Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computational Statistics & Data Analysis 51 (2006) 1243 – 1259

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Simulation-based approach to estimation of latent variable models

Zhiguang Qian, Alexander Shapiro*, [1]

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Received 12 January 2006; received in revised form 16 February 2006; accepted 18 February 2006
Available online 20 March 2006

Abstract

We propose a simulation-based method for calculating maximum likelihood estimators in latent variable models. The proposed method integrates a recently developed sampling strategy, the so-called Sample Average Approximation method, to efficiently compute high quality solutions of the estimation problem. Theoretical and algorithmic properties of the method are discussed. A computational study, involving two numerical examples, is presented to highlight a significant improvement of the proposed approach over existing methods.

Keywords: Simulation-based inference; Latent variable models; Stochastic programming; Sample average approximation

1. Introduction

In recent years a considerable attention was attracted to parametric statistical models in which the probability distributions are defined in terms of multivariate integrals. Since, typically, the involved integrals cannot be written in a closed form, a popular approach to solving the obtained estimation problems is by using Monte Carlo (MC) simulation techniques (e.g., Bhat, 2001; Geyer and Thompson, 1992; Gouriéroux and Monfort, 1996; Lee, 1997). At the same time it was shown theoretically, and verified in numerical experiments, that MC simulation techniques can be surprisingly efficient in solving large scale stochastic programming problems (see Shapiro, 2003; Shapiro and Nemirovski, 2005, and references therein).

In this paper we discuss applications of the methodology and some theoretical results borrowed from stochastic programming to a class of latent variable models. Recent years have witnessed an increasing interest in using latent variable models in a large spectrum of applications, ranging from business/social science to medical science to engineering (cf., Bartholomew and Knott, 1999; Joreskog and Moustaki, 2001; Medel and Kamakura, 2001; Moustaki and Knott, 2000). As a defining feature of such models, latent variables are incorporated for modeling variance, correlation, dependence or other unobserved quantities. Maximum likelihood (ML) method and variants are standard techniques for parameter estimation in these models. The likelihood of such latent variable models are in forms of multivariate integrals, which makes calculating the ML estimators a challenging task.

* Corresponding author.
E-mail addresses: zqian@isye.gatech.edu (Z. Qian), ashapiro@isye.gatech.edu (A. Shapiro).

[1] Research of this author was partly supported by the NSF Grant DMS-0510324.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.02.016

The EM algorithm has been widely used as a solution to mitigate these computational difficulties (cf., Bartholomew and Knott, 1999; Moustaki and Knott, 2000). It starts with approximating the involved integrals by some deterministic methods, like Gaussian quadratures, and then iterates between E and M steps to calculate the estimators based on the approximated integrals. There are two major difficulties associated with an implementation of the EM algorithm for the considered class of problems. First, it is known that numerical approximations for multi-dimensional integrals are quite inaccurate and deterministic methods do not work in high dimensions (see, e.g., Evans and Swartz, 2000). Actually this would be a problem with any numerical procedure for solving such type of problems. Second, the iterative E and M schemes in the EM method do not make much sense from the optimization point of view, and it is known that convergence rates of the EM algorithm are rather slow. When dealing with small optimization problems, the drawback of slow convergence of the EM algorithm could be counterbalanced by simplicity of its implementation. However, for larger problems this slow convergence could be prohibitively expensive.

In this paper we study the ML estimation problem from a *stochastic programming* perspective. Stochastic programming is an emerging and important area in modern optimization that studies mathematical programming problems involving uncertainty. We show that the estimation problem can be framed as a generalized stochastic program. By integrating a sampling methodology, which became known in the area of stochastic programming as the sample average approximation (SAA) method, together with modern optimization techniques, we intend to show that it is possible to solve such type of problems in a considerably faster and more reliable way, and, furthermore, to validate quality of the obtained solutions by estimating errors resulting from the sampling approximations.

Although the proposed method is motivated and developed for latent variable models, similar estimation problems exist in many other areas, such as mixed-logit models in transportation (cf., Bhat, 2001) and generalized linear mixed models (cf., McCulloch and Searle, 2001), to which the proposed methodology can also be applied.

The remainder of the paper will unfold as follows. Section 2 discusses stochastic programming and SAA method. Section 3 introduces binary latent variable models. We discuss approaches to solving the ML problem by MC simulation in Section 4. Convergence analysis of the SAA method is provided in Section 5. In Section 6 we discuss methods for validation of SAA estimators. Section 7 illustrates the proposed method with two numerical examples involving binary latent trait models. A brief summary and some conclusions are outlined in Section 8.

2. Stochastic programming and sample average approximation method

Two-stage stochastic programming (with recourse), as an area in the field of optimization, can be traced back to Beale (1955) and Dantzig (1955). For an overview of recent theoretical and algorithmic developments in this field the interested reader can be referred to Ruszczyński and Shapiro (2003). In an abstract form a stochastic programming problem can be written in the form

$$\operatorname*{Min}_{\boldsymbol{x}\in X}\{f(\boldsymbol{x}) := \mathbb{E}_P[F(\boldsymbol{x},\boldsymbol{\xi})]\}, \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^m$ is a vector of decision variable, $X \subset \mathbb{R}^m$ is a given feasible set and $\boldsymbol{\xi}$ is a random vector having probability distribution P supported on a set $\Xi \subset \mathbb{R}^d$.

Since, typically, the expected value in (2.1) cannot be written in a closed form and, moreover in a multivariate case, with say $d \geqslant 3$, cannot be numerically calculated with a high accuracy, it should be approximated. In that respect one can use MC sampling techniques. Of course, there are various ways how sampling can be incorporated into an optimization procedure. A natural and simple way is the following. First, a random sample $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^S$ of S realizations from P is generated. By replacing $f(\cdot)$ with an approximation based on this sample, problem (2.1) can be approximated by

$$\operatorname*{Min}_{\boldsymbol{x}\in X}\left\{\hat{f}_S(\boldsymbol{x}) := \frac{1}{S}\sum_{j=1}^{S} F\left(\boldsymbol{x},\boldsymbol{\xi}^j\right)\right\}. \tag{2.2}$$

This approach was discovered and rediscovered, under different names and in different fields, by many authors. In the current stochastic programming literature it is known as the SAA method. By the Law of Large Numbers we have that as S tends to infinity, the sample average function $\hat{f}_S(\boldsymbol{x})$ converges w.p.1 to the expected value function $f(\boldsymbol{x})$. However, it is well known that the convergence is notoriously slow. Therefore, it is somewhat surprising that the SAA method turned out to be quite efficient in solving certain classes of stochastic programming problems (cf., Shapiro, 2003; Shapiro and Nemirovski, 2005). In this paper we apply some methodology recently developed in stochastic programming literature to binary latent variable models.

It should be noted that the SAA method is *not* an algorithm. After the random sample is generated, the obtained problem (2.2) should be solved by an appropriate (deterministic) algorithm. We also would like to point out that the SAA method is essentially different, in philosophy and implementations, from the classical Stochastic Approximation (SA) algorithms (see, e.g., Lai, 2003, for a recent survey of the SA, Gu and Zhu, 2001; Zhu and Lee, 2002, for applications of SA algorithms coupled with MC simulation to solving ML estimation problems, and Deylon et al., 1999, for a discussion of a SA version of the EM algorithm). It is possible to show that, for smooth unconstrained optimization problems, the SAA and SA methods converge at the same asymptotic rate, provided that the SA method is implemented with an optimal step size procedure (cf., Shapiro, 1996). This, however, is asymptotics, and in our numerical experience the SAA method coupled with a good (deterministic) algorithm exploiting a particular structure of a considered problem, works much better than the SA method.

3. Binary latent variable models

Although the proposed method may work for general latent trait models, we specifically apply it to binary logit/normal models described as follows. Let $\boldsymbol{x}$ is a p-dimensional vector representing observed variables, which take binary values 0 or 1, and $\boldsymbol{y}$ be an m-dimensional vector of latent variables. The conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$g(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\lambda}) = \prod_{i=1}^{p} \{\pi_i(\boldsymbol{y};\boldsymbol{\lambda})\}^{x_i}\{1-\pi_i(\boldsymbol{y};\boldsymbol{\lambda})\}^{1-x_i}, \tag{3.1}$$

where the logit link is given as

$$\pi_i(\boldsymbol{y};\boldsymbol{\lambda}) := \frac{\exp\left(\lambda_{i0}+\sum_{j=1}^{m}\lambda_{ij}y_j\right)}{1+\exp\left(\lambda_{i0}+\sum_{j=1}^{m}\lambda_{ij}y_j\right)}, \tag{3.2}$$

$\boldsymbol{\lambda} = \left(\lambda_{ij}\right)_{1\leqslant i\leqslant p, 0\leqslant j\leqslant m}$ are unknown parameters. Now the vector of latent variables is viewed as a random vector, denoted by $\boldsymbol{Y}$. It will be assumed that $\boldsymbol{Y} \sim N(0, \boldsymbol{I}_m)$ has the standard normal distribution. This model was introduced by Bartholomew (1987) in an attempt to extend the classical Factor Analysis model to situations where observed variables are binary. For a more recent discussion of this model, its motivation etc., we refer the interested reader to Bartholomew and Knott (1999).

Let us observe that the p-dimensional vector $\boldsymbol{x}$ of binary variables has 2^p different patterns denotes $\boldsymbol{z}_1,\dots,\boldsymbol{z}_{2^p}$. Therefore we can view the above as a multinomial parametric model with probability $\gamma_k = \mathbb{P}(\boldsymbol{x}=\boldsymbol{z}_k)$ of pattern $\boldsymbol{z}_k$, $k=1,\dots,2^p$, parameterized by functions

$$\phi_k(\boldsymbol{\lambda}) := \mathbb{E}[g(\boldsymbol{z}_k|\boldsymbol{Y},\boldsymbol{\lambda})] = \int g(\boldsymbol{z}_k|\boldsymbol{y},\boldsymbol{\lambda})h(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}, \quad k=1,\dots,2^p, \tag{3.3}$$

where $h(\cdot)$ is the pdf of random vector $\boldsymbol{Y}$. That is, we consider the multinomial parametric model

$$\boldsymbol{\gamma} = \boldsymbol{\phi}(\boldsymbol{\lambda}), \tag{3.4}$$

where $\boldsymbol{\gamma} = \left(\gamma_1,\dots,\gamma_{2^p}\right)$ and $\boldsymbol{\phi}(\boldsymbol{\lambda}) = \left(\phi_1(\boldsymbol{\lambda}),\dots,\phi_{2^p}(\boldsymbol{\lambda})\right)$ is a function of the parameter vector $\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}$. The above model can be considered in a general framework of the so-called *moment structures* models (see, e.g., Shapiro, 2005, for a recent survey of a statistical inference of such models). Of course, the difficulty in handling the above parametric model (3.4) is that it is not defined explicitly and involves calculations of multivariate integrals. Note that $\sum_{k=1}^{2^p} g\left(\boldsymbol{z}_k|\boldsymbol{y},\boldsymbol{\lambda}\right) = 1$ for any (fixed) $\boldsymbol{y}$ and $\boldsymbol{\lambda}$, and hence $\sum_{k=1}^{2^p}\phi_k(\boldsymbol{\lambda}) = 1$ as well. Note also that for any $\boldsymbol{z}_k$ and $\boldsymbol{y}$, the function $g(\boldsymbol{z}_k|\boldsymbol{y},\boldsymbol{\lambda})$ is analytic as a function of $\boldsymbol{\lambda}$. It follows then that the integral functions $\phi_k(\boldsymbol{\lambda})$ are also analytic.

Suppose now that we have a random sample $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ of observable variables, and let c_k, $k=1,\dots,2^p$, be the observed frequency of the respective pattern $\boldsymbol{z}_k$, with $\sum_{k=1}^{2^p} c_k = N$. Of course, we have here that $\hat{\gamma}_k = c_k/N$ is an unbiased estimate of the true (population) value of γ_k, $k = 1, \dots, 2^p$. The log-likelihood function, up to a constant independent of $\boldsymbol{\lambda}$, can be written here as follows

$$L(\boldsymbol{\lambda}) = \sum_{k=1}^{2^p} c_k \log \phi_k(\boldsymbol{\lambda}). \tag{3.5}$$

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\lambda}}$ is obtained by maximizing $L(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}$, i.e., by solving the optimization problem

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}} L(\boldsymbol{\lambda}). \tag{3.6}$$

Equivalently, $\hat{\boldsymbol{\lambda}}$ can be obtained as a solution of the minimization problem

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}} F\left(\hat{\boldsymbol{\gamma}}, \boldsymbol{\phi}(\boldsymbol{\lambda})\right), \tag{3.7}$$

where

$$F(\boldsymbol{\gamma}, \boldsymbol{\phi}) := 2 \sum_{\gamma_k \neq 0} \gamma_k \log \gamma_k - 2 \sum_{k=1}^{2^p} \gamma_k \log \phi_k = 2 \sum_{\gamma_k \neq 0} \gamma_k \log \left(\gamma_k / \phi_k\right) \tag{3.8}$$

is called a *discrepancy function* in the terminology of the moment structures analysis. Since the above discrepancy function is associated with the ML method (it corresponds to the likelihood ratio chi-square statistic), we sometimes denote it by $F_{\mathrm{ML}}(\boldsymbol{\gamma}, \boldsymbol{\phi})$.

Observe that for all $\boldsymbol{\gamma}, \boldsymbol{\phi} \in \Theta$, where

$$\Theta := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{2^p} : \sum_{k=1}^{2^p} \theta_k = 1, \ \theta_k > 0, \ k = 1, \dots, 2^p \right\},$$

we have that $F(\boldsymbol{\gamma}, \boldsymbol{\phi}) \geqslant 0$ and that $F(\boldsymbol{\gamma}, \boldsymbol{\phi}) = 0$ iff $\boldsymbol{\gamma} = \boldsymbol{\phi}$. Note also that it follows from the definition of $\boldsymbol{\phi}(\boldsymbol{\lambda})$ that all its components $\phi_k(\boldsymbol{\lambda})$, $k = 1, \dots, 2^p$, are positive for any $\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}$. It is also possible to consider the following discrepancy function

$$F_{\mathrm{GLS}}(\boldsymbol{\gamma}, \boldsymbol{\phi}) := \sum_{k=1}^{2^p} \frac{\left(\gamma_k - \phi_k\right)^2}{\phi_k}. \tag{3.9}$$

This function corresponds to Pearson's chi-square statistic and can be viewed as a generalized least squares discrepancy function. We denote by $\hat{F}_N$ the optimal value of problem (3.7), i.e., $\hat{F}_N = F\left(\hat{\boldsymbol{\gamma}}, \boldsymbol{\phi}\left(\hat{\boldsymbol{\lambda}}\right)\right)$, for a chosen discrepancy function. Unless stated otherwise, we use in the subsequent analysis the ML discrepancy function.

One of the basic theoretical questions related to the model (3.4) is its identifiability, i.e., whether the parameter vector $\boldsymbol{\lambda}$ is defined uniquely by Eq. (3.4). Unfortunately, for $m \geqslant 1$, the answer to this question is negative. In order to see this, let us observe that the model can be reformulated as follows. For a given set $\left(\lambda_{ij}\right)_{1 \leqslant i \leqslant p, 0 \leqslant j \leqslant m}$ of parameters and $\boldsymbol{Y} \sim N\left(0, \boldsymbol{I}_m\right)$ define $\boldsymbol{Z} := \boldsymbol{\mu} + \boldsymbol{\Lambda Y}$, where $\boldsymbol{\mu} := \left(\lambda_{10}, \dots, \lambda_{p0}\right)'$ and $\boldsymbol{\Lambda} := \left(\lambda_{ij}\right)_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant m}$ is the corresponding $p \times m$ matrix. Then $\boldsymbol{Z} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Lambda\Lambda}'\right)$ and the probabilities π_i in (3.2) can be considered as functions of the components of $\boldsymbol{Z}$. Now the distribution of $\boldsymbol{Z}$ does not change if we replace the random vector $\boldsymbol{Y}$ by $\boldsymbol{QY}$, where $\boldsymbol{Q}$ can be any $m \times m$ orthogonal matrix. That is, we have that $\boldsymbol{\phi}(\boldsymbol{\lambda}) = \boldsymbol{\phi}\left(\boldsymbol{\lambda}_*\right)$, where $\boldsymbol{\lambda}_*$ is obtained from $\boldsymbol{\lambda}$ by rotating the corresponding $p \times m$ matrix $\boldsymbol{\Lambda}$ by an $m \times m$ orthogonal matrix $\boldsymbol{Q}$. This is the so-called indeterminacy of the factor analysis model. For $m = 1$ this corresponds simply to changing the sign of vector $\left(\lambda_{11}, \dots, \lambda_{p1}\right)$. By a general theory of moment structures we have that with model (3.4), is associated a positive integer r, called the *characteristic rank* of the model, equal to the rank of the $2^p \times p(m+1)$ Jacobian matrix $\boldsymbol{\Delta}(\boldsymbol{\lambda}) := \partial \boldsymbol{\phi}(\boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}$ for almost every $\boldsymbol{\lambda}$ (cf., Shapiro, 1986). Because of the above indeterminacy of the model, we have here that

$$r \leqslant p(m+1) - m(m-1)/2, \tag{3.10}$$

provided that

$$p(m+1)-m(m-1)/2 \leqslant 2^p-1. \tag{3.11}$$

It is an open question whether actually the equality in (3.10) holds (under condition (3.11)). In the later case the model can be identified, at least locally, by setting $m(m-1)/2$ components of $\left(\lambda_{ij}\right)_{1\leqslant i\leqslant p, 1\leqslant j\leqslant m}$ to zero.

By the Law of Large Numbers we have that $\hat{\gamma}$ converges w.p.1 to its population (true) value γ^* as N tends to infinity. It follows then that the ML estimator $\hat{\boldsymbol{\lambda}}$ converges w.p.1 to $\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^{p(m+1)}} F(\gamma^*, \boldsymbol{\phi}(\boldsymbol{\lambda}))$, provided that this minimizer $\boldsymbol{\lambda}^*$ is unique and $\hat{\boldsymbol{\lambda}}$ stays w.p.1 in a compact subset of $\mathbb{R}^{p(m+1)}$. In particular, if the model is correct (i.e., there exists $\boldsymbol{\lambda}_0 \in \mathbb{R}^{p(m+1)}$ such that $\gamma^* = \boldsymbol{\phi}(\boldsymbol{\lambda}_0)$) and is identified, then $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_0$ and $\hat{\boldsymbol{\lambda}}$ converges w.p.1 to $\boldsymbol{\lambda}_0$. Recall that, for any $\boldsymbol{\lambda}$, all components of $\boldsymbol{\phi}(\boldsymbol{\lambda})$ are positive. Therefore, the assumption that the model is correct implies that all elements of the population vector γ^* are positive.

We also have here that if the model is correct (i.e., under the null hypothesis), then $N\hat{F}_N$ converges in distribution to χ^2_v with the number of degrees of freedom $v = 2^p - 1 - r$ (this holds for both the ML and GLS discrepancy functions). Moreover, under a sequence of local alternatives (i.e., under the assumption of Pitman's parameter drift), distribution of $N\hat{F}_N$ can be approximated by a noncentral chi-square distribution with the same number v of degrees of freedom and the noncentrality parameter $\delta = N\min_{\boldsymbol{\lambda}} F(\gamma^*, \boldsymbol{\phi}(\boldsymbol{\lambda}))$. This suggests a way for testing fit of the model to the observed data. There are several problems, however, with an application of this test. First, the characteristic rank r, and hence the number of degrees of freedom v, is not known. If actually the equality in (3.10) holds, then, of course, we have $v = 2^p - 1 - p(m+1) + m(m+1)/2$. Second, for larger values of p, typically, some of the empirical frequencies $\hat{\gamma}_k$ are very small or even zeros. In such cases the chi-square approximation of the statistic $N\hat{F}_N$ can be problematic (see Maydeu-Olivares and Joe, 2005, for a discussion of that problem and some alternatives for testing 2^p contingency tables). Finally, the minimal value $\hat{F}_N$ of the discrepancy function can be calculated only approximately. In the subsequent analysis we show how the MC sampling error in calculation of $\hat{F}_N$ can be evaluated. If that error, relative to the data sampling error (variability), is small, the corresponding chi-square tests still can be reasonably applied.

It also could be important to test nested models. For example, one can be interested in testing a one-factor against two-factor models. If there are two nested models with respective discrepancy test statistics $N\hat{F}_N^{(1)}$ and $N\hat{F}_N^{(2)}$, then in order to test one model against the other the difference test statistic $N\hat{F}_N^{(2)} - N\hat{F}_N^{(1)}$ can be employed. Under the null hypothesis (that the smaller model is correct), and mild regularity conditions, this difference test statistic has asymptotically a chi-square distribution with $r_2 - r_1$ degrees of freedom, where r_1 and r_2 are characteristic ranks of the respective models.

4. Solving the ML problem by Monte Carlo simulation

One of the main difficulties in solving the ML optimization problem (3.6) is that the functions $\phi_k(\boldsymbol{\lambda})$ are not given explicitly and involve calculations of multivariate integrals. In this section we discuss a numerical approach to solving (3.6) by using MC sampling techniques coupled with modern optimization algorithms. As it was discussed in Section 2, the basic idea of our approach is quite simple, in order to evaluate the expected value function $\phi_k(\boldsymbol{\lambda})$ we generate a random sample $\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^{S_k}$ of S_k realizations of the random vector $\boldsymbol{Y} \sim N(0, \boldsymbol{I}_m)$ and estimate $\phi_k(\boldsymbol{\lambda})$ by the sample average

$$\hat{\phi}_{S_k}(\boldsymbol{\lambda}) := \frac{1}{S_k}\sum_{j=1}^{S_k} g(z_k|\boldsymbol{Y}^j, \boldsymbol{\lambda}), \quad k = 1, \dots, 2^p. \tag{4.1}$$

Then $L(\boldsymbol{\lambda})$ can be approximated by

$$\hat{L}_S(\boldsymbol{\lambda}) := \sum_{k=1}^{2^p} c_k \log \hat{\phi}_{S_k}(\boldsymbol{\lambda}). \tag{4.2}$$

Consequently, the (true) ML problem (3.6) is approximated by the SAA problem

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}} \hat{L}_S(\boldsymbol{\lambda}), \tag{4.3}$$

where $S := (S_1, \dots, S_{2^p})$.

There are several issues that should be worked out for this approach to work. Namely, how to generate and of what size to take the corresponding random samples, and how to evaluate an error of the MC SAA. A simple and straightforward way will be to generate independent of each other i.i.d. random samples of the same size $S_k = S^*$, $k = 1, \dots, 2^p$. However, intuitively it looks more advantageous to put more effort in evaluating those functions $\phi_k(\boldsymbol{\lambda})$ corresponding to larger values of the empirical frequencies c_k. For $\hat{\phi}_k = \hat{\phi}_{S_k}(\boldsymbol{\lambda})$ close to the corresponding "true" probability γ_k^*, we can use the approximation

$$c_k \log \hat{\phi}_k - c_k \log \gamma_k^* \approx \frac{c_k}{\hat{\phi}_k} \left(\hat{\phi}_k - \gamma_k^* \right) \approx N \left(\hat{\phi}_k - \gamma_k^* \right).$$

Now variance of $\hat{\phi}_k$ can be approximated by $\gamma_k^*(1 - \gamma_k^*)/S_k \approx \gamma_k^*/S_k$, for reasonably small values of γ_k^*. This suggests to take the sample size S_k proportional to γ_k^*, i.e., proportional to c_k, $k = 1, \dots, 2^p$. Some of our numerical experiments confirmed that such choice of the sample sizes leads to computational savings for approximately the same level of achieved accuracy. Since for numerical experiments, reported in Section 7, computational time was not an issue, for the sake of simplicity we used the same sample size S^*.

It is also possible to use various variance reduction techniques to enhance convergence of the SAA estimators. It was found theoretically and confirmed in numerical experiments that quasi-MC methods, for generating the sample of realizations of random vector $\boldsymbol{Y}$, could significantly improve accuracy of SAA estimators, especially when dimension m of $\boldsymbol{Y}$ is small. In the present study we use the latin hypercube (LH) method (as compared with the standard MC approach) for generation of the corresponding random sample (see, e.g., Avramidis and Wilson, 1996, for a discussion and comparison of some variance reduction techniques).

5. Convergence analysis of SAA algorithm

In this section, we present a convergence analysis for the SAA program (4.3). Denote by $\hat{v}_S$ and $\hat{\boldsymbol{\lambda}}_S$ the optimal value and an optimal solution of problem (4.3), respectively. We view $\hat{v}_S$ and $\hat{\boldsymbol{\lambda}}_S$ as estimates (approximations) of their counterparts v^* and $\boldsymbol{\lambda}^*$ of the "true" ML problem (3.6). Note that we treat now the ML problem (3.6) as fixed (deterministic) and view (4.3) as its SAA. In that framework the estimates $\hat{v}_S$ and $\hat{\boldsymbol{\lambda}}_S$ depend on the generated MC sample and therefore are treated as random variables.

We have that as the sample sizes $S_1, \dots, S_{2^p}$ tend to ∞, the estimate $\hat{v}_S$ converges w.p.1 to v^*. This holds if $\hat{\boldsymbol{\lambda}}_S$ stays w.p.1 in a compact set $\Theta \subset \mathbb{R}^{p(m+1)}$ and $\hat{L}_S(\boldsymbol{\lambda})$ converges w.p.1 to $L(\boldsymbol{\lambda})$ uniformly in $\boldsymbol{\lambda} \in \Theta$. Moreover, $\hat{\boldsymbol{\lambda}}_S$ converges w.p.1 to $\boldsymbol{\lambda}^*$, provided that the optimal solution $\boldsymbol{\lambda}^*$ of problem (3.6) is unique (may be after imposing identifiability constraints). Let us observe that the difference between the objective functions of SAA (4.3) and the true ML problem (3.6) can be written as follows

$$\hat{L}_S(\boldsymbol{\lambda}) - L(\boldsymbol{\lambda}) = \sum_{k=1}^{2^p} c_k \left(\log \hat{\phi}_{S_k}(\boldsymbol{\lambda}) - \log \phi_k(\boldsymbol{\lambda}) \right). \tag{5.1}$$

It follows that $\hat{L}_S(\boldsymbol{\lambda})$ converges w.p.1 to $L(\boldsymbol{\lambda})$ uniformly in $\boldsymbol{\lambda} \in \Theta$ if each $\log \hat{\phi}_{S_k}(\boldsymbol{\lambda})$ converges w.p.1 to $\log \phi_k(\boldsymbol{\lambda})$, as $S_k \to \infty$, uniformly on the set Θ. This, in turn, holds if each $\hat{\phi}_{S_k}(\boldsymbol{\lambda})$ converges w.p.1 to $\phi_k(\boldsymbol{\lambda})$ uniformly on the set Θ and $\phi_k(\boldsymbol{\lambda})$ is bounded from zero for every $\boldsymbol{\lambda} \in \Theta$. It is well known that under mild regularity conditions such uniform convergence holds on any compact set Θ (see, e.g., Section 2.6 in Rubinstein and Shapiro, 1993). In the present case the required regularity conditions can be easily verified provided that $0 \notin \Theta$.

This type of arguments settles consistency of the estimators $\hat{v}_S$ and $\hat{\boldsymbol{\lambda}}_S$ with increase of the sample sizes. Moreover, it is possible to show that in a sense the convergence is exponentially fast. That is, let Θ be a compact set in $\mathbb{R}^{p(m+1)}$

such that $0 \notin \Theta$. Then for $k = 1, \ldots, 2^p$ and any $\varepsilon > 0$, there exist positive constants $C_k = C_k(\varepsilon)$ and $\beta_k = \beta_k(\varepsilon)$, independent of S_k, such that

$$\Pr\left\{\sup_{\boldsymbol{\lambda}\in\Theta}\left|\hat{\phi}_{S_k}(\boldsymbol{\lambda}) - \phi_{S_k}(\boldsymbol{\lambda})\right| \geqslant \varepsilon\right\} \leqslant C_k \mathrm{e}^{-S_k \beta_k}, \tag{5.2}$$

e.g., Shapiro and Xu (2005). (Recall that the random vector $\boldsymbol{Y}$ is assumed to have a (standard) normal distribution and functions $g(z_k|\boldsymbol{y}, \cdot)$ are continuously differentiable. Therefore, the required regularity conditions hold here.) It follows that $\hat{L}_S(\boldsymbol{\lambda})$ converges exponentially fast to $L(\boldsymbol{\lambda})$ uniformly in $\boldsymbol{\lambda} \in \Theta$. This, in turn, implies an exponentially fast convergence of the optimal value and optimal solutions of the SAA problem to their counterparts of the true ML problem (cf., Shapiro, 2003).

6. Validation of SAA estimators

In the Econometric (and Statistics) literature, statistical inference of MC simulation-based estimators is discussed in an asymptotic sense as the sample sizes tend to infinity (e.g., Gouriéroux and Monfort, 1996; Lee, 1997). One of the advantages of the proposed methodology is the ability to assess the accuracy of calculated solutions for a *given* (generated) MC sample. This is in contrast with existing methods, like the EM algorithm (Bartholomew and Knott, 1999; Moustaki and Knott, 2000), which can only provide an estimator of the model parameters without giving much information regarding the accuracy of the solution of the ML estimation problem. Two different methods for evaluating SAA solutions will be discussed in this paper. One approach is based on constructing statistical upper and lower bounds on the optimal value v^* of the true ML problem (3.6), and also an upper bound for the gap between the optimal value v^* and value $L\left(\bar{\boldsymbol{\lambda}}\right)$ at a considered point $\bar{\boldsymbol{\lambda}}$. The other approach is based on statistical testing of the first-order optimality conditions.

6.1. Estimating the optimal value

Although as it was discussed in Section 5, $\hat{v}_S$ converges w.p.1 to the optimal value v^* of the true problem with increase of the sample sizes, for any finite sample there is an error in estimating v^* by $\hat{v}_S$. We give now a brief description of MC sampling methodology for constructing lower and upper bounds on the true value v^*. This methodology was introduced in Norkin et al. (1998) and Mak et al. (1999).

We have that $\sup_{\boldsymbol{\lambda}'} \hat{L}_S\left(\boldsymbol{\lambda}'\right) \geqslant \hat{L}_S(\boldsymbol{\lambda})$, and hence $\mathbb{E}\left[\hat{v}_S\right] \geqslant \mathbb{E}\left[\hat{L}_S(\boldsymbol{\lambda})\right]$, for any $\boldsymbol{\lambda} \in \mathbb{R}^{p(m+1)}$. It follows that $\mathbb{E}\left[\hat{v}_S\right] \geqslant v_S^0$, where

$$v_S^0 := \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{p(m+1)}} \mathbb{E}\left[\hat{L}_S(\boldsymbol{\lambda})\right]. \tag{6.1}$$

The expected value $\mathbb{E}\left[\hat{v}_S\right]$ can be estimated by solving the corresponding SAA problem (4.3) several, say M, times for independently generated samples of the same size $S = (S_1, \ldots, S_{2^p})$. Let $\hat{v}_S^m$, $m = 1, \ldots, M$, be the calculated optimal values of these SAA problems. We have that

$$\bar{\hat{v}}_{S,M} := \frac{1}{M}\sum_{m=1}^{M} \hat{v}_S^m \tag{6.2}$$

provides an unbiased estimator of $\mathbb{E}\left[\hat{v}_S\right]$, and hence can be used as a statistical upper bound for the optimal value v_S^0. An approximate $(1-\alpha)$-confidence upper bound for $\mathbb{E}\left[\hat{v}_S\right]$, and hence for v_S^0, can be written then as

$$UB_{S,M} := \bar{\hat{v}}_{S,M} + \frac{t_{\alpha,M-1}s_M}{\sqrt{M}}, \tag{6.3}$$

where

$$s_M^2 := \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{v}_S^m - \bar{\hat{v}}_{S,M}\right)^2 \tag{6.4}$$

is the corresponding sample variance. The number M of replications does not need to be large (often 5–7 replications are sufficient to get an idea about variability of $\hat{v}_S$), and therefore we use critical value $t_{\alpha,M-1}$, given by the quantile of the t-distribution, which is more conservative than the corresponding quantile z_α of standard normal distribution.

We have that $\mathbb{E}\left[\hat{\phi}_{S_k}(\boldsymbol{\lambda})\right] = \phi_k(\boldsymbol{\lambda})$ for any fixed $\boldsymbol{\lambda}$. Therefore, by Jensen's inequality (since the function $\log x$ is concave) it follows that $\mathbb{E}\left[\log \hat{\phi}_{S_k}(\boldsymbol{\lambda})\right] < \log \phi_k(\boldsymbol{\lambda})$. Consequently, by (5.1) this implies

$$\mathbb{E}\left[\hat{L}_S(\boldsymbol{\lambda})\right] - L(\boldsymbol{\lambda}) = \sum_{k=1}^{2^p} c_k \left(\mathbb{E}\left[\log \hat{\phi}_{S_k}(\boldsymbol{\lambda})\right] - \log \phi_k(\boldsymbol{\lambda})\right) < 0, \tag{6.5}$$

and hence $v_S^0 < v^*$. Therefore, strictly speaking, the upper bound $UB_{S,M}$, for v_S^0, cannot be used as an upper bound for v^* in a straightforward way. By using the second-order Taylor expansion $\log x \approx \log \mu + \frac{1}{\mu}(x-\mu) - \left(1/(2\mu^2)\right)(x-\mu)^2$, at the point $\mu = \phi_k(\boldsymbol{\lambda})$, we can approximate

$$\mathbb{E}\left[\log \hat{\phi}_{S_k}(\boldsymbol{\lambda})\right] - \log \phi_k(\boldsymbol{\lambda}) \approx -\frac{\operatorname{Var}\left[\hat{\phi}_{S_k}(\boldsymbol{\lambda})\right]}{2\phi_k(\boldsymbol{\lambda})^2}. \tag{6.6}$$

Moreover, for $\hat{\phi}_{S_k}(\boldsymbol{\lambda})$ close to the true probability γ_k^* the corresponding variance can be approximated by γ_k^*/S_k, which implies the approximation

$$v^* \approx v_S^0 + N \sum_{k=1}^{2^p} \frac{1}{2S_k}. \tag{6.7}$$

Now let $\bar{\boldsymbol{\lambda}}$ be a candidate for an optimal solution for the true ML problem. Such a candidate point can be obtained by solving a corresponding SAA problem. Clearly $L\left(\bar{\boldsymbol{\lambda}}\right) \leqslant v^*$ for any $\bar{\boldsymbol{\lambda}}$ and, of course, $L\left(\bar{\boldsymbol{\lambda}}\right) = v^*$ iff $\bar{\boldsymbol{\lambda}}$ is an optimal solution of the true ML problem. Hence, we can obtain a lower bound for v^* by accurately estimating $L\left(\bar{\boldsymbol{\lambda}}\right)$, which can be achieved by using MC sampling with a relatively large sample size (it is feasible here to use a large sample since the procedure does not require to solve an optimization problem). However, because of the inequality $v_S^0 < v^*$, we rather use the following procedure for estimating the gap $v_S^0 - \mathbb{E}\left[\hat{L}_S\left(\bar{\boldsymbol{\lambda}}\right)\right]$, which by (6.7) gives an approximation of the gap $v^* - L\left(\bar{\boldsymbol{\lambda}}\right)$.

The idea is similar to the above considerations and is based on solving the corresponding SAA problem M times for independently generated samples. Let $\hat{v}_S^m$, $m = 1, \ldots, M$, be the calculated optimal values of these SAA problems and $\hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right)$ be the corresponding values of the SAA of the ML function, calculated at the point $\bar{\boldsymbol{\lambda}}$, by using the *same* sample which was used in calculation of the corresponding optimal value $\hat{v}_S^m$. Consider the differences

$$\widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right) := \hat{v}_S^m - \hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right). \tag{6.8}$$

Note that because the same sample is used in solving the mth SAA problem and calculating $\hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right)$, it is always true that $\widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right) \geqslant 0$. Also we have that

$$\mathbb{E}\left[\widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right)\right] = \mathbb{E}\left[\hat{v}_S^m\right] - \mathbb{E}\left[\hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right)\right] \geqslant v_S^0 - \mathbb{E}\left[\hat{L}_S\left(\bar{\boldsymbol{\lambda}}\right)\right]. \tag{6.9}$$

Therefore an approximate $(1-\alpha)$-confidence upper bound for the gap $v_S^0 - \mathbb{E}\left[\hat{L}_S\left(\bar{\boldsymbol{\lambda}}\right)\right]$ can be written as follows:

$$\frac{1}{M}\sum_{m=1}^{M} \widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right) + \frac{t_{\alpha,M-1} S_M}{\sqrt{M}}, \tag{6.10}$$

where S_M^2 is the sample variance of $\widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right), m = 1, \ldots, M$. Note that typically $\hat{v}_S^m$ and $\hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right)$ are positively correlated random variables, and hence the variance of $\widehat{\text{gap}}_S^m\left(\bar{\boldsymbol{\lambda}}\right)$ is smaller (sometimes significantly smaller) than the sum of the variances of $\hat{v}_S^m$ and $\hat{L}_S^m\left(\bar{\boldsymbol{\lambda}}\right)$. This is the classical idea of the common random numbers generation well documented in the MC simulation literature. The above statistical upper bound (6.10) was suggested by Mak et al. (1999).

6.2. Testing optimality conditions

In the preceding section, we discussed one method to assess the quality of an SAA solution by constructing a statistical upper bound for the optimality gap corresponding the considered solution. In this section we discuss statistical testing of the first-order optimality conditions for the true ML problem (3.6). Since the maximization problem here is unconstrained we have that if $\boldsymbol{\lambda}^*$ is an optimal solution of (3.6), then $\nabla L(\boldsymbol{\lambda}^*) = 0$. The idea is to test this condition at a considered candidate point $\bar{\boldsymbol{\lambda}}$ by generating a reasonably large sample and consequently estimating $\nabla L\left(\bar{\boldsymbol{\lambda}}\right)$. The approach of statistical testing of first-order (KKT) optimality conditions was developed in Shapiro and Homem-de-Mello (1998), for a more general situation where the optimization is performed subject to constraints.

We have that

$$\nabla L(\boldsymbol{\lambda}) = \sum_{k=1}^{2^p} \frac{c_k}{\phi_k(\boldsymbol{\lambda})} \nabla \phi_k(\boldsymbol{\lambda}), \tag{6.11}$$

and $\nabla \phi_k(\boldsymbol{\lambda}) = \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} g(z_k|\boldsymbol{Y}, \boldsymbol{\lambda})\right]$, $k = 1, \ldots, 2^p$. By generating independent samples $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^{S'_k}$ of S'_k realizations of $\boldsymbol{Y}$, $k = 1, \ldots, 2^p$, we can estimate $\phi_k\left(\bar{\boldsymbol{\lambda}}\right)$ in accordance with (4.1). Note that since this procedure does not require solution of an optimization problem, one can use a relatively large sample sizes S'_k, $k = 1, \ldots, 2^p$. At the same time we can estimate $\nabla \phi_k\left(\bar{\boldsymbol{\lambda}}\right)$, and hence to estimate $\nabla L\left(\bar{\boldsymbol{\lambda}}\right)$. By repeating this procedure M' times (for independently generated samples) we obtain estimates $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{M'}$ of $\nabla L\left(\bar{\boldsymbol{\lambda}}\right)$. Finally, we estimate $\nabla L\left(\bar{\boldsymbol{\lambda}}\right)$by $\bar{\boldsymbol{G}} := (1/M') \sum_{m=1}^{M'} \boldsymbol{G}_m$ and the covariance matrix of $\bar{\boldsymbol{G}}$ by $\boldsymbol{S}/M'$, where $\boldsymbol{S} := (1/(M'-1)) \sum_{m=1}^{M'} \left(\boldsymbol{G}_m - \bar{\boldsymbol{G}}\right)\left(\boldsymbol{G}_m - \bar{\boldsymbol{G}}\right)'$ is the corresponding sample covariance matrix (see Shapiro and Homem-de-Mello, 1998, for details and a discussion of the corresponding statistical test).

7. Numerical examples

Our computation in the following sections are implemented to run in an IBM PC with Intel (R) Pentium (R) 4 CPU 2.66 GHz and Microsoft Windows 2000 operational system. We use the nonlinear optimization solver Systems Optimization Laboratory at Stanford University, MINOS (2005) in GAMS (2005) for numerical calculations and use R (2005) for MC and LH sampling. MINOS solves nonlinear optimization problems using a *reduced-gradient* algorithm (Murtagh and Saunders, 1978) combined with a *quasi-Newton* algorithm that is described in Wolfe (1962).

Calculation times of the following examples: for the problem of Section 7.1 it took, for each SAA iteration, 5–15 s in R to generate MC or LH samples with $S^* = 10$, 50, 250, 1250, and 15–40 s in GAMS to solve the corresponding SAA optimization problem. For the example of Section 7.2 it took 10–30 s to generate MC samples with $S^* = 10$, 50, 250, 500 and 40–90 s in GAMS to solve the corresponding SAA optimization problem.

7.1. Law school admission test example

In this section we analyze the law school admission test example from Bartholomew and Knott (1999), which has five binary manifest variables. Table 4.2 on Page 96 of Bartholomew and Knott (1999) presents response patterns and observed frequencies for these manifest variables. Following the analysis in Bartholomew and Knott (1999), we fit a binary latent variable model with one factor to the data of this example. Parameters of this model that need to be estimated are $\boldsymbol{\lambda} = \left(\lambda_{ij}\right)_{1 \leqslant i \leqslant 5, 0 \leqslant j \leqslant 1}$. As discussed in Section 3, indeterminacy of this one-factor model is simple, it only involves change of sign of vector $(\lambda_{11}, \ldots, \lambda_{51})'$. In actual numerical calculations we need to restrict the set of permissible values of $\boldsymbol{\lambda}$ to a bounded subset of $\mathbb{R}^{p(m+1)}$. For the ease of computation, we impose the following (box) constraints $-10 \leqslant \lambda_{ij} \leqslant 10$ on every component of $\boldsymbol{\lambda}$.

Frequencies of all patterns except Patterns 9 and 15 for this example are nonzero. The computational procedures for evaluating bounds on the optimal value and optimality gaps are described in Section 6.1. We generate independent of each other random samples of the same sizes for the 30 patterns with nonzero frequencies in computing SAA estimators. MC method and LH method are used for selecting the samples. The use of the LH method produces significant improvements over the MC sampling in this problem in terms of lower bounds on $\mathbb{E}\left[\hat{v}_S\right]$, the upper bounds on $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$ as well as the

Table 1
Summary of upper bounds on $\mathbb{E}\left[\hat{v}_S\right]$, lower bounds on $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$, and upper bounds on optimality gaps for SAA estimators using LH in the law school admission test example

S^*	m	$\hat{v}_S^m$	$\mathbb{E}\left(\hat{v}_S\right)$ (95% confidence upper bound)	$L\left(\hat{\lambda}_S^m\right)$ (97.5% confidence lower bound)	Best	Optimality gap (95% confidence upper bound)
10	1	−2451.89		−2494.55 − 6.14	*	
	2	−2459.36		−2490.30 − 15.21		
	3	−2461.20		−2475.35 − 7.68		
	4	−2470.00		−2490.21 − 9.28		
	5	−2452.80		−2488.36 − 14.97		
	6	−2450.92		−2475.78 − 7.29		
	7	−2462.71		−2475.98 − 8.41		
	8	−2458.96		−2464.97 − 13.02		
	9	−2452.60		−2467.45 − 12.26		
	10	−2454.92		−2476.34 − 8.17		
			−2457.54 + 3.50			13.87 + 8.23
50	1	−2466.50		−2473.44 − 7.56	*	
	2	−2466.05		−2464.88 − 7.18		
	3	−2465.91		−2464.91 − 7.23		
	4	−2466.23		−2473.92 − 7.42		
	5	−2469.55		−2464.90 − 7.27		
	6	−2466.39		−2473.43 − 7.59		
	7	−2465.38		−2473.51 − 7.52		
	8	−2469.62		−2465.00 − 7.28		
	9	−2465.83		−2464.84 − 7.19		
	10	−2463.47		−2473.34 − 7.46		
			−2466.49 + 1.07			0.95 + 0.85
250	1	−2466.79		−2473.44 − 7.62	*	
	2	−2466.53		−2464.76 − 7.26		
	3	−2466.40		−2464.76 − 7.26		
	4	−2466.76		−2464.76 − 7.27		
	5	−2466.95		−2473.48 − 7.64		
	6	−2466.62		−2473.47 − 7.64		
	7	−2466.07		−2464.78 − 7.24		
	8	−2466.48		−2473.48 − 7.65		
	9	−2466.59		−2473.46 − 7.64		
	10	−2466.65		−2473.49 − 7.64		
			−2466.59 + 0.14			0.01 + 0.11
1250	1	−2466.60		−2464.76 − 7.26	*	
	2	−2466.68		−2473.45 − 7.63		
	3	−2466.71		−2464.76 − 7.27		
	4	−2466.57		−2464.76 − 7.27		
	5	−2466.65		−2473.45 − 7.63		
	6	−2466.62		−2464.76 − 7.27		
	7	−2466.55		−2473.46 − 7.62		
	8	−2466.76		−2464.76 − 7.27		
	9	−2466.68		−2464.76 − 7.27		
	10	−2466.69		−2464.76 − 7.27		
			−2466.65 + 0.04			0.01 + 0.02

upper bounds on optimality gaps. This should be not surprising since we deal here with just *one* random variable. For brevity, we only report the results of the LH method here. Table 1 summarizes upper bounds on $\mathbb{E}\left[\hat{v}_S\right]$, lower bounds on $L(\cdot)$ and upper bounds on optimality gaps for the resulting SAA estimators using the LH method. For upper bounds, we calculate values of $\hat{v}_S^m$ for $m = 1, \ldots, 10$, and (equal) sample sizes $S^* = 10, 50, 250, 1250$. The 95% confidence upper bound on $\mathbb{E}\left[\hat{v}_S\right]$, given in (6.3), for each value of S^* is tabulated in this table. For the computed maximizers

Table 2
Values of the SAA and EM estimators for the law school admission test example

Parameter	λ_{10}	λ_{20}	λ_{30}	λ_{40}	λ_{50}	λ_{11}	λ_{21}	λ_{31}	λ_{41}	λ_{51}
SAA estimator	2.77	0.99	0.25	1.28	2.05	0.82	0.72	0.89	0.69	0.66
EM estimator	2.75	0.99	0.24	1.27	2.09	0.83	0.72	0.89	0.69	0.66

Table 3
The 97.5% confidence lower bounds on L at the SAA and EM solutions for the law school admission test example

	97.5% confidence lower bound on L
SAA solution	$-2464.76 - 7.26$
EM solution	$-2464.91 - 7.29$

$\hat{\boldsymbol{\lambda}}_S^m$ in each of the upper-bound trials, we estimate $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$by sampling $T = 20$ batches of independent of each other i.i.d. MC random samples of the same size $\bar{S}^* = 100$, thereby obtaining a 97.5% confidence lower bound $LB^m_{\bar{S},T}$ on $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$. In the table, we mark the *best* one among the ten SAA estimators computed using random samples of size S^* that produce the *highest* value of $LB^m_{\bar{S},T}$. The last column of the table gives estimated optimality gaps, of the form given in (6.10), for each of the four best SAA estimators with $S^* = 10, 50, 250, 1250$ respectively.

Several observations emerging from Table 1 are discussed below. The estimated gap for the best SAA estimator using the LH method with $S^* = 1250$ is nearly zero, indicating convergence of the SAA method at this point. This observed convergence could be further confirmed by calculating the distance matrix of the ten SAA estimators using LH with $S^* = 1250$. That is, we take SAA estimators $\hat{\boldsymbol{\lambda}}_S^m$, $m = 1, \ldots, 10$, using LH with $S^* = 1250$, and compute pairwise distances between these solutions. Because of the model indeterminacy, discussed at the beginning of this section, absolute (instead of original) values of the solutions are used in the calculations. For this problem, elements in the resulting distance matrix for ten SAA estimators using LH with $S^* = 1250$ are all close to zero, suggesting that the SAA estimators may be converging to a unique value.

Now we compare quality of the SAA and EM methods for this example. We choose the best $\hat{\boldsymbol{\lambda}}_S^m$ using LH with $S^* = 1250$ identified in Table 1 as the SAA solution used for comparison. A description of the EM algorithm for binary latent variable model can be found in Bartholomew and Knott (1999). The EM solution for this example is provided in Bartholomew and Knott (1999). Table 2 presents the SAA and EM estimators. This table indicates that the EM solution is quite close to the SAA solution; the Euclidean distance between them is 0.05. Similarity of these solutions may be due to the fact that the likelihood of the model for the current example is in a form of one-dimensional integral. The Gauss–Hermite method, underlying the EM approach, may produce comparable result with the LH method, underlying the SAA approach, in approximating integrals of one-dimension. And similar approximation results may therefore lead to similar solutions. However, the example to be given in Section 7.2 shall clearly demonstrate the significant advantages of the proposed SAA method over the EM method in estimation for models with more than one factors. In addition, the KKT test results to be given later in this section shall show that the SAA solution has much better optimality than the EM solution.

For the comparison purpose, we use the same MC procedure to estimate the 95% confidence lower bound on $L(\cdot)$ at the EM solution as we computed the lower bounds in Table 1. Moreover, the same random samples used for estimating the lower bound with the SAA solution are reused here to compute the lower bound with the EM solution. Table 3 presents the resulting lower bound with the EM solution in addition to the lower bound with the SAA solution taken from Table 1. This table indicates that the lower bound with the SAA solution is slightly better than that with the EM solution.

We now report on the optimality test discussed in Section 6.2, applied to the SAA and EM estimators. We use MC sampling with sizes $M = 50, 500, 1000, 2000$ to estimate values of gradients of $L(\cdot)$ at these solutions. Results are shown for different values of M in Tables 4 and 5 for the two solutions, respectively. The last columns in these tables give values, denoted $\hat{\delta}$, of the estimated Euclidean norm of the gradient of $L(\cdot)$ at the considered point. Note

Table 4
Estimated values of gradient ∇L and KKT discrepancy δ at the SAA solution in the law school admission test example

M	i	$j=0$	$j=1$	$\hat{\delta}$
50	1	−1.33	−1.56	
	2	−1.27	−3.70	
	3	−0.30	19.82	
	4	−0.41	1.13	
	5	−0.66	0.49	
				20.36
500	1	−0.74	1.38	
	2	−2.08	4.22	
	3	−3.23	2.07	
	4	−0.78	3.84	
	5	−0.48	−0.22	
				7.41
1000	1	−0.22	0.10	
	2	0.47	−0.86	
	3	1.20	−2.63	
	4	1.23	2.75	
	5	0.53	−0.66	
				4.37
5000	1	−0.07	0.23	
	2	−0.00	−0.75	
	3	0.07	−0.45	
	4	0.93	−0.18	
	5	0.50	0.49	
				1.49

that the estimated values of $\nabla L(\cdot)$ at the SAA solution are significantly better than those at the EM solution, clearly demonstrating the advantages the proposed SAA method.

7.2. *Workplace industrial relations example*

In this section we consider the workplace industrial relations example from Bartholomew and Knott (1999). For this example, response patterns and their frequencies for six binary manifest variables are available at http://www.arnoldpublishers.com/support/lvmfa2.htm. We fit a binary latent variable model with two factors, suggested by Bartholomew and Knott (1999), to the data for this example. This model involves unknown parameters $\boldsymbol{\lambda} = \left(\lambda_{ij}\right)_{1\leqslant i\leqslant 6, 0\leqslant j\leqslant 2}$. Following the discussion in Section 3, we fix λ_{12}, the upper triangular part of $\boldsymbol{\lambda}$, at 0 to obtain an identifiable model. A box constraint $[-10, 10]$ is imposed on each of the remaining parameters of $\boldsymbol{\lambda}$ that need to be estimated. The same procedure used in the example in Section 7.1 is applied to the present example with some omission of details.

For this example, 58 patterns have nonzero frequencies and Patterns 8, 9, 23, 24, 49 and 57 have zero frequency. As we did in the example in Section 7.1, we generate independent of each other random samples of the same sizes for the 58 patterns with non-zero frequencies to compute SAA estimators. Summaries of upper bounds on $\mathbb{E}\left[\hat{v}_S\right]$, lower bounds on $L(\cdot)$ and upper bounds on optimality gaps for the resulting SAA estimators are given in Table 6. The MC sampling is used for selecting samples for this table. The upper bounds in the table are calculated based on values of $\hat{v}_S^m$ for $m = 1, \ldots, 10$ and $S^* = 10, 50, 250, 500$. This table shows the 95% confidence upper bound on $\mathbb{E}\left[\hat{v}_S\right]$, given in (6.3), for each value of S^*. Lower bound estimates are also provided in table. We estimate $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$ for the computed maximizers $\hat{\boldsymbol{\lambda}}_S^m$ in each of the upper-bound trials by sampling $T = 20$ batches of independent of each other i.i.d. MC random samples of the same size $\bar{S}^* = 100$, thereby obtaining a 97.5% confidence lower bound $LB^m_{\bar{S},T}$ on

Table 5
Estimated values of gradient ∇L and KKT discrepancy δ at the EM solution in the law school admission test example

M	i	$j=0$	$j=1$	$\hat{\delta}$
50	1	0.33	−2.69	
	2	−1.42	−3.70	
	3	1.64	19.73	
	4	1.15	0.64	
	5	−4.99	2.27	
				21.14
500	1	0.94	0.16	
	2	−2.24	4.24	
	3	−1.34	1.95	
	4	0.75	3.34	
	5	−4.80	1.63	
				8.18
1000	1	1.48	−1.10	
	2	0.31	−0.84	
	3	3.10	−2.79	
	4	2.78	2.24	
	5	−3.82	1.22	
				7.10
5000	1	1.63	−1.00	
	2	−0.16	−0.72	
	3	1.97	−0.61	
	4	2.47	−0.69	
	5	−3.85	2.39	
				5.96

$L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$. In the table, we mark the *best* one that produces the *highest* value of $LB^m_{\bar{S},T}$ among the ten SAA estimators $\hat{\boldsymbol{\lambda}}_S^m$, $m=1,\ldots,10$, computed using random sample of size S^*, as we did in the previous example. Estimated optimality gaps are presented in the last column of the table in the form of (6.10) for each of the identified best SAA estimators with $S^*=10, 50, 250, 500$.

Some observations made from Table 6 are now in order. The lower bounds on $\mathbb{E}\left[\hat{v}_S\right]$, upper bounds on $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$ and upper bounds on optimality gaps tend to decrease as the value of S^* increases. The estimated gap for the best SAA estimator with $S^*=500$ is quite small in comparison with the function values of $L(\cdot)$, which is of the order from -3400 to -3300, indicating convergence of this SAA estimator. Optimality gaps of SAA estimators for this example can be enhanced by using some variance reduction techniques, like the LH sampling.

Taking the SAA estimators $\hat{\boldsymbol{\lambda}}_S^m$, $m=1,\ldots,10$, with $S^*=500$, we compute pairwise distances between these solutions. Table 7 presents the resulting distance matrix. This table indicates that these solutions differ considerably from each other, suggesting that they may not converge to a unique estimator for the current two-factor model. On the other hand, Table 6 shows that values of $\hat{v}_S^m$ at these solutions are close to each other and the coefficient of variation for these function values is of the order of 1% only. These observations indicate that the likelihood function for the present model may be quite flat resulting in a numerical instability of optimal solutions.

Next we compare results of the SAA and EM methods, applied to this example. We choose the best $\hat{\boldsymbol{\lambda}}_S^m$, with $S^*=500$, marked in Table 6 for the SAA solution to be used in comparison. The EM solution for this example is given by Bartholomew and Knott (1999). Tables 8 and 9 present the SAA and EM estimators, respectively. These tables indicate that the EM solution differs considerably from the SAA solution, resulting the Euclidean distance of 13.18 between them.

Table 10 presents the estimated 95% confidence lower bounds on $L(\cdot)$ at the SAA and EM solutions. Unlike the preceding example, this table shows that for this example the lower bound with the SAA solution is significantly better

Table 6
Summary of upper bounds on $\mathbb{E}\left[\hat{v}_S\right]$, lower bounds on $L\left(\hat{\boldsymbol{\lambda}}_S^m\right)$ and upper bounds on optimality gaps for SAA estimators in the workplace industrial relations example

S^*	m	$\hat{v}_S^m$	$\mathbb{E}\left[\hat{v}_S\right]$ (95% confidence upper bound)	$L\left(\hat{\lambda}_S^m\right)$ (97.5% confidence lower bound)	Best	Optimality gap (95% confidence upper bound)
10	1	−3177.12		−3462.21 − 7.24	*	
	2	−3377.77		−3390.02 − 9.91		
	3	−3263.02		−3395.03 − 11.13		
	4	−3203.74		−3439.89 − 11.24		
	5	−3265.23		−3522.69 − 13.96		
	6	−3221.55		−3385.31 − 12.95		
	7	−3271.96		−3435.23 − 7.08		
	8	−3204.81		−3394.43 − 11.13		
	9	−3297.64		−3464.88 − 10.28		
	10	−3175.80		−3566.97 − 14.56		
			−3245.87 + 105.32			303.84 + 122.81
50	1	−3310.98		−3431.76 − 8.17	*	
	2	−3159.26		−3364.49 − 13.07		
	3	−3306.31		−3358.89 − 9.58		
	4	−3260.51		−3352.44 − 12.08		
	5	−3321.26		−3364.80 − 10.83		
	6	−3328.40		−3420.05 − 6.58		
	7	−3354.46		−3403.47 − 12.25		
	8	−3311.75		−3434.62 − 5.34		
	9	−3249.09		−3400.38 − 6.75		
	10	−3364.75		−3424.04 − 10.28		
			−3296.68 + 101.40			86.78 + 69.94
250	1	−3310.04		−3381.14 − 5.28	*	
	2	−3320.46		−3368.97 − 11.59		
	3	−3314.31		−3354.98 − 10.03		
	4	−3319.89		−3358.12 − 9.63		
	5	−3320.11		−3355.30 − 14.11		
	6	−3346.81		−3367.39 − 8.95		
	7	−3347.48		−3354.44 − 9.29		
	8	−3334.59		−3345.44 − 11.00		
	9	−3348.65		−3395.97 − 11.80		
	10	−3279.51		−3346.29 − 11.90		
			−3324.18 + 36.01			50.11 + 53.93
500	1	−3324.96		−3368.16 − 7.61	*	
	2	−3334.16		−3354.78 − 13.78		
	3	−3308.08		−3349.53 − 13.04		
	4	−3336.26		−3359.33 − 9.57		
	5	−3350.59		−3357.25 − 8.77		
	6	−3337.67		−3359.87 − 8.39		
	7	−3313.64		−3353.42 − 6.60		
	8	−3341.19		−3372.65 − 11.64		
	9	−3313.27		−3352.57 − 6.32		
	10	−3337.77		−3361.95 − 9.61		
			−3329.76 + 23.69			36.33 + 33.59

than that with the EM solution, clearly demonstrating an advantage of the proposed SAA method over the EM method. The superiority of the SAA method shall be further supported by the result of optimality tests below.

Now we report results on the optimality test described in Section 6.2, applied to the SAA and EM estimators. Tables 11 and 12 summarize estimated values of $\nabla L(\cdot)$ at these solutions using MC sampling of size $M = 10\,000$. These tables also show calculated (Euclidean) norm of $\nabla L(\cdot)$, denoted $\hat{\delta}$. In comparison, the estimated values of $\nabla L(\cdot)$

Table 7
Distance matrix for ten SAA estimators with $S^* = 500$ for the workplace industrial relations example

	1	2	3	4	5	6	7	8	9	10
1	0.00	22.77	20.40	19.75	16.48	15.62	8.30	16.37	8.63	18.46
2	22.77	0.00	20.41	17.52	25.91	24.26	19.86	30.48	19.78	21.95
3	20.40	20.41	0.00	26.66	15.85	13.15	17.16	23.94	17.19	8.71
4	19.75	17.52	26.66	0.00	20.32	20.75	17.72	22.27	17.61	22.44
5	16.48	25.91	15.85	20.32	0.00	3.25	14.19	12.70	14.22	8.91
6	15.62	24.26	13.15	20.75	3.25	0.00	13.36	14.08	13.41	7.23
7	8.30	19.86	17.16	17.72	14.19	13.36	0.00	18.99	0.46	14.62
8	16.37	30.48	23.94	22.27	12.70	14.08	18.99	0.00	19.19	19.02
9	8.63	19.78	17.19	17.61	14.22	13.41	0.46	19.19	0.00	14.59
10	18.46	21.95	8.71	22.44	8.91	7.23	14.62	19.02	14.59	0.00

Table 8
Values of the SAA solution for the workplace industrial relations example

λ_{ij}	$j = 0$	$j = 1$	$j = 2$
$i = 1$	−0.71	1.26	0
$i = 2$	3.85	−10	−8.30
$i = 3$	1.39	0.85	1.52
$i = 4$	−1.38	−0.09	−1.09
$i = 5$	−0.92	0.82	−1.76
$i = 6$	−2.42	0.78	−1.33

Table 9
Values of the EM solution for the workplace industrial relations example

λ_{ij}	$j = 0$	$j = 1$	$j = 2$
$i = 1$	−0.93	0.97	2.13
$i = 2$	0.54	1.51	−0.96
$i = 3$	−1.40	1.31	1.11
$i = 4$	−1.47	1.22	0.12
$i = 5$	−0.97	1.58	1.24
$i = 6$	−2.39	1.05	1.06

Table 10
The 97.5% confidence lower bounds on L at the SAA and EM solutions for the workplace industrial relations example

	97.5% confidence lower bound on L
SAA solution	−3349.53 −13.04
EM solution	−3637.985 −7.082

at the SAA solution are consistently smaller than those at the EM solution. Moreover, the value of estimated KKT discrepancy $\hat{\delta}$ at the SAA solution is 9.02, which is only 9.7% of the estimated discrepancy at the EM solution (92.54). These results further demonstrate the advantage of the proposed SAA method over the EM method in the present example.

8. Summary and conclusions

In this paper, we have developed a simulation-based method for calculating maximum likelihood estimators in latent variable models. The proposed method makes use of a sampling strategy recently developed in stochastic

Table 11
Estimated values of gradient ∇L and KKT discrepancy δ at the SAA solution using $M = 10\,000$ in the workplace industrial relations example. Gradient for λ_{12} is not given because this parameter is fixed at 0 in calculation

$M = 10\,000$	$j = 0$	$j = 1$	$j = 2$	$\hat{\delta}$
$i = 1$	2.43	−1.81	*	
$i = 2$	−1.39	−0.53	5.84	
$i = 3$	−1.89	1.13	−0.71	
$i = 4$	−0.02	0.60	0.46	
$i = 5$	−1.19	0.83	−1.05	
$i = 6$	−2.20	3.86	−2.72	
				9.02

Table 12
Estimated values of gradient ∇L and KKT discrepancy δ at the EM solution with $M = 10\,000$ in the workplace industrial relations example

$M = 10\,000$	$j = 0$	$j = 1$	$j = 2$	$\hat{\delta}$
$i = 1$	13.51	−19.86	50.32	
$i = 2$	−14.62	−47.51	77.28	
$i = 3$	3.18	6.70	−16.73	
$i = 4$	0.09	−9.02	19.14	
$i = 5$	2.98	9.99	−23.12	
$i = 6$	1.53	5.71	−18.40	
				92.54

programming, namely the Sample Average Approximation (SAA) method, to efficiently compute the solutions for the estimation problem. Algorithmic and theoretical issues related to the convergence analysis of SAA method and validation of the SAA estimators have been discussed. The advantages of the proposed simulation-based method over the existing EM method are demonstrated with two examples from the latent variable literature. Furthermore, aided by the use of Latin Hypercube method, the SAA estimators in the first example are found to converge to a unique solution. Although, in this paper the proposed method is applied to latent variable models, the method is rather general and can be potentially applied to other statistical problems where likelihoods are modeled in a form of high-dimensional integrals.

Acknowledgement

We are grateful to Dr. Shabbir Ahmed for his suggestions and assistance in computing.

References

Avramidis, A.N., Wilson, J.R., 1996. Integrated variance reduction strategies for simulation. Oper. Res. 44, 327–346.

Bartholomew, D.J., 1987. Latent variable models and factor analysis. Griffin's Statistical Monographs and Courses, London.

Bartholomew, D.J., Knott, M., 1999. Latent Variable Models and Factor Analysis. Arnold, New York.

Beale, E.M.L., 1955. On minimizing a convex function subject to linear inequalities. J. Roy. Statist. Soc. Ser. B 17, 173–184.

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Res. 35B, 677–693.

Dantzig, G.B., 1955. Linear programming under uncertainty. Management Sci. 1, 197–206.

Deylon, B., Lavielle, E., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27, 94–128.

Evans, M., Swartz, T., 2000. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford.

GAMS Development Corporation, 2005. GAMS: A General Algebraic Modeling System, Version 2.5, available at ⟨http://www.gams.com/, 2005⟩.

Geyer, C.J., Thompson, E.A., 1992. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). J. Roy. Statist. Soc. Ser. B 54, 657–699.

Gouriéroux, C., Monfort, A., 1996. Simulation-Based Econometric Methods. Oxford University Press, New York.

Gu, M.-gao., Zhu, H.T., 2001. Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. J. Roy. Statist. Soc. Ser. B 63, 339–355.

Joreskog, K., Moustaki, I., 2001. Factor analysis of ordinal variables: a comparsion of three approaches. Multivariate Behavioral Res. 36, 347–387.

Lai, T.L., 2003. Stochastic approximation: invited paper. Ann. Statist. 31, 391–406.

Lee, L.F., 1997. Simulated maximum likelihood estimation of dynamic discrete choice statistical models, Some Monte Carlo results. J. Econometrics 82, 1–35.

Mak, W.K., Morton, D.P., Wood, R.K., 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Oper. Res. Lett. 24, 47–56.

Maydeu-Olivares, A., Joe, H., 2005. Limited and full-information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. J. Amer. Statist. Assoc. 100, 1009–1020.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. Wiley, New York.

Medel, M., Kamakura, W., 2001. Factor analysis with (mixed) observed and latent variables in the exponential family. Psychometrika 66, 515–530.

Moustaki, I., Knott, M., 2000. Generalized latent trait models. Psychometrika 65, 391–411.

Murtagh, B.A., Saunders, M.A., 1978. Large-scale linearly constrined optimization. Math. Programming 14, 41–72.

Norkin, V.I., Pflug, G.C., Ruszczyński, A., 1998. A branch and bound method for stochastic global optimization. Math. Programming 83, 425–450.

R Development Team, 2005. R: A Project for Statistical Computing, Version 2.0, Available at ⟨http://www.r-project.org/, 2005⟩.

Rubinstein, R.Y., Shapiro, A., 1993. Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. Wiley, New York.

Ruszczyński, A., Shapiro, A. (Eds.), 2003. Stochastic programming. Handbook in OR & MS, vol. 10. North-Holland Publishing Company, Amsterdam.

Shapiro, A., 1986. Asymptotic theory of overparametrized structural models. J. Amer. Statist. Assoc. 81, 142–149.

Shapiro, A., 1996. Simulation based optimization—convergence analysis and statistical inference. Stochastic Models 12, 425–454.

Shapiro, A., 2003. Monte Carlo sampling methods. In: Rusczyński, A., Shapiro, A. (Eds.), Stochastic Programming, Handbooks in OR & MS, vol. 10. North-Holland Publishing Company, Amsterdam, pp. 353–425.

Shapiro, A., 2005. Statistical inference of moment structures. In: Sik-Yum Lee (Ed.), Handbook on Structural Equation Models. Elsevier, Amsterdam. to be published.

Shapiro, A., Homem-de-Mello, T., 1998. A simulation-based approach to two-stage stochastic programming with recourse. Math. Programming 81, 301–325.

Shapiro, A., Nemirovski, A., 2005. On complexity of stochastic programming problems. In: Jeyakumar, V., Rubinov, A.M. (Eds.), Continuous Optimization. Springer, New York, pp. 111–144.

Shapiro, A., Xu, H., 2005. Stochastic mathematical programs with equilibrium constraints, modeling and sample average approximation E-print available at: ⟨http://www.optimization-online.org, 2005⟩.

Systems Optimization Laboratory at Stanford University, 2005. MINOS: A Solver for Large-Scale Non-Linear Optimization Problems.

Wolfe, P., 1962. The reduced-gradient method, Unpublished Manuscript. RAND Corporation.

Zhu, H.T., Lee, S.Y., 2002. Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte Carlo method. Statist. Comput. 12, 175–183.