (4)$$

Equation (4) represents, with the appropriate normalizing constant, the proposed new Multinomial model.

4 Estimation of model parameters and applications

Now our aim is to estimate the parameters of the new Multinomial model.

The posterior distribution of equation (4) is known up to a normalizing constant that is not possible to compute. One of the algorithms that allows to sample from a distribution without the knowledge of the normalizing constant is the *Metropolis-Hastings (M-H) algorithm* (Hastings, 1970). In particular for sampling from the distribution of equation (4) we use a version of the M-H algorithm that is called Independent Metropolis Hastings algorithm. The n-th step of this algorithm works as follow. Suppose that $X_n = x$ and that π is the distribution that we want to sample, known up to a normalizing constant (usually is called *target distribution*). A distribution q from what we are able to sample is chosen (usually is called *proposal distribution*). Hence,

1. A point y is sampled from q
2. The so called *acceptance probability* is computed:

$$A = \min\left\{1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right\}$$

For Peer Review Only

3. With probability A the point y is accepted and $X_{n+1} = y$. Otherwise, with probability $1 - A$ the point y is rejected and $X_{n+1} = x$.

In the case of equation (4) we choose a Dirichlet proposal with parameters $(\alpha_1, \ldots, \alpha_n)$:

$$q(\theta_1, \ldots, \theta_n) \propto \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdot \ldots \cdot \theta_n^{\alpha_n - 1}$$

where a convenient choice in terms of calculations for $(\alpha_1, \ldots, \alpha_n)$ is the following:

$$\begin{aligned} \alpha_1 &= a_1 + x_1 \\ \alpha_2 &= a_2 + x_2 \\ \ldots &= \ldots \\ \alpha_{n-1} &= a_{n-1} + x_{n-1} \\ \alpha_n &= x_n - a_{n-1}. \end{aligned} \tag{5}$$

Note that the determination of the acceptance probability requires rewriting the posterior as a n-dimensional density.

In order to study the behavior of our model compared to the traditional Multinomial-Dirichlet model, we apply it to two toy examples.

1. The first example considers a total of $N = 18$ categorical response data, observed in $n = 4$ possible outcomes, where N is the number of trials and n is the number of categories. Data are denoted by $\underline{x} = (x_1, x_2, x_3, x_4)$, where x_i $(i = 1, \ldots, n)$ represents the number of observations over the N trials that falls into the i-th category. In this example suppose that $\underline{x} = (2, 3, 4, 9)$, meaning that 2 observations of the 18 trials fall into the first category, 3 observations of the 18 trials fall into the second category, and so on. Moreover, let $\underline{a} = (a_1, a_2, a_3, a_4)$ denote the hyperparameters vector of the Multivariate Beta prior distribution (3). Here we assume that the vector $\underline{a}$ takes the values: $\underline{a} = (1, 1, 0, 0)$. Note that (recalling the definition of the Dirichlet proposal (5)) this choice of hyperparameters is suitable for a comparison of the Multinomial-Dirichlet with the new Multinomial model.
2. The second toy example is based on the data provided by Engeman and Swanson (1991) from a 2×2 contingency table. They are mortality data from two anaesthesia techniques for elderly patients receiving emergency hip surgery. Here the categorical response data are $N = 76$, observed in $n = 4$ possible outcomes. Data are denoted by $\underline{x} = (34, 3, 30, 9)$ and the Multivariate Beta hyperparameters vector is denoted by $\underline{a} = (1, 1, 0, 0)$, like in toy example 1.

For Peer Review Only

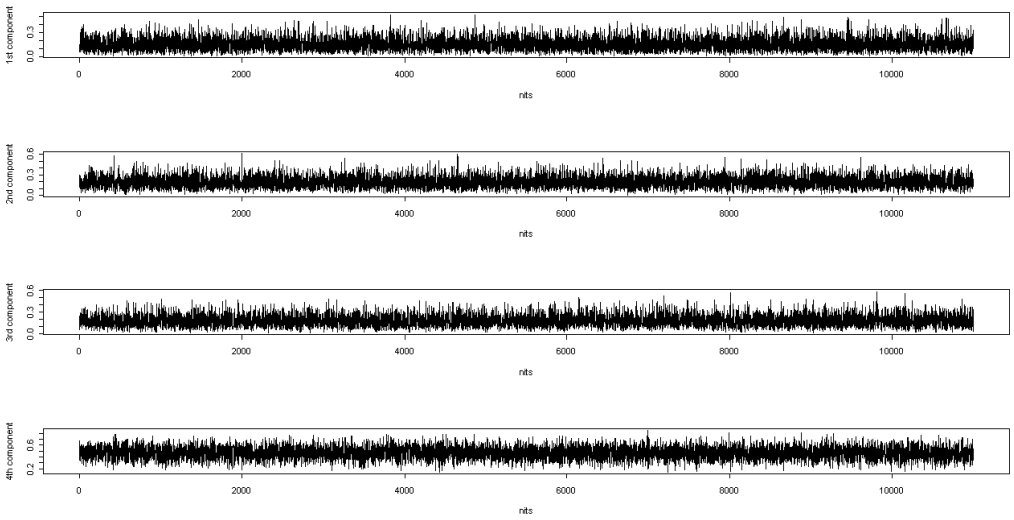

Figure 1: Sample paths for toy example 1

4.1 Simulation results for the first toy example

Figure (1) shows the sample paths of the four parameters $(\theta_1, \ldots, \theta_4)$ from top to bottom. These are the plots of the random variables being generated versus the number of iterations. This plot is helpful not only to understand if the chain can move around the state space, but also to assess the convergence to the target distribution. As we can see from the figure, the chains of all parameters are well mixing, since there are no flat periods and the jumps are not too close. In fact, flat periods indicate that the proposal generates candidate observations that are too often rejected, while very close jumps denote that candidate observations are too often accepted, suggesting in both cases that the proposal is not suitable to describe the target distribution and that there is not convergence to the target itself. However, in our case the chains mix very well, exploring freely the parameter space and being centered in correspondence to the parameter estimate, showing in this way convergence to the target distribution. We run 10000 iterations, after discarding 1000 values as burn-in period, considering in this way only the elements approaching to the chain's stationarity. The acceptance rate is 0.84164, which is satisfactory as stated by Besag et al. (1995), suggesting an acceptance rate between 20% and 80%.

Another test to assess whether the sample has reached its stationary distribution is given by autocorrelations (Gelfand and Smith, 1990). Generally, we expect observa-

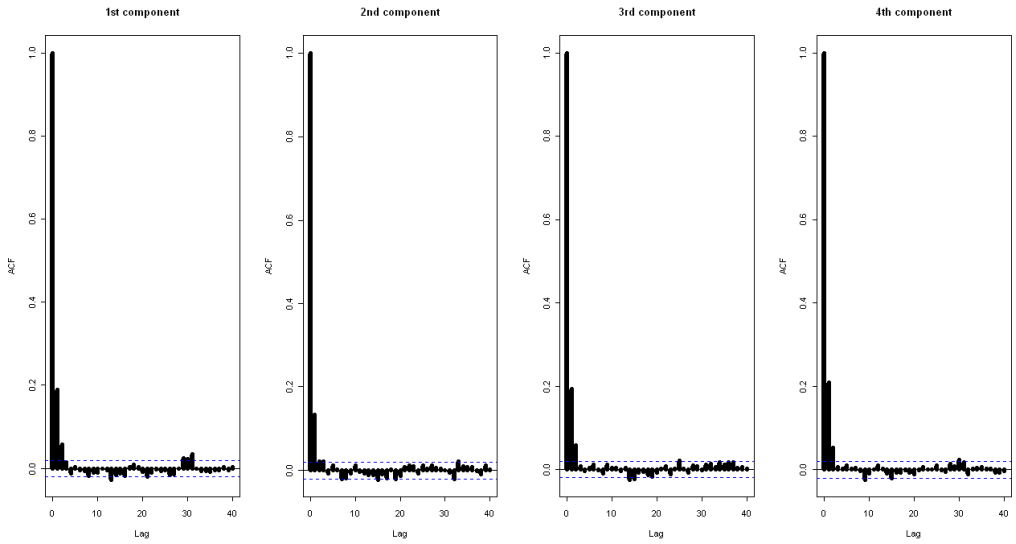

Figure 2: Autocorrelation plots for toy example 1

tions generated through the Metropolis-Hastings algorithm to be positively correlated, but this correlation has not to be too high. We can quantify this correlation by using an autocorrelation function. This helps us to detect lack of convergence in situations where the sample trace appears to be well mixing, but small jumps induce autocorrelations between successive observations. Considering a sequence $\theta^1, \ldots, \theta^n$ of length n, $\rho_k(\theta^i, \theta^{i+k})$ is the k-th order autocorrelation between observations θ^i and θ^{i+k}, where k is the time lag. The k-th order autocorrelation can be estimated by $\hat{\rho}_k = \frac{Cov(\theta^i, \theta^{i+k})}{Var(\theta^i)}$. Autocorrelation plots show the k-th order autocorrelation as a function of the time lag k. Figure (2) shows the autocorrelation plots for the parameters estimated in toy example 1. For all the four parameters $(\theta_1, \ldots, \theta_4)$ autocorrelations is very close to zero indicating stability of convergence to the target distribution.

4.2 Simulation results for the second toy example

In the second toy example, as in the previous case, we run 10000 iterations, after a burn-in period of 1000. In figure (3) the sample paths of the parameters $(\theta_1, \ldots, \theta_4)$ are displayed form top to bottom. The acceptance rate of the algorithm is 0.73273, lying perfectly in Besag's interval $[20\%; 80\%]$ (Besag et al., 1995). Again the picture shows well-mixing chains, that converge to the target distribution.

For Review Only

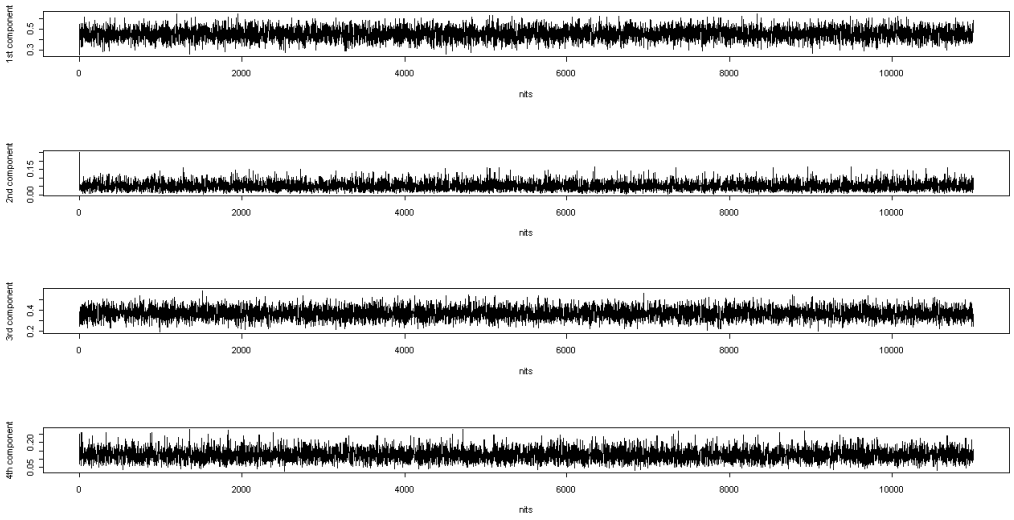

Figure 3: Sample paths for toy example 2

Figure (4) illustrates the autocorrelations for the second example. Results are good and suggest stable convergence.

5 The Zero Variance Principle

Now our purpose is to get more efficient estimates of the model parameters. This problem is particularly important when the number of observations is very low, as in many applications of the Multinomial model. The *Zero Variance Principle* allows us to construct a different estimator with the same mean but with a huge reduction of the variance.

Suppose that $X_1, ..., X_k$ is an i.i.d. sample from a probability distribution π defined on a space $E \subset \mathbb{R}^n$ and let f be a real function defined on E, then, under suitable conditions, an estimator of

$$E_\pi(f) = \int_E f(x)\pi(x)dx$$

is the empirical mean

$$S_n(f) = \frac{1}{k}\sum_{i=1}^{k} f(x). \tag{6}$$

For Review Only

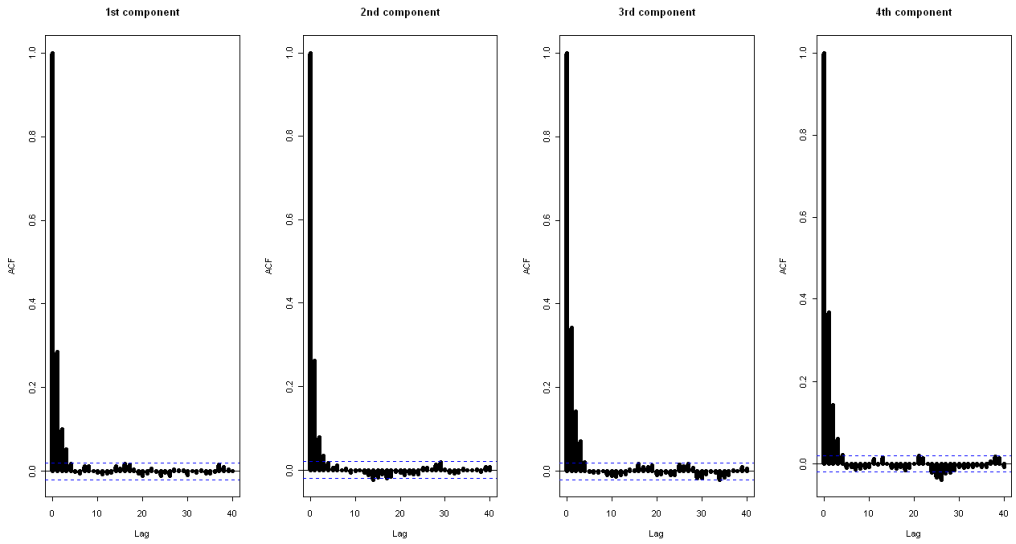

Figure 4: Autocorrelation plots for toy example 2

Assaraf and Caffarel (1999) have introduced a technique to estimate more efficiently moments in Monte Carlo simulations. The idea is to construct an estimator of the moments $S_n(\tilde{f})$ through a function $\tilde{f}$ with the following properties: i) $E_\pi(f) = E_\pi(\tilde{f})$, ii) $\sigma_f^2 \leq \sigma_{\tilde{f}}^2$, where σ_f^2, $\sigma_{\tilde{f}}^2$ are the variances with respect to π, respectively of f and $\tilde{f}$.

Therefore, instead of reducing the estimator variance by changing the updating law of the Markov Chain, Assaraf and Caffarel (1999) suggest to modify the estimator by replacing the function f with $\tilde{f}$. In this way, the expected values of f and $\tilde{f}$ are equal, but the variance of $\tilde{f}$ is always lower.

In order to construct $\tilde{f}$, the Zero Variance principle requires the definition of an operator H and a function ψ. H has to be an hermitian operator and it has to satisfy the following equation:

$$\int H(x,y)\sqrt{\pi(x)}dx = 0.$$

We also assume that ψ is a twice differentiable function.

Then, once defined H and ψ, we construct $\tilde{f}$ in such a way that

$$\tilde{f}(x) = f(x) + \frac{\int H(x,y)\psi(y)dy}{\sqrt{\pi(x)}}. \tag{7}$$

This choice of $\tilde{f}$, as explained above, is such that the function has the properties:

For Review Only

1. $E_\pi(f) = E_\pi(\tilde{f})$
2. $\sigma^2_f \leq \sigma^2_{\tilde{f}}$.

In order to have optimal outcomes, our goal would be to construct the function $\tilde{f}$ in such a way that $\sigma^2_{\tilde{f}}$ is exactly equal to zero. In this case $\tilde{f}$ is equal to a constant which is the expected value of f under π. By setting $\tilde{f} = E_\pi(f)$ into the (7) we obtain

$$\int H(x,y)\psi(y)dy = -\sqrt{\pi(x)}[f(x) - E_\pi(f)]. \tag{8}$$

The solution of the previous equation leads to the definition of the best choice for ψ.

Optimal choices for H differ in the discrete and continuous case. A choice for a discrete space could be:

$$H(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}[P(x,y) - \delta_x(y)]$$

where $P(x,y)$ is the transition matrix of a Markov Chain, reversible with respect to π and $\delta_x(y)$ is the Dirac delta, which is equal to 1 if $x = y$ and zero otherwise.

However, here we focus on a continuous space $E \subset \mathbb{R}^n$ and thus we choose a H operator such that

$$H = -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2} + \frac{1}{2\sqrt{\pi(x)}}\sum_{i=1}^{n}\frac{\partial^2\sqrt{\pi(x)}}{\partial x_i^2}.$$

The last operator satisfies that

$$H(\sqrt{\pi(x)}) = 0$$

and equation (7) can be written in the following way

$$\tilde{f}(x) = f(x) + \frac{H\psi(x)}{\sqrt{\pi(x)}}.$$

Note that, for every function ψ, we have that $\sigma^2_{\tilde{f}} \leq \sigma^2_f$. In particular, if ψ satisfies the equation

$$(H\psi)(x) = \sqrt{\pi}(f(x) - E_\pi(f)), \tag{9}$$

from the (8) in the continuous case, we have that $\sigma^2_{\tilde{f}} = 0$. Clearly, in most of practical application, a "zero variance ψ" is not available, so it is necessary to construct ψ's that ensure a great reduction of variance. To achieve this objective for the applications studied in this paper, in the next section we calculate some zero variance ψ's for the multivariate Dirichlet distribution. The hope is to have some suggestions on the

For Peer Review Only

structure of the ψ's functions. Throughout the paper we will call the zero variance ψ's as "exact ψ's".

Note that in the continuous case, the determination of $(H\psi)(x)$ is a time consuming process, requiring the numerical computation of the first and second order derivatives of the target distribution. However, we can avoid this problem by reducing the evaluation of $(H\psi)(x)$ to the first derivative. If the ψ function has the form

$$\psi(x) = P(x)\sqrt{\pi(x)}$$

where $P(x)$ is a polynomial, then the expression of $(H\psi)(x)$ may be written in the following form

$$(H\psi)(x) = -\frac{1}{2}\sum_{i=1}^{n}\left[\sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right].$$

The previous expression for $(H\psi)(x)$ simplifies the calculations allowing to speed up the simulation.

5.1 Exact ψ for the Multivariate Dirichlet distribution

In this section we calculate the exact ψ's for the *Multivariate Dirichlet* distribution for the functions

- $f_{1i}(\underline{x}) = x_i$
- $f_{2i}(\underline{x}) = x_i^2$

where $i = 1, \ldots, n$ and $\underline{x} = (x_1, \ldots, x_n)$.

The target distribution is, up to a normalizing constant,

$$\pi(\underline{x}) \propto x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

where $\alpha_i > 0$ for all $i = 1, \ldots, n$. If $\alpha = \sum_{i=1}^{n} \alpha_i$, then the ψ functions that solve equation (8) are, respectively for $f_{1i}(\underline{x})$ and $f_{2i}(\underline{x})$,

$$\begin{aligned}
\psi_{1i}(\underline{x}) &= \left(\frac{2}{3}\frac{1}{\alpha_i + 1}x_i^3 - \frac{1}{n-1}\frac{\alpha_i}{\alpha}\sum_{j=1, j\neq i}^{n}\frac{x_j^2}{\alpha_j}\right)\sqrt{\pi(\underline{x})} \\
\psi_{2i}(\underline{x}) &= \left(\frac{1}{2}\frac{1}{\alpha_i + 2}x_i^4 - \frac{1}{n-1}\frac{\alpha_i(1+\alpha_i)}{\alpha(1+\alpha)}\sum_{j=1, j\neq i}^{n}\frac{x_j^2}{\alpha_j}\right)\sqrt{\pi(\underline{x})}.
\end{aligned} \tag{10}$$

For Peer Review Only

6 Application of the Zero Variance Principle

Now, we apply the Zero Variance technique to the two toy examples introduced in section 4. Our aim is to estimate the first moment of each component. The target distribution $\pi(\underline{x})$ is the posterior defined in the new Multinomial Model. Since the zero variance ψ is not available for our model, we use the exact $\psi_{1i}(\underline{x})$ (expressed by the (10)) calculated for the Multivariate Dirichlet distribution in the previous section.

6.1 Simulation results for the first toy example

Table (1) lists the final outcomes for the first toy example we have introduced in section 4.1. The table displays results for the four components $(\theta_1, \ldots, \theta_4)$ on the rows and for the classical Multinomial-Dirichlet model and the new Multinomial model on the columns. For each component the table shows the parameter estimate S_n and its variance $\hat{\sigma}^2$. For the new Multinomial model we computed the estimates S_n and the variances $\widehat{\sigma}^2$ for f and $\tilde{f}$, listing the outcomes in the last two columns respectively. Results show a sensible difference in the estimates of the new Multinomial model com-

		Multinomial-Dirichlet	**New Multinomial Model**	
			f	$\tilde{f}$
θ_1 **first component**	S_n	0.15	0.15488	0.13952
	$\widehat{\sigma}^2$	0.00679	0.006395	5.3655e-05
θ_2 **second component**	S_n	0.2	0.1956	0.20581
	$\widehat{\sigma}^2$	0.00762	0.00717	1.14549e-04
θ_3 **third component**	S_n	0.2	0.17941	0.22121
	$\widehat{\sigma}^2$	0.00762	0.00615	3.54589e-04
θ_4 **fourth component**	S_n	0.45	0.47011	0.41675
	$\widehat{\sigma}^2$	0.01179	0.01124	1.48421e-04

Table 1: Final results for toy example 1

pared to the Multinomial-Dirichlet, thus indicating the different behavior of the new model. Moreover (as illustrated in the last column of table (1)), the application of the Zero Variance principle allows a great reduction of the variance values. The estimates S_n are more precise and they are different from the estimates computed with the function f.

For Peer Review Only

6.2 Simulation results for the second toy example

Table (2) illustrates the final results for the second toy example. The table lists parameter estimates (S_n) and variances ($\hat{\sigma}^2$) for the four components ($\theta_1, \ldots, \theta_4$). The first column shows the outcomes for the traditional Multinomial-Dirichlet model, the second column illustrates the results for the new Multinomial model and the third column shows the resulting values for the new Multinomial model with the application of the Zero Variance principle. The analysis of the outcomes bring us to similar conclusions to the previous example. As it is clear from the last column of table (2), variances $\hat{\sigma}^2_{\tilde{f}}$ fall down to zero and the differences among parameter values indicate that the computation of efficient estimates through the Zero Variance principle is able to show the real value of the parameters.

		Multinomial-Dirichlet	**New Multinomial Model**	
			f	$\tilde{f}$
θ_1	S_n	0.44872	0.45741	0.45029
first component	$\widehat{\sigma}^2$	0.00313	0.00329	2.7171e-05
θ_2	S_n	0.05128	0.05145	0.05268
second component	$\widehat{\sigma}^2$	0.00062	0.00060	5.286e-07
θ_3	S_n	0.38462	0.36250	0.42155
third component	$\widehat{\sigma}^2$	0.00299	0.00285	4.4314e-05
θ_4	S_n	0.11538	0.12864	0.10082
fourth component	$\widehat{\sigma}^2$	0.00129	0.00133	1.12e-05

Table 2: Final results for toy example 2

7 Concluding remarks

We proposed a new Multinomial Model with the specification of an *a priori* distribution which is different from the traditional conjugate prior. We employed the "Stick Breaking transformation" of the Multivariate Beta of Olkin (see Olkin and Liu (2003)) to construct a new prior distribution. We derived the posterior and we implemented the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)) to sample from the posterior distribution. The application of our Multinomial model to toy examples shew excellent results in terms of convergence to the target distribution: chains are well-mixing and autocorrelations are close to zero. Then we applied the Zero Variance

For Peer Review Only

Principle allowing us to estimate the parameters of the distribution with a dramatic reduction of their variances.

The efficient estimation of model parameters is a great advantage in every application, but especially in case of lack of data, where it is easier to obtain less precise estimates.

Moreover, we remark that the differences in the resulting values of parameter estimates denote the different characteristic of our Multinomial model compared the traditional Multinomial-Dirichlet model.

Further work will concern the testing of our model on real data in order to go into more depth study of both its limitations and its potential compared to other models.

8 Acknowledgements

We are grateful to the anonymous referee for the many detailed suggestions, which led to significant improvement in the presentation of the paper.

References

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", Journal of the American Statistical Association, 88, (422), 669–679.

Aitchison, J. and Shen, S.M. (1980), "Logistic-Normal Distributions: Some Properties and Uses", Biometrika, 67, (2), 261–272.

Assaraf, R. and Caffarel, M. (1999), "Zero-Variance principle for Monte Carlo Algorithms", Physical Review letters, 83, (23), 4682–4685.

Balakrishnan, N., Johnson, N. L. and Kotz, S. (1997), "Discrete Multivariate Distributions", John Wiley and Sons.

Berger, J.O. (1985), "Statistical decision theory and Bayesian analysis", Springer.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995), "Bayesian computation and stochastic systems", Statistical Science, 10, 3–66.

Boender, C.G.E. and Rinnooy Kan, A.H.G. (1987), "A Multinomial Bayesian Approach to the Estimation of Population and Vocabulary Size", Biometrika, 74, (4), 849–856.

Brier S.S. (1980), "Analysis of Contingency Tables Under Cluster Sampling", Biometrika, 67, (3), 591–596.

For Peer Review Only

Casella G. and Robert C. P. (1996), "Rao Blackwellization of sampling scheme", Biometrika, 83, (1), 81-94.

Delmas, J. F. and Jourdain, B. (2006), "Does waste recycling really improve Metropolis Hastings Monte Carlo algorithm?", Rapport du Recherce du CERMICS 2006-331.

Engeman, R. M. and Swanson, G. D. (1991), "On Analytical Methods and Inferences for 2×2 Contingency Table Data from Medical Studies", Computers & Biomedical Research, 24, 509–513.

Ferguson, T.S. (1973), "A Bayesian analysis of some nonparametric problems", Ann. Statist., 1, 209–230.

Fine, L. J., Philogene, G. S., Gramling, R., Coups, E. J. and Sinha, S. (2004), "Prevalence of multiple chronic disease risk factors: 2001 National Health Interview Survey", American Journal of Preventive Medicine, 27, (2), 18–24.

Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling based approaches to calculating marginal densities", Journal of the American Statistical Association, 85, 398–409.

Gelman, A, Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004), "Bayesian Data Analysis", Chapman & Hall.

Ghosh, J.K., Delampady, M. and Samanta, T. (2006), "An Introduction to Bayesian Analysis: Theory and Methods", Springer.

Goodhardt,G.J., Ehrenberg, A.S.C. and Chatfield, C. (1984), "The Dirichlet: A Comprehensive Model of Buying Behaviour", Journal of the Royal Statistical Society. Series A, 147, 621–655.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", Biometrika, 57, 97–109.

Hobert, J. P. and Marchev, D. (2007), "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms", Annals of Statistics, 36, (2), 532–554.

Ishwaran H. and James L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors", Journal of the American Statistical Association, 96, (453), 161–173.

Kamakura, W. A., Kim, B-D. and Lee, J. (1996), "Modeling Preference and Structural Heterogeneity in Consumer Choice", Marketing Science, 15, (2), 152–172.

Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000), "Continuous Multivariate Distributions", 2nd Edition, Vol. 1. Wiley, New York.

Metropolis, N., Rosenbluth, A. E., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines", Journal of Chemical Physics, 21, 1087–1092.

For Peer Review Only

Morel, J.G. and Nagaraj, N.K. (1993), "A Finite Mixture Distribution for Modelling Multinomial Extra Variation", Biometrika, 80, (2), 363–371.

Mosimann, J. E. (1962), "On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions", Biometrika, 49, (1/2), 65–82.

Olkin, I. and Liu, R. (2003), "A Bivariate Beta Distribution", Statistics & Probability Letters, 62, 407–412.

Peskun, P. H. (1973), "Optimum Monte Carlo sampling using Markov chains", Biometrika, 60, 607–612.

Rannala, B. and Mountain, J.L. (1997), "Detecting immigration by using multilocus genotypes", Proceedings of the National Academy of Sciences of the United States of America, 94, 9197–9201.

Seaman, S.R. and Richardson S. (2001), "Bayesian Analysis of Case-Control Studies with Categorical Covariates", Biometrika, 88, (4), 1073–1088.

Seaman, S.R. and Richardson S. (2004), "Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies", Biometrika, 91, (1), 15–25.

Walley, P. (1996), "Inferences from Multinomial Data: Learning about a Bag of Marbles", Journal of the Royal Statistical Society. Series B, 58, 3–57.

Wei Y., Liu, B. and Liu, X. (2005), "Entry modes of foreign direct investment in China: a multinomial logit approach", Journal of Business Research, 58, (11), 1495–1505.

For Peer Review Only