MANAGEMENT SCIENCE
Vol. 55, No. 3, March 2009, pp. 453–469
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5503 | 0453
DOI 10.1287/mnsc.1080.0951

Persistency Model and Its Applications in Choice Modeling

Karthik Natarajan
Department of Mathematics, National University of Singapore, Singapore 117543 and Singapore-MIT Alliance, matkbn@nus.edu.sg

Miao Song
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, msong@mit.edu

Chung-Piaw Teo
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore 117591, bizteocp@nus.edu.sg

Given a discrete maximization problem with a linear objective function where the coefficients are chosen randomly from a distribution, we would like to evaluate the expected optimal value and the marginal distribution of the optimal solution. We call this the persistency problem for a discrete optimization problem under uncertain objective, and the marginal probability mass function of the optimal solution is named the persistence value. In general, this is a difficult problem to solve, even if the distribution of the objective coefficient is well specified. In this paper, we solve a subclass of this problem when the distribution is assumed to belong to the class of distributions defined by given marginal distributions, or given marginal moment conditions. Under this model, we show that the persistency problem maximizing the expected objective value over the set of distributions can be solved via a concave maximization model. The persistency model solved using this formulation can be used to obtain important qualitative insights to the behavior of stochastic discrete optimization problems. We demonstrate how the approach can be used to obtain insights to problems in discrete choice modeling. Using a set of survey data from a transport choice modeling study, we calibrate the random utility model with choice probabilities obtained from the persistency model. Numerical results suggest that our persistency model is capable of obtaining estimates that perform as well, if not better, than classical methods, such as logit and cross-nested logit models. We can also use the persistency model to obtain choice probability estimates for more complex choice problems. We illustrate this on a stochastic knapsack problem, which is essentially a discrete choice problem under budget constraint.

Key words: probability distribution; integer programming; utility preference; choice functions
History: Accepted by Dorit Hochbaum, optimization and modeling; received September 28, 2006. This paper was with the authors 10 months for 2 revisions. Published online in *Articles in Advance* December 17, 2008.

1. Introduction

Consider a discrete optimization problem:

$$Z(\tilde{\mathbf{c}}) = \max\{\tilde{\mathbf{c}}'\mathbf{x}\colon \mathbf{x} \in \mathscr{X}\}, \tag{1}$$

where the feasible region $\mathscr{X}$ is given as

$$\mathscr{X} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{b},\ x_i \in \mathscr{X}_i \subseteq \mathscr{Z}^+\ \forall\, i \in \mathscr{N}\}. \tag{2}$$

The decision variables x_i are indexed in $i \in \mathscr{N} = \{1, 2, \ldots, n\}$, where the set $\mathscr{X}_i$ consists of nonnegative integer values from α_i to β_i that x_i can take

$$\mathscr{X}_i = \{\alpha_i, \alpha_i + 1, \ldots, \beta_i - 1, \beta_i\}. \tag{3}$$

For a given distribution θ of the objective coefficients, the expected value of the discrete optimization problem in (1) can be expressed as

$$E_\theta(Z(\tilde{\mathbf{c}})) = E_\theta\left(\sum_{i\in\mathscr{N}} \tilde{c}_i x_i^*(\tilde{\mathbf{c}})\right),$$

where $x_i^*(\tilde{\mathbf{c}})$ is an optimal value for the x_i decision variable for the objective $\tilde{\mathbf{c}}$. When $\tilde{\mathbf{c}}$ is random, $x_i^*(\tilde{\mathbf{c}})$ is a random variable. For ease of exposition, we will assume that the set of $\tilde{\mathbf{c}}$ such that $Z(\tilde{\mathbf{c}})$ has multiple optimal solutions has a support with measure zero. We can then rewrite the expected value as

$$E_\theta(Z(\tilde{\mathbf{c}})) = \sum_{i\in\mathscr{N}} \sum_{k\in\mathscr{X}_i} k E_\theta(\tilde{c}_i \mid x_i^*(\tilde{\mathbf{c}}) = k) P_\theta(x_i^*(\tilde{\mathbf{c}}) = k).$$

In this paper, we are interested in finding the value $E_\theta(Z(\tilde{\mathbf{c}}))$ and in particular $P_\theta(x_i^*(\tilde{\mathbf{c}}) = k)$, where the latter is called the *persistence value* of x_i at the value k.[1]

[1] In an earlier work, Adams et al. (1998) studied the question of to what extent the solution to a linear programming relaxation can be used to fix the value of 0-1 discrete optimization problems. They termed this the persistency problem of the 0-1 programming model. The motivation of their work, however, is different from ours.

We call this the persistency problem for discrete optimization under uncertain objective function. The persistence values correspond to the concept of choice probabilities in the discrete choice literature (see §4 for more details) and to the criticality indices in the project management literature (see Bertsimas et al. 2006).

For many fixed choices of the distribution θ (e.g., $\tilde{c}_i$s are independently and uniformly generated in $[0,1]$), there is by now a huge literature on finding approximations and bounds to the expected value of a stochastic discrete optimization problem (cf. Bertsimas et al. 2004, Lyons et al. 1999, Lovász 2001, and Steele 1997). However, finding precise persistence values for stochastic discrete optimization problems appears to be even harder, because we now need probabilistic information on the support of the optimal solutions. We propose instead an approach to compute $E_{\theta^*}(Z(\tilde{\mathbf{c}}))$ and $P_{\theta^*}(x_i^*(\tilde{\mathbf{c}})=k)$, where

$$\theta^* \in \arg\max_{\theta\in\Theta} E_\theta(Z(\tilde{\mathbf{c}}))$$

and Θ is the set of joint distributions with a prescribed set of information on the marginal distributions. In this regard, θ^* can be viewed as an "optimistic" solution in the set Θ, as it is the distribution that attains the largest expected objective value. We guard against overoptimism by prescribing the marginal distributions (termed as marginal distribution model or MDM), or a finite set of marginal moments (marginal moment model or MMM), as boundary conditions.

2. Problem Formulation

In this section, we discuss the model for the set of distributions Θ. The basic model assumes that the distribution function $F_i(c)$ is known for each objective coefficient $\tilde{c}_i$. No assumptions on independence of the marginal distributions are made. We term this as the MDM. The earliest study of combinatorial optimization problems under this model was carried out by Meilijson and Nadas (1979) and Nadas (1979) for the stochastic project management problem. Their main results show that the problem of finding the tight upper bound on the expected project duration can be obtained by solving a convex optimization problem. Weiss (1986) generalized this result to network flow and reliability problems. Bertsimas et al. (2006) considered 0-1 optimization problems under partial moment information on each objective coefficient and initiated the study of the persistency problem. In a similar spirit, the central problem that we focus on is

$$Z^* = \max_{\theta\in\Theta} E_\theta(\max\{\tilde{\mathbf{c}}'\mathbf{x}\colon \mathbf{x}\in\mathscr{X}\}), \tag{4}$$

where $\mathscr{X}$ is the feasible region of a general discrete optimization problem, and we evaluate the persistence of variables under this model.

Because our interest lies in general discrete optimization problems, a formulation that is particularly useful is the full binary reformulation of integer variables. We can represent any variable x_i in the set $\mathscr{X}_i$ as

$$x_i = \sum_{k\in\mathscr{X}_i} k\hat{y}_{ik},$$
$$1 = \sum_{k\in\mathscr{X}_i} \hat{y}_{ik},$$
$$\hat{y}_{ik}\in\{0,1\} \quad \forall\, k\in\mathscr{X}_i.$$

Defining $\hat{\mathbf{y}} = (\hat{y}_{ik})_{k\in\mathscr{X}_i,\, i\in\mathscr{N}}$, the feasible region $\mathscr{X}$ for the discrete optimization problem can be transformed to $\mathscr{Y}$ using the full binary expansion:

$$\mathscr{Y} = \left\{\hat{\mathbf{y}} \;\middle|\; \mathbf{A}\left(\sum_{k\in\mathscr{X}_i} k\hat{y}_{ik}\right)_{i\in\mathscr{N}} \le \mathbf{b},\ \sum_{k\in\mathscr{X}_i}\hat{y}_{ik} = 1\ \forall\, i\in\mathscr{N},\ \hat{y}_{ik}\in\{0,1\}\ \forall\, k\in\mathscr{X}_i\ \forall\, i\in\mathscr{N}\right\}.$$

There is a unique, one-to-one correspondence between the extreme points of $\mathscr{X}$ and $\mathscr{Y}$; that is $x_i = k$ if and only if $\hat{y}_{ik} = 1$. Such binary reformulations have also been used in Sherali and Adams (1999), though in a different context.

We expand on the earlier work by Meilijson and Nadas (1979), Weiss (1986), and Bertsimas et al. (2006) in three ways. First, we generalize the persistency results to arbitrary discrete optimization problems. In particular, the decision variables can assume more than two values, which allows for more complicated problems to be modeled. The key idea is to map the variables in the original space $\mathscr{X}$ into the higher dimensional space $\mathscr{Y}$ using a binary expansion, so that we can impose restrictions such that the expanded set of variables assumes only two possible values. For the formulation to be tight, however, we require an explicit characterization of the convex hull of the space $\mathscr{Y}$.

Second, the work of Bertsimas et al. (2006) focuses on the case when the information given is marginal moments. Our work generalizes the approach to complete marginal distribution information. In this setting, our work can be viewed as a dual approach to the results of Meilijson and Nadas (1979) in the 0-1 case. More importantly, instead of invoking duality results from infinite-dimensional optimization, we obtain the results through a more direct constructive approach. We also indicate techniques to incorporate information on the shape of the distributions into our formulation.

Third, we investigate the use of the persistency model on a variety of new applications. For a discrete choice model, we use the persistency model to perform both estimation and prediction using a real-life transportation data set. The predicted choice probabilities are comparable to logit and cross-nested logit models under weaker assumptions. In the case of choice modeling under budget constraints, the problem reduces to a stochastic knapsack problem. Through numerical simulations, we show that in this instance the persistence values are close to the choice probabilities found under some standard distributions, including the normal, uniform, two-point, and exponential distributions.

3. Main Results

3.1. Marginal Distribution Model

Assume that each objective coefficient $\tilde{c}_i$ in the discrete optimization problem is a continuously distributed random variable with marginal distribution function $F_i(c)$ and marginal density function $f_i(c)$: i.e., $\tilde{c}_i \sim f_i(c)$. Let Θ denote the set of multivariate distributions θ for the objective coefficients with the given marginal distributions.

Define the following sets of nonnegative variables:

$$y_{ik}(c) = P(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{c}_i = c) \quad \text{and}$$
$$y_{ik} = P(x_i^*(\tilde{\mathbf{c}}) = k) \quad \forall k \in \mathscr{X}_i \; \forall i \in \mathscr{N}.$$

Using conditional expectations, the objective function in (4) can be expressed as

$$\begin{aligned} E(Z(\tilde{\mathbf{c}})) &= E\left(\sum_{i\in\mathscr{N}} \tilde{c}_i x_i^*(\tilde{\mathbf{c}})\right) = \sum_{i\in\mathscr{N}} \int c E(x_i^*(\tilde{\mathbf{c}}) \mid \tilde{c}_i = c) f_i(c)\,dc \\ &= \sum_{i\in\mathscr{N}} \sum_{k\in\mathscr{X}_i} \int c E(x_i^*(\tilde{\mathbf{c}}) \mid \tilde{c}_i = c, x_i^*(\tilde{\mathbf{c}}) = k) \\ &\quad \cdot P(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{c}_i = c) f_i(c)\,dc \\ &= \sum_{i\in\mathscr{N}} \sum_{k\in\mathscr{X}_i} \int kc P(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{c}_i = c) f_i(c)\,dc \\ &= \sum_{i\in\mathscr{N}} \sum_{k\in\mathscr{X}_i} k \int c y_{ik}(c) f_i(c)\,dc. \end{aligned}$$

By the definition of the variables as conditional expectations, we have

$$y_{ik} = \int y_{ik}(c) f_i(c)\,dc \quad \forall k \in \mathscr{X}_i \; \forall i \in \mathscr{N}.$$

Furthermore, for each realization c, the optimal value for x_i is one of the values in $\mathscr{X}_i$, implying

$$\sum_{k\in\mathscr{X}_i} y_{ik}(c) = 1 \quad \forall c \; \forall i \in \mathscr{N}.$$

Last, because the decision variables lie in the convex hull of $\mathscr{Y}$ (denoted as $CH(\mathscr{Y})$) for all realizations, taking expectations we have

$$\mathbf{y} = (y_{ik})_{k\in\mathscr{X}_i,\, i\in\mathscr{N}} = (P(x_i^*(\tilde{\mathbf{c}}) = k))_{k\in\mathscr{X}_i,\, i\in\mathscr{N}} \in CH(\mathscr{Y}).$$

It is clear now that we can obtain an upper bound on (4) by solving

$$\begin{aligned} Z^* \le \max \quad & \sum_{i\in\mathscr{N}} \sum_{k\in\mathscr{X}_i} k \int c y_{ik}(c) f_i(c)\,dc \\ \text{s.t.} \quad & \int y_{ik}(c) f_i(c)\,dc = y_{ik} \quad \forall k \in \mathscr{X}_i \; \forall i \in \mathscr{N} \\ & \sum_{k\in\mathscr{X}_i} y_{ik}(c) = 1 \quad \forall c \; \forall i \in \mathscr{N} \\ & y_{ik}(c) \ge 0 \quad \forall c \; \forall k \in \mathscr{X}_i \; \forall i \in \mathscr{N} \\ & \mathbf{y} \in CH(\mathscr{Y}). \end{aligned} \tag{5}$$

For a given set of values of $\mathbf{y} \in CH(\mathscr{Y})$, the upper bound in (5) is separable across i. The ith subproblem that we need to solve is

$$\begin{aligned} \max \quad & \sum_{k\in\mathscr{X}_i} k \int c y_{ik}(c) f_i(c)\,dc \\ \text{s.t.} \quad & \int y_{ik}(c) f_i(c)\,dc = y_{ik} \quad \forall k \in \mathscr{X}_i \\ & \sum_{k\in\mathscr{X}_i} y_{ik}(c) = 1 \quad \forall c \\ & y_{ik}(c) \ge 0 \quad \forall c \; \forall k \in \mathscr{X}_i. \end{aligned}$$

By relabeling the points in $\mathscr{X}_i$ if necessary, we may without loss of generality assume that $y_{ik} > 0$ for all $k \in \mathscr{X}_i$. The optimal values for the variables $y_{ik}(c)$ in the subproblem can be found using a greedy argument. Introducing the Lagrange multipliers λ_{ik} for the first set of constraints, we obtain

$$\begin{aligned} Z(\lambda) = \max \quad & \int \sum_{k\in\mathscr{X}_i} (ck - \lambda_{ik}) y_{ik}(c) f_i(c)\,dc + \sum_{k\in\mathscr{X}_i} \lambda_{ik} y_{ik} \\ \text{s.t.} \quad & \sum_{k\in\mathscr{X}_i} y_{ik}(c) = 1 \quad \forall c \\ & y_{ik}(c) \ge 0 \quad \forall c \; \forall k \in \mathscr{X}_i. \end{aligned}$$

Thus, for a given value of c, we need to solve a linear program with a single budget constraint and nonnegativity restrictions. Given a set of values in increasing order $\mathscr{X}_i = \{\alpha_i, \ldots, \beta_i\}$, the optimal solution is given by $y_{ik}(c) = \mathbb{I}(ck - \lambda_{ik} \ge cj - \lambda_{ij} \; \forall j \in \mathscr{X}_i)$, where $\mathbb{I}$ is the indicator function.

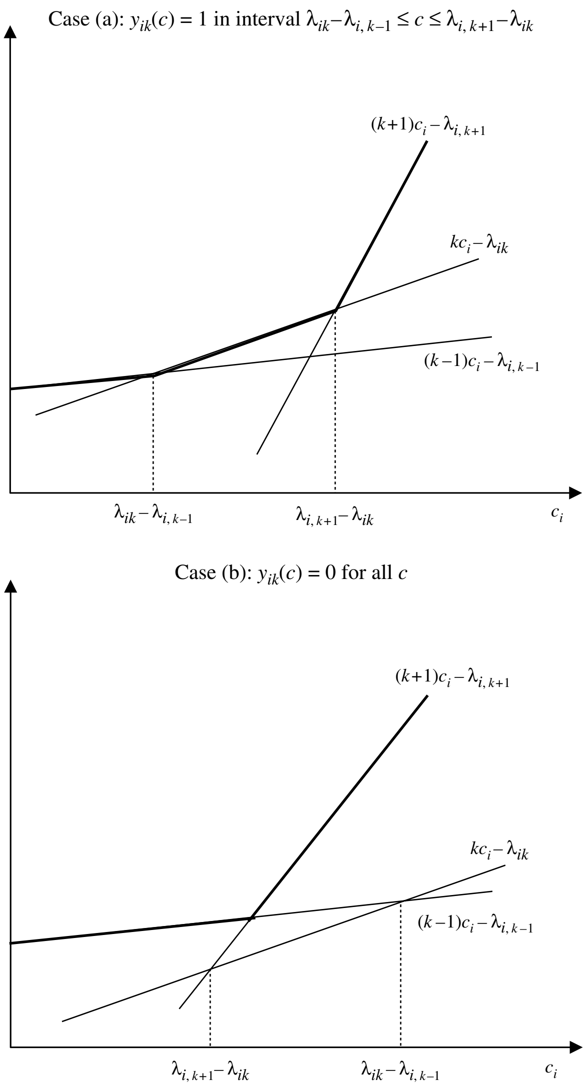
Let $\lambda^* = (\lambda_{ik}^*)$ denote the optimal Lagrange multipliers. In the solution to $Z(\lambda^*)$, our assumption that $y_{ik} > 0$ ensures that for all k, $y_{ik}(c) > 0$ for some c. Hence, we must have

$$y_{ik}(c) = \mathbb{I}(\lambda_{ik}^* - \lambda_{i,k-1}^* \le c \le \lambda_{i,k+1}^* - \lambda_{ik}^*) \quad \forall k \in \mathscr{X}_i,$$

where $\lambda_{i,\alpha_i-1}^* = -\infty$, $\lambda_{i,\beta_i+1}^* = \infty$ (see Figure 1(a)). The Lagrange multipliers are chosen as

$$y_{ik} = F_i(\lambda_{i,k+1}^* - \lambda_{ik}^*) - F_i(\lambda_{ik}^* - \lambda_{i,k-1}^*) \quad \forall k \in \mathscr{X}_i,$$

Figure 1 Optimal Values for Variables $y_{ik}(c)$



with $\sum_{k\in\mathscr{X}_i} y_{ik} = 1$. Note that it is not possible for $\lambda^*_{ik} - \lambda^*_{i,k-1} \ge \lambda^*_{i,k+1} - \lambda^*_{ik}$ wherein $y_{ik}(c) = 0$ for all c and $y_{ik} = 0$ as indicated in Figure 1(b). The values of the multipliers λ^*_{ik} that satisfy these set of equations are given as

$$\lambda^*_{ik} - \lambda^*_{i,k-1} = F_i^{-1}\left(\sum_{j\le k-1} y_k\right).$$

Hence, the upper bound on Z^* in (5) reduces to solving the problem

$$Z^* \le \max\left\{\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i}\left(k\int_{F_i^{-1}(\sum_{j\le k-1} y_{ij})}^{F_i^{-1}(\sum_{j\le k} y_{ij})} c f_i(c)\,dc\right):\ \mathbf{y}\in CH(\mathscr{Y})\right\}.$$

Using the substitution $t = F_i(c)$ and $dt = f_i(c)dc$, we can rewrite the upper bound as

$$Z^* \le \max\left\{\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i}\left(k\int_{\sum_{j\le k-1} y_{ij}}^{\sum_{j\le k} y_{ij}} F_i^{-1}(t)\,dt\right):\ \mathbf{y}\in CH(\mathscr{Y})\right\}.$$

In fact, this bound is tight under the MDM.

Theorem 1. *Under the MDM, Z^* and the persistence values are computed by solving the concave maximization problem*:

$$Z^* = \max\left\{\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i}\left(k\int_{\sum_{j\le k-1} y_{ij}}^{\sum_{j\le k} y_{ij}} F_i^{-1}(t)\,dt\right):\ \mathbf{y}\in CH(\mathscr{Y})\right\}. \quad (6)$$

Proof. We use a randomized rounding argument to construct the probability distribution attaining the bound in the optimal solution. For details, we refer the readers to the appendix. □

Theorem 1 indicates that the difficulty of solving the persistency problem for discrete optimization problems is related to characterizing the convex hull of the extreme points in the binary reformulation. We compare this with a compact relaxation that uses the convex hull of $\mathscr{X}$ directly:

$$Z_0^* = \max\left\{\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i}\left(k\int_{\sum_{j\le k-1} y_{ij}}^{\sum_{j\le k} y_{ij}} F_i^{-1}(t)\,dt\right):\ \left(\sum_{k\in\mathscr{X}_i} k y_{ik}\right)_{i\in\mathscr{N}} \in \mathscr{CH}(\mathscr{X})\right\}. \quad (7)$$

Clearly, $Z^* \le Z_0^*$. Note that instead of the space $\mathscr{CH}(\mathscr{Y})$, which may not be readily available, we are now solving in the space $\mathscr{CH}(\mathscr{X})$, which is easier to characterize. However, the relaxation is weak with the optimal values for each variable concentrated at the two extreme values α_i and β_i.

Proposition 1. *Under the MDM, the upper bound Z_0^* on Z^* is computed by solving the concave maximization problem*:

$$Z_0^* = \max\left\{\sum_{i\in\mathscr{N}}\left(\alpha_i\int_0^{y_{i\alpha_i}} F_i^{-1}(t)\,dt + \beta_i\int_{1-y_{i\beta_i}}^{1} F_i^{-1}(t)\,dt\right):\ (\alpha_i y_{i\alpha_i} + \beta_i y_{i\beta_i})_{i\in\mathscr{N}} \in \mathscr{CH}(\mathscr{X})\right\}. \quad (8)$$

Proof. See the appendix. □

It is clear that the upper bound in Proposition 1 is tight for 0-1 optimization problems. This result can also be interpreted as the dual of the formulations obtained in Meilijson and Nadas (1979) and Weiss (1986). Our constructive approach, however, appears to be more direct and elegant. For completeness, we provide the equivalent results for minimization problems next. Consider the tight lower bound on the minimization version of a discrete optimization problem under MDM:

$$Z^*_{\min} = \min_{\theta\in\Theta} E_\theta(\min\{\tilde{\mathbf{c}}'\mathbf{x}:\ \mathbf{x}\in\mathscr{X}\}).$$

Lemma 1. *Under the MDM, $Z^*_{\min}$ and the persistence values are computed by solving the convex minimization problem*:

$$Z^*_{\min} = \min\left\{\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i}\left(k\int_{\sum_{j\ge k+1} y_{ij}}^{\sum_{j\ge k} y_{ij}} F_i^{-1}(t)\,dt\right):\ \mathbf{y}\in CH(\mathscr{Y})\right\}.$$

A lower bound on $Z^*_{\min}$ is computed by solving the convex minimization problem:

$$Z^*_{\min} \geq \min\left\{\sum_{i\in\mathcal{N}}\left(\alpha_i \int_{y_{i\beta_i}}^{1} F_i^{-1}(t)\,dt + \beta_i \int_0^{1-y_{i\alpha_i}} F_i^{-1}(t)\,dt\right): (\alpha_i y_{i\alpha_i} + \beta_i y_{i\beta_i})_{i\in\mathcal{N}} \in CH(\mathcal{X})\right\},$$

where the inequality is tight for 0-1 *optimization problems*.

We now show that a direct application of this result with F_i set to exponential distributions leads to a bound for large classes of stochastic discrete optimization problems.

Example 1. Lyons et al. (1999) provided a bound for the stochastic shortest path problem with arc lengths $\tilde{c}_i$ that are independent, exponentially distributed random variables with means μ_i. They proved that the expected length of the shortest path between two nodes is bounded from below by the resistance between the nodes, where the resistance of an edge is μ_i. This bound is obtained by solving the quadratic program

$$E\left(\min\left\{\sum_{i\in\mathcal{N}} \tilde{c}_i x_i : \mathbf{x}\in\mathcal{X}\right\}\right) \geq \min\left\{\sum_{i\in\mathcal{N}} \mu_i x_i^2 : \mathbf{x}\in\mathcal{X}\right\},$$

where $\mathcal{X}$ denotes the dominant of all $s-t$ paths in a graph G (refer to Lovász 2001 for the generalization to log-concave distributions). As our model focuses on finding a "best case" probability distribution over all joint distributions, we can obtain a similar lower bound for this stochastic combinatorial optimization problem. Let $\mathcal{X}$ denote a polytope with 0-1 vertices (the dominant of all $s-t$ path solutions being a special case) and Θ denote the set of all joint distributions such that the marginal distributions for $\tilde{c}_i$ are exponential distributions with means μ_i. No other assumptions, such as independence among the $\tilde{c}_i$s, are made. For any $\theta\in\Theta$, a lower bound on the expected optimal cost of a minimization problem can be obtained by evaluating $Z^*_{\min}$:

$$E_\theta\left(\min\left\{\sum_{i\in\mathcal{N}} \tilde{c}_i x_i : x\in\mathcal{X}\right\}\right) \geq Z^*_{\min}.$$

Applying Lemma 1 with $\alpha_i = 0$, $\beta_i = 1$, $F_i(t) = 1 - e^{-t/\mu_i}$, and $F_i^{-1}(t) = -\mu_i \log(1-t)$, we obtain

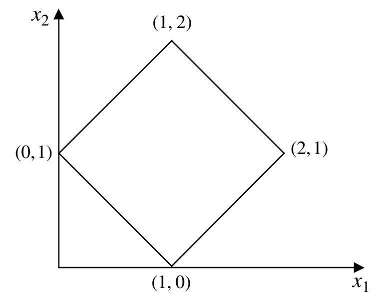
$$Z^*_{\min} = \min\left\{\sum_{i\in\mathcal{N}}\left(\int_0^{x_i} -\mu_i \log(1-t)\,dt\right): \mathbf{x}\in\mathcal{X}\right\},$$

which reduces to

$$Z^*_{\min} = \min\left\{\sum_{i\in\mathcal{N}} \mu_i (x_i + (1-x_i)\log(1-x_i)): \mathbf{x}\in\mathcal{X}\right\}.$$

Although our bound applies to general 0-1 problems, it is nevertheless weaker than the bound proposed in Lyons et al. (1999) because we drop the assumption of independence. Our bounds have the advantage, though, that they can be shown to be tight; i.e., we can find the distribution, with suitable correlations built into the random variables, that approaches the objective bound.

Figure 2 Diamond-Shaped Polyhedron



We next illustrate the bound for a simple integer program where the decision variables can take more than two values.

Example 2. Consider the two-dimensional integer polytope given as

$$\mathcal{X} := \{(x_1, x_2)\in\Re^2 \mid x_1 + x_2 \geq 1,\ -x_1 - x_2 \geq -3,\ x_1 - x_2 \geq -1,\ -x_1 + x_2 \geq -1\}.$$

The four extreme points of this polyhedron (see Figure 2) are $\{(0,1),(1,0),(1,2),(2,1)\}$ with $\mathcal{X}_1 = \{0,1,2\}$ and $\mathcal{X}_2 = \{0,1,2\}$. Assume that both $\tilde{c}_1$ and $\tilde{c}_2$ are independently and uniformly distributed on $[0,1]$. Note that under this model, the extreme points $\{(1,2),(2,1)\}$ have equal chances of attaining the optimal solution under the random objective function. We solve the case with only the marginal distribution conditions imposed on the problem. In this case, the upper bound in Proposition 1 reduces to

$$\begin{aligned} Z_0^* = \max\ & 2 - 0.5y_{10}^2 - 0.5(y_{10}+y_{11})^2 - 0.5y_{20}^2 - 0.5(y_{20}+y_{21})^2 \\ \text{s.t.}\ & (y_{11}+2y_{12}) + (y_{21}+2y_{22}) \geq 1, \\ & -(y_{11}+2y_{12}) - (y_{21}+2y_{22}) \geq -3, \\ & (y_{11}+2y_{12}) - (y_{21}+2y_{22}) \geq -1, \\ & -(y_{11}+2y_{12}) + (y_{21}+2y_{22}) \geq -1, \\ & y_{10}+y_{11}+y_{12} = 1, \quad y_{20}+y_{21}+y_{22} = 1, \\ & y_{10}, y_{11}, y_{12}, y_{20}, y_{21}, y_{22} \geq 0. \end{aligned}$$

This is a concave maximization problem over linear constraints. Solving this with CPLEX Version 9.1, OPL Studio Version 4.1 yields $Z_0^* = 1.875$ with persistence values $(y_{10}, y_{11}, y_{12}, y_{20}, y_{21}, y_{22}) = (0.25, 0, 0.75, 0.25, 0, 0.75)$. We can use the notion of cutting planes to strengthen the above formulation.

One valid equality that is obtained from Figure 2 (or by looking at the $\mathcal{Y}$ polytope), which is not satisfied by the current solution, is $y_{10} + y_{12} + y_{20} + y_{22} = 1$. This is a facet of the $\mathcal{Y}$ polytope and says that in the optimal solution, we must have either $x_1 = 1$ or $x_2 = 1$. This can be expressed as the probabilistic equality: $P(x_1 = 0) + P(x_1 = 2) + P(x_2 = 0) + P(x_2 = 2) = 1$. Solving the quadratic optimization problem with this added equality yields $Z^* = 1.75$, which is the tight bound. The corresponding optimal persistence values are $(y_{10}, y_{11}, y_{12}, y_{20}, y_{21}, y_{22}) = (0, 0.5, 0.5, 0, 0.5, 0.5)$, with the optimal solutions concentrated at $(x_1, x_2) = (1, 2)$ and $(x_1, x_2) = (2, 1)$, respectively. Interestingly, the persistence values obtained from our model is precisely the values under independence.

3.2. Marginal Moment Model

We now relax the assumption on the knowledge of complete marginal distributions to the knowledge of a finite set of marginal moments. For simplicity, we focus on the case when only the mean μ_i and variance $\sigma_i^2 > 0$ of each objective coefficient $\tilde{c}_i$ is known with the support over the entire real line. For generalization to higher-order moments, the reader is referred to Bertsimas et al. (2004, 2006). Let Θ denote the set of multivariate distributions θ for the objective coefficients such that they satisfy the given mean and variances for each $\tilde{c}_i$.

Under the MMM, with mean and variance information, Z^* is computed by solving

$$
\begin{aligned}
Z^* = \max \quad & \sum_{i\in\mathcal{N}} \sum_{k\in\mathcal{X}_i} k w_{ik} \\
\text{s.t.} \quad & \sum_{k\in\mathcal{X}_i} z_{ik} = \mu_i^2 + \sigma_i^2 \quad \forall\, i \in \mathcal{N} \\
& \sum_{k\in\mathcal{X}_i} w_{ik} = \mu_i \quad \forall\, i \in \mathcal{N} \\
& \sum_{k\in\mathcal{X}_i} y_{ik} = 1 \quad \forall\, i \in \mathcal{N} \\
& z_{ik} y_{ik} \geq w_{ik}^2 \quad \forall\, k \in \mathcal{X}_i\ \forall\, i \in \mathcal{N} \\
& \mathbf{y} \in \mathcal{CH}(\mathcal{Y}).
\end{aligned} \tag{9}
$$

This formulation is obtained using an argument similar to §3.1 and is skipped. The variables can be interpreted as (scaled) conditional moments (refer to Bertsimas et al. 2006):

$$
\begin{pmatrix} z_{ik} \\ w_{ik} \\ y_{ik} \end{pmatrix} = \begin{pmatrix} E_\theta(\tilde{c}_i^2 \mid x_i^*(\tilde{\mathbf{c}}) = k) P_\theta(x_i^*(\tilde{\mathbf{c}}) = k) \\ E_\theta(\tilde{c}_i \mid x_i^*(\tilde{\mathbf{c}}) = k) P_\theta(x_i^*(\tilde{\mathbf{c}}) = k) \\ P_\theta(x_i^*(\tilde{\mathbf{c}}) = k) \end{pmatrix},
$$

with the y_{ik} variables denoting the persistency values. The first three constraints model the marginal moment conditions. The constraints $z_{ik} y_{ik} \geq w_{ik}^2$ where $y_{ik} \in [0, 1]$ correspond to the moment feasibility conditions (the Cauchy-Schwarz inequality in this case).

Proposition 2. *Under the MMM, with mean and variance information, Z^* and the persistence values are computed by solving the concave maximization problem*:

$$
Z^* = \max\left\{ \sum_{i\in\mathcal{N}} \left(\mu_i \sum_{k\in\mathcal{X}_i} k y_{ik} + \sigma_i \sqrt{\sum_{k\in\mathcal{X}_i} k^2 y_{ik} - \left(\sum_{k\in\mathcal{X}_i} k y_{ik} \right)^2} \right) : \mathbf{y} \in \mathcal{CH}(\mathcal{Y}) \right\}. \tag{10}
$$

Proof. See the appendix. □

The above proposition can be simplified further in the case of 0-1 optimization problem.

Lemma 2. *Given only mean and variance information, Z^* and the persistence values for 0-1 optimization problems are computed by solving the concave maximization problem*:

$$
Z^* = \max\left\{ \sum_{i\in\mathcal{N}} (\mu_i x_i + \sigma_i \sqrt{x_i(1-x_i)}) : \mathbf{x} \in \mathcal{CH}(\mathcal{X}) \right\}. \tag{11}
$$

We next provide a simple application of Lemma 2 for approximating a normal distribution.

Example 3. For a standard normal random variable, the distribution function is

$$
P(\tilde{c} \leq z) = \phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt,
$$

for which no closed form expression exists but numerical estimates are easily available. Consider the following simple analytic approximation to the normal distribution. Let $\tilde{c}$ denote a random variable with mean μ and variance σ^2. Without loss of generality, let $\mu = 0$ and $\sigma^2 = 1$. Using Lemma 2, the persistency model for $Z(\tilde{c}) = \max(\tilde{c}, z)$ reduces to the following problem:

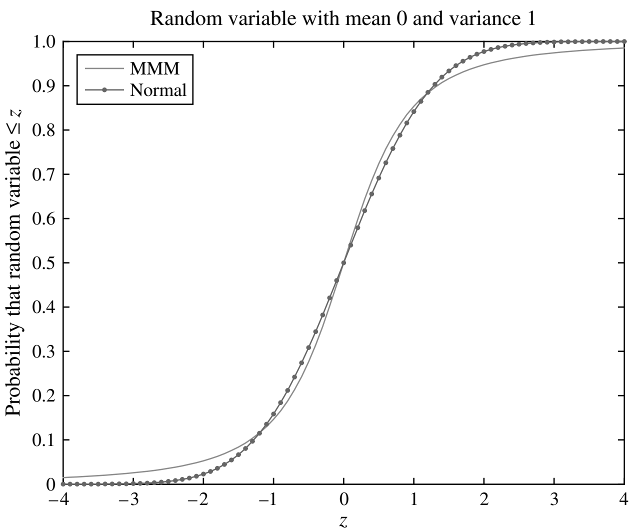
$$
Z^* = \max\{ (\sqrt{x_1(1-x_1)} + z x_2) : x_1 + x_2 = 1,\, x_1 \geq 0,\, x_2 \geq 0 \}.
$$

Solving this optimization problem, we obtain

$$
P_{\theta^*}(\tilde{c} \leq z) = P_{\theta^*}(x_2(\tilde{c}) = 1) = \frac{1}{2}\left(1 + \frac{z}{\sqrt{z^2+1}} \right).
$$

Figure 3 compares these values for $P(\tilde{c} \leq z)$ and $P_{\theta^*}(\tilde{c} \leq z)$ as a function of z. Clearly, these two values are observed to be in close agreement with an absolute error of at most 0.0321. This suggests that the extremal distribution obtained from the persistency model for a single variable is a good approximation to the normal distribution.

Figure 3 Comparison of $P(\tilde{c} \le z)$ for Normal Distribution and MMM



3.3. Range and Unimodal Distribution

The approach described in this paper is flexible and can be enhanced in different ways. For instance, the MMM can be strengthened further if we know the range of support for each random variable. For instance, if we know that $\tilde{c}_i \in [\underline{c}_i, \bar{c}_i]$, then, using the condition

$$(\bar{c}_i - \tilde{c}_i)(\tilde{c}_i - \underline{c}_i) \ge 0,$$

we can capture the range conditions into the persistency model (9) through the addition of the following inequality:

$$(\underline{c}_i + \bar{c}_i) w_{ik} \ge \underline{c}_i \bar{c}_i y_{ik} + z_{ik}.$$

Other information such as the shape of the distribution can also be introduced into the formulation, as we describe next.

A random variable is said to be unimodal with mode 0 if it is a mixture of a Dirac distribution δ_0 and a distribution with density function $f(\cdot)$ that is nondecreasing on $(-\infty, 0]$ and nonincreasing on $[0, \infty)$. Some examples of unimodal distributions are the standard normal, exponential, Cauchy probability densities, and the class of stable distributions. The main result driving our enhanced reformulation is the following result from Khintchine: A random variable $\tilde{c}$ has a unimodal distribution if and only if there exists a random variable $\tilde{d}$ such that $\tilde{c} \sim \tilde{u}\tilde{d}$, where $\tilde{u}$ is a uniform $[0, 1]$ random variable independent of $\tilde{d}$.

Assume that the objective coefficient $\tilde{c}_i$ is unimodal with mean μ_i and variance σ_i^2. We can use Khintchine's Theorem to refine our persistency model under the moments approach. Let $\tilde{c}_i = \tilde{u}_i \tilde{d}_i$, where $\tilde{u}_i$ is uniformly generated in $[0, 1]$ and independent of $\tilde{d}_i$. The first and second moments for $\tilde{d}_i$ are then given as

$$E_\theta(\tilde{d}_i) = 2\mu_i \quad \text{and} \quad E_\theta(\tilde{d}_i^2) = 3(\mu_i^2 + \sigma_i^2).$$

Furthermore, we have

$$\begin{aligned} E_\theta(\tilde{c}_i x_i(\tilde{\mathbf{c}})) &= \int_0^1 E_\theta(\tilde{u}_i \tilde{d}_i x_i(\tilde{\mathbf{c}}) \mid \tilde{u}_i = u)\, du \\ &= \int_0^1 u E_\theta(\tilde{d}_i x_i(\tilde{\mathbf{c}}) \mid \tilde{u}_i = u)\, du \\ &= \int_0^1 u \sum_{k \in \mathscr{X}_i} (k E_\theta(\tilde{d}_i \mid x_i(\tilde{\mathbf{c}}) = k, \tilde{u}_i = u) \\ &\qquad \cdot P_\theta(x_i(\tilde{\mathbf{c}}) = k \mid \tilde{u}_i = u))\, du. \end{aligned}$$

Define the variables as (scaled) conditional moments:

$$\begin{pmatrix} z_{iku} \\ w_{iku} \\ y_{iku} \end{pmatrix} = \begin{pmatrix} E_\theta(\tilde{d}_i^2 \mid x_i^*(\tilde{\mathbf{c}}) = k, \tilde{u}_i = u) P_\theta(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{u}_i = u) \\ E_\theta(\tilde{d}_i \mid x_i^*(\tilde{\mathbf{c}}) = k, \tilde{u}_i = u) P_\theta(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{u}_i = u) \\ P_\theta(x_i^*(\tilde{\mathbf{c}}) = k \mid \tilde{u}_i = u) \end{pmatrix}.$$

We can then reformulate the persistency model using

$$\int_0^1 P_\theta(x_i(\tilde{\mathbf{c}}) = k \mid \tilde{u}_i = u)\, du = P_\theta(x_i(\tilde{\mathbf{c}}) = k).$$

Proposition 3. *Under the MMM given mean, variances, and unimodal distributions, Z^* and the persistence values are computed by solving the problem*

$$\begin{aligned} Z^* = \max \ & \sum_{i \in \mathscr{N}} \sum_{k \in \mathscr{X}_i} k \int_0^1 u w_{iku}\, du \\ \text{s.t.} \ & \sum_{k \in \mathscr{X}_i} z_{iku} = 3(\mu_i^2 + \sigma_i^2) \quad \forall\, u \in [0, 1]\ \forall\, i \in \mathscr{N} \\ & \sum_{k \in \mathscr{X}_i} w_{iku} = 2\mu_i \quad \forall\, u \in [0, 1]\ \forall\, i \in \mathscr{N} \\ & \sum_{k \in \mathscr{X}_i} y_{iku} = 1 \quad \forall\, u \in [0, 1]\ \forall\, i \in \mathscr{N} \\ & z_{iku} y_{iku} \ge w_{iku}^2, \quad \forall\, u \in [0, 1],\ \forall\, k \in \mathscr{X}_i\ \forall\, i \in \mathscr{N} \\ & \left(\int_0^1 y_{iku}\, du \right)_{k \in \mathscr{X}_i,\, i \in \mathscr{N}} \in \mathscr{CH}(\mathscr{Y}). \end{aligned}$$

One disadvantage of this model is that there are infinitely many variables, depending on $u \in [0, 1]$. Numerically, this problem can be solved by discretization. The increase in the number of variables, however, allows us to incorporate the characteristics of the shape of the marginal distribution.

4. Applications: Discrete Choice Modeling

In this section, we test the performance of the persistency model on two canonical problems chosen from the literature: discrete choice models and the integer knapsack problem (i.e., discrete choice under budget constraint). The former is selected to demonstrate estimation and prediction of choice probabilities for discrete choice models using commercial nonlinear solvers. The latter problem is selected to demonstrate how the mapping to a higher dimensional polytope can be executed for a NP-hard discrete optimization problem.

4.1. Choice Probability in Discrete Choice Models

Discrete choice models deal with the problem of predicting the probability that a random customer will choose a selected alternative from a finite set of alternatives. Consider a set of alternatives $\mathcal{N} = \{1, 2, \ldots, n\}$, where the utility that an individual assigns to alternative j can be expressed as

$$\tilde{U}_j = V_j + \tilde{\epsilon}_j,$$

where V_j is the deterministic component that relates to the known attributes of the alternative and $\tilde{\epsilon}_j$ is the random error associated with a model because of factors not considered. The random utility maximization problem is then formulated as

$$Z(\tilde{\mathbf{U}}) = \max\left\{\sum_{j\in\mathcal{N}} \tilde{U}_j x_j\colon \sum_{j\in\mathcal{N}} x_j = 1,\ x_j \in \{0,1\}\ \forall\, j \in \mathcal{N}\right\}.$$

Let P_j denote the probability that alternative j is selected by the individual. This choice probability is the persistence value for the problem $Z(\tilde{\mathbf{U}})$:

$$P_j = P(x_j^*(\tilde{\mathbf{U}}) = 1) = P(\tilde{U}_j \geq \tilde{U}_k\ \forall\, k \in \mathcal{N}).$$

The classical logit model (see McFadden 1974) starts with the assumption that the error terms $\tilde{\epsilon}_j$s are modeled by independent extreme value distributions, so that the following elegant closed-form solution for the choice probabilities can be obtained:

$$P_j = \frac{e^{V_j}}{\sum_{k\in\mathcal{N}} e^{V_k}}.$$

However, this approach has some drawbacks. For example, the formula implies the independence of irrelevant alternatives property wherein the relative ratio of the choice probabilities for two alternatives is independent of the remaining alternatives. This property is not always observed in practice wherein the entire choice set helps determine the relative probabilities. The probit model, using correlated normal distributions, can be used to overcome this shortcoming, but at the added cost of finding choice probabilities through extensive simulation. Using the models proposed in this paper, we can obtain simple formulas for choice probabilities. For example, suppose the cumulative distribution function for utility $\tilde{U}_j$ is $F_j(c)$. From Theorem 1, the persistency model is

$$\max\left\{\sum_{j\in\mathcal{N}}\left(\int_{1-x_j}^{1} F_j^{-1}(t)\,dt\right)\colon \sum_{j\in\mathcal{N}} x_j = 1, x_j \geq 0\ \forall\, j\in\mathcal{N}\right\}. \quad (12)$$

The optimality conditions yields the choice probability as

$$P_{\theta^*}(x_j^*(\tilde{\mathbf{U}}) = 1) = 1 - F_j(\lambda^*),$$

where the Lagrange multiplier λ^* is found by solving the equation

$$\sum_{j\in\mathcal{N}} (1 - F_j(\lambda^*)) = 1.$$

Alternatively, suppose we know the mean V_j and variance σ_j^2 of the utility $\tilde{U}_j$, but the exact distribution function is unknown. Then from Lemma 2, we need to solve

$$\max\left\{\sum_{j\in\mathcal{N}}\left(V_j x_j + \sigma_j\sqrt{x_j(1-x_j)}\right)\colon \sum_{j\in\mathcal{N}} x_j = 1,\ x_j \geq 0\ \forall\, j \in \mathcal{N}\right\}. \quad (13)$$

The optimality conditions yield the choice probability as

$$P_{\theta^*}(x_j^*(\tilde{\mathbf{U}}) = 1) = \frac{1}{2}\left(1 + \frac{V_j - \lambda^*}{\sqrt{(V_j - \lambda^*)^2 + \sigma_j^2}}\right),$$

where the Lagrange multiplier λ^* is found by solving the equation

$$\sum_{j\in\mathcal{N}} \frac{1}{2}\left(1 + \frac{V_j - \lambda^*}{\sqrt{(V_j - \lambda^*)^2 + \sigma_j^2}}\right) = 1.$$

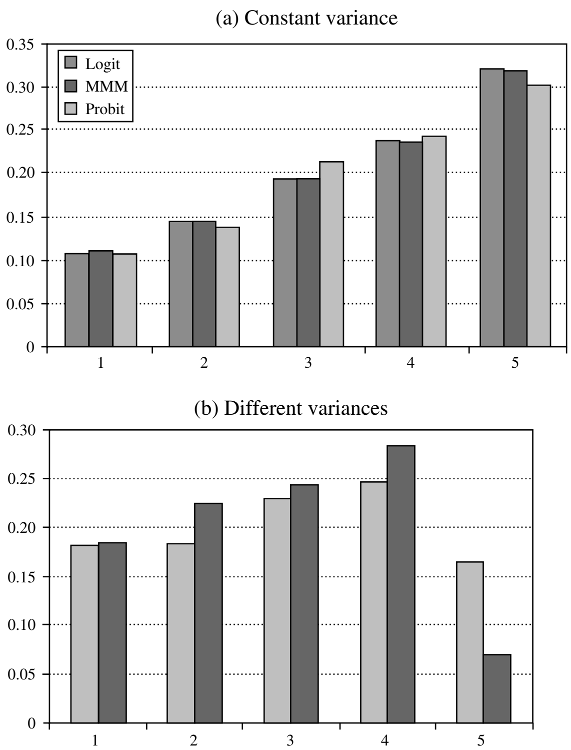
It is easy to see that the independence of irrelevant alternatives property need not hold under our model because of the dependence on λ^*. Interestingly, the choice probabilities obtained under the persistence model and many discrete choice models are strikingly similar for the numerical examples on which we have experimented.

Example 4. We compare the choice probabilities for a simple five-product example using logit (closed form), probit (simulation), and MMM (root finding). The first example assumes that each product has a common variance of $\pi^2/6$, with means

$$V_1 = 1.2,\quad V_2 = 1.5,\quad V_3 = 1.8,\quad V_4 = 2,\quad V_5 = 2.3.$$

Figure 4(a) plots the choice probabilities obtained from the three models. For the probit model, the simulations were carried out under the assumption that the utilities are independent of each other. The results are strikingly similar, although they are obtained under very different assumptions. Note that product 5 with the higher mean attribute has a higher choice probability. The second example (see Figure 4(b)) predicts choice probabilities when the error terms related

Figure 4 Choice Models Comparison



to the products are not identically distributed. We use the mean values as before, but the variances are set to

$$\sigma_1^2 = 4, \quad \sigma_2^2 = 4, \quad \sigma_3^2 = 3, \quad \sigma_4^2 = 3, \quad \sigma_5^2 = 0.1.$$

Note that now product 5 has the smallest variance. Unfortunately, this proves detrimental to the probability of product 5 being selected by the consumer. In fact, in this example, product 5 with the highest mean attribute and the smallest variance is now the product least likely to be selected. This can be observed under the probit model by performing simulations. Our model predicts that the preferences for the products are in the same order—4, 3, 2, 1, and 5—as the probit model.

We next consider the estimation problem in discrete choice models. Consider a set of individuals $\mathcal{I}$, where the utility that an individual $i \in \mathcal{I}$ assigns to alternative $j \in \mathcal{N}$ is expressed as

$$\widetilde{U}_{ij} = V_{ij} + \tilde{\epsilon}_{ij}.$$

The deterministic component of the utility is represented as $V_{ij} = \beta' \mathbf{x}_{ij}$, where $\mathbf{x}_{ij}$ is the vector of attributes characterizing the individual and the alternative, and β is the parameters of the model that need to be estimated. The traditional approach to estimating β is to use a log-likelihood estimation technique (see Ben-Akiva and Lerman 1985). Let $y_{ij} = 1$ if alternative j is selected by individual i and zero otherwise. The probability that each person in $\mathcal{I}$ chooses the alternative that they were actually observed to choose is

$$\mathcal{L}(\beta) = \prod_{i \in \mathcal{I}} \prod_{j \in \mathcal{N}} (P_{ij})^{y_{ij}},$$

where the probability P_{ij} that individual i chooses j is

$$P_{ij} = P(V_{ij} + \beta' \mathbf{x}_{ij} + \tilde{\epsilon}_{ij} \ge V_{ik} + \beta' \mathbf{x}_{ik} + \tilde{\epsilon}_{ik} \ \ \forall\, k \in \mathcal{N}).$$

The maximum log-likelihood estimator is obtained by solving the optimization problem:

$$\max_{\beta} \mathcal{L}\mathcal{L}(\beta) = \max_{\beta} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{N}} y_{ij} \ln P_{ij}.$$

We compare three different models for the estimation procedure—the popular multinomial logit model (MNL), the cross-nested logit model (CNL), and our MMM:

For MNL (see McFadden 1974), the error terms are assumed to be independent and identically distributed as extreme value distributions. The distribution for the error term is

$$F(\epsilon_{ij}) = e^{e^{-\epsilon_{ij}}}.$$

The choice probabilities are given as

$$P_{ij} = \frac{e^{\beta' \mathbf{x}_{ij}}}{\sum_{k \in \mathcal{N}} e^{\beta' \mathbf{x}_{ik}}}.$$

The log-likelihood objective is globally concave in this model with efficient convex optimization solvers available to estimate the optimal β. A freely available package BIOGEME (developed by Bierlaire 2003) was used for the computations. The maximum log-likelihood estimation was performed using the sequential equality constrained quadratic programming method (cf. DONLP2 in Spellucci 1993) that is incorporated in BIOGEME.

For CNL (see Ben-Akiva and Bierlaire 1999), the choice set is partitioned into a set of nests denoted as $\mathcal{M}$. The parameter μ_m is the scale parameter for nest $m \in \mathcal{M}$, and α_{jm} is the cross-nesting parameter for alternative j in nest m such that $\sum_{m \in \mathcal{M}} \alpha_{jm} = 1$. The joint distribution for the error terms in CNL is given as

$$F(\epsilon_{i1}, \ldots, \epsilon_{in}) = \exp\left(- \sum_{m \in \mathcal{M}} \left(\sum_{j} (\alpha_{jm}^{\mu_m} e^{-\epsilon_{ij} \mu_m}) \right)^{1/\mu_m} \right).$$

The choice probabilities under this model is

$$P_{ij} = \sum_{m \in \mathcal{M}} \left(\frac{\left(\sum_{k \in \mathcal{N}} \alpha_{km}^{\mu_m} e^{\mu_m \beta' \mathbf{x}_{ik}} \right)^{1/\mu_m}}{\sum_{l \in \mathcal{M}} \left(\sum_{k \in \mathcal{N}} \alpha_{kl}^{\mu_l} e^{\mu_l \beta' \mathbf{x}_{ik}} \right)^{1/\mu_l}} \times \frac{\alpha_{jm}^{\mu_m} e^{\mu_m \beta' \mathbf{x}_{ij}}}{\sum_{k \in \mathcal{N}} \alpha_{km}^{\mu_m} e^{\mu_m \beta' \mathbf{x}_{ik}}} \right).$$

The log-likelihood function is unfortunately not concave under this model, with no guarantee of finding the global optimal solution efficiently. For our computations, we used BIOGEME with the DONLP2 algorithm to estimate the coefficients.

For MMM, we assume that the error term $\tilde{\epsilon}_{ij}$ has mean 0 and variance σ_{ij}^2. The random utility $\widetilde{U}_{ij}$ is then distributed with mean $\beta'\mathbf{x}_{ij}$ and variance σ_{ij}^2. The variance term is used to capture the variability caused by factors not captured by the specified attributes of the model. The choice probabilities are given as

$$P_{ij} = \frac{1}{2}\left(1 + \frac{\beta'\mathbf{x}_{ij} - \lambda_i}{\sqrt{(\beta'\mathbf{x}_{ij} - \lambda_i)^2 + \sigma_{ij}^2}}\right),$$

where λ_i is found by solving the equation

$$\sum_{j\in\mathcal{N}} \frac{1}{2}\left(1 + \frac{\beta'\mathbf{x}_{ij} - \lambda_i}{\sqrt{(\beta'\mathbf{x}_{ij} - \lambda_i)^2 + \sigma_{ij}^2}}\right) = 1.$$

The maximum log-likelihood estimation problem under MMM is

$$\begin{aligned} \max\ & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{N}} y_{ij} \ln \frac{1}{2}\left(1 + \frac{\beta'\mathbf{x}_{ij} - \lambda_i}{\sqrt{(\beta'\mathbf{x}_{ij} - \lambda_i)^2 + \sigma_{ij}^2}}\right) \\ \text{s.t.}\ & \sum_{j\in\mathcal{N}} \frac{1}{2}\left(1 + \frac{\beta'\mathbf{x}_{ij} - \lambda_i}{\sqrt{(\beta'\mathbf{x}_{ij} - \lambda_i)^2 + \sigma_{ij}^2}}\right) = 1 \quad \forall i \in \mathcal{I}. \end{aligned} \tag{14}$$

The log-likelihood function again is not necessarily concave under this model. For our computations, we considered two versions of MMM. The first version assumed a constant variance $\sigma_{ij}^2 = \pi^2/6$ for all the error terms, as in logit. The second version assumed that the variances are related only to the alternatives and not the individuals. Thus, the variances $\sigma_{ij}^2 = \sigma_j^2$ are unknown and estimated in the log-likelihood problem. We used LOQO (a system to solve smooth constrained optimization problems), developed by Vanderbei (2006), to estimate the parameters. LOQO uses an infeasible primal-dual interior point method applied to quadratic approximations to solve the original problem.

Example 5. We compare the performance of the three discrete choice models, using a real-life data set taken from Bierlaire et al. (2001). This is a transportation mode choice problem with three available alternatives: train, car, and the swiss metro (SM). The attributes that were modeled for each alternative are indicated in Table 1.

The data consisted of a total of 10,710 preferences (including revealed and stated preferences) for the three alternatives. The parameters for the model were estimated using part of the data (N preferences) and tested for accuracy using the remaining part of the data (10,710-N preferences). The value of N was varied from 10% (1,071 preferences) to 50% (5,355 preferences) of the total number of preferences with increments of about 5% each, in the order listed in the data set. For the MNL model, a total of 9 parameters were estimated. For the CNL model, two nests were used and a total of 14 parameters were estimated, including μ-Exist, μ-Future, α-Exist-Car, α-Exist-SM, and α-Exist-Train. These additional parameters captured the scale and cross-nesting parameters for CNL.

For the MMM, there were a total of nine parameters with two additional parameters in the version with different variances. The two additional parameters captured the Variance-Train and Variance-SM with the variance for the Car alternative set to the default of $\pi^2/6$. A nonzero lower bound of 0.0001 was set on the variance terms to ensure fast convergence for the LOQO solver. The parameters estimated using the maximum log-likelihood formulation are shown in Table 2 for the instance with $N = 5{,}355$. In terms of the log-likelihood objective, MMM with different variances is clearly better than CNL, which is better

Table 1 Attributes and Alternatives

	Alternatives		
Attributes	Train	Swiss metro	Car
Age	B-Age		
Cost	B-Cost	B-Cost	B-Cost
Frequency	B-Freq.	B-Freq.	
Luggage			B-Luggage
Seats		B-Seats	
Time	B-Time	B-Time	B-Time
GA (season ticket)	B-GA	B-GA	
ASC (alternative specific constant)		ASC-SM	ASC-Car

Table 2 Estimation Results Using MNL, CNL, and MMM

Parameter	MNL	CNL	MMM (constant variance)	MMM (different variances)
ASC-car	0.5942	0.3574	0.6826	0.1003
ASC-SM	0.5310	0.2746	0.5826	0.0642
B-Age	0.0639	0.0506	0.1276	0.1240
B-Cost	−0.0062	−0.0047	−0.0103	−0.0105
B-Freq	−0.0065	−0.0041	−0.0010	−0.0059
B-GA	1.8439	1.0261	2.3200	2.1808
B-Luggage	−0.1367	−0.1453	−0.1452	−0.2464
B-Seats	−0.1816	0.0013	−0.0186	−0.1312
B-Time	−0.0132	−0.0098	−0.0190	−0.0170
μ-Exist		2.5857		
μ-Future		1.0000		
α-Exist-Car		0.9665		
α-Exist-SM		0.0000		
α-Exist-Train		0.9548		
Variance-Car			1.6450	1.6450
Variance-Train			1.6450	0.1619
Variance-SM			1.6450	3.1465
Log-likelihood	−4,242.05	−4,117.23	−4,146.43	−4,065.58
Time (seconds)	3	64	13.6	42.5
N	5,355	5,355	5,355	5,355

Note. ASC, alternative specific constant; GA, season ticket.

Table 3 Prediction Results Using MNL, CNL, and MMM

N	1,071	1,606	2,142	2,677	3,213	3,748	4,284	4,819	5,355
				Actual choice (in 10,710-*N* instances)					
Train	0.122	0.106	0.095	0.099	0.105	0.111	0.116	0.123	0.132
SM	0.564	0.564	0.564	0.561	0.557	0.547	0.553	0.559	0.574
Car	0.314	0.330	0.341	0.340	0.338	0.342	0.331	0.318	0.294
				Predicted choice using in-sample average					
Train	0.223	0.279	0.277	0.231	0.195	0.172	0.156	0.144	0.132
SM	0.732	0.672	0.647	0.640	0.635	0.642	0.621	0.607	0.587
Car	0.045	0.049	0.076	0.129	0.170	0.186	0.223	0.249	0.281
LL	−12,872	−12,352	−10,670	−8,786	−7,674	−7,050	−6,270	−5,652	−5,065
Error	0.538	0.562	0.530	0.422	0.336	0.312	0.216	0.138	0.026
				Predicted choice using MNL					
Train	0.167	0.241	0.247	0.198	0.167	0.151	0.143	0.136	0.131
SM	0.760	0.686	0.616	0.615	0.612	0.616	0.604	0.595	0.585
Car	0.073	0.073	0.137	0.187	0.221	0.233	0.253	0.269	0.284
LL	−10,603	−10,319	−8,621	−7,171	−6,337	−5,827	−5,232	−4,739	−4,279
Error	0.482	0.514	0.408	0.306	0.234	0.218	0.156	0.098	0.022*
				Predicted choice using CNL					
Train	0.151	0.213	0.246	0.156	0.137	0.128	0.129	0.127	0.127
SM	0.778	0.691	0.617	0.635	0.626	0.628	0.609	0.599	0.587
Car	0.071	0.096	0.137	0.209	0.237	0.244	0.262	0.274	0.286
LL	−10,502	−9,514	−8,599	−6,853*	−6,165*	−5,696	−5,128	−4,652	−4,192
Error	0.486	0.468	0.408	0.262	0.202*	0.196*	0.138*	0.088*	0.026
				Predicted choice using MMM with constant variance					
Train	0.150	0.223	0.221	0.178	0.152	0.140	0.135	0.129	0.126
SM	0.771	0.692	0.632	0.628	0.622	0.621	0.606	0.599	0.587
Car	0.079	0.085	0.147	0.194	0.226	0.239	0.259	0.272	0.287
LL	−10,277	−9,813	−8,244	−6,957	−6,207	−5,717	−5,109	−4,629	−4,171
Error	0.470	0.490	0.388	0.292	0.224	0.206	0.144	0.092	0.026
				Predicted choice using MMM with different variances					
Train	0.124	0.183	0.167	0.142	0.127	0.122	0.125	0.125	0.126
SM	0.777	0.699	0.657	0.646	0.638	0.635	0.615	0.601	0.585
Car	0.099	0.118	0.176	0.212	0.235	0.243	0.260	0.274	0.289
LL	−9,730*	−9,022*	−7,856*	−6,886	−6,249	−5,614*	−5,049*	−4,563*	−4,115*
Error	0.430*	0.424*	0.330*	0.256*	0.206	0.198	0.142	0.088*	0.022*

*Indicates the best among the models.

than MMM with constant variance, which is better than MNL. This is achieved in MMM without capturing the rather complex structure of the CNL model. Similar patterns are observed for the other values of N and hence are not displayed because of space restrictions.

In Table 3, we compare the choice probabilities in the out-of-sample data (10,710-N preferences) with the in-sample data (N preferences). As expected, as N increases, the prediction using the in-sample average become better. The log-likelihood (LL) value is calculated by adding up the logarithms of the predicted choice probabilities for the actual alternative that was observed to be chosen in the out-of-sample data. The error measures the aggregate deviation in the choice probabilities. We next test the accuracy of the estimations from the more sophisticated discrete choice models (see Table 3). In terms of the log-likelihood value, MMM with different variances outperforms the CNL in seven of the nine cases and seems to be a good choice model for this data set.

4.2. Discrete Choice Under Budget Constraint

Consider the discrete choice problem, where a set of items $\mathcal{N} = \{1, 2, \ldots, n\}$ is considered. Let $\tilde{c}_i$ denote the utility value accrued for each unit of item i and a_i denote the weight of any item $i \in \mathcal{N}$. Assume also that the total weight available to the consumer is b. Let x_i be the units of item i that the consumer needs to choose to maximize her utility values. This problem can be formulated as the following integer knapsack problem:

$$\begin{aligned} \max \quad & \sum_{i\in\mathcal{N}} \tilde{c}_i x_i \\ \text{s.t.} \quad & \sum_{i\in\mathcal{N}} a_i x_i \le b \qquad (15) \\ & x_i \in \{0, 1, \ldots, \lfloor b/a_i \rfloor\} \quad \forall\, i \in \mathcal{N}. \end{aligned}$$

In this section, we focus on the stochastic integer knapsack problem, where the utility $\tilde{c}_i$ is uncertain, and investigate the persistency of the variable x_i under mean μ_i, variance σ_i^2, and range $[\underline{c}_i, \bar{c}_i]$.

The persistence values under the MMM are obtained by solving

$$\begin{aligned} Z^* = \max\ & \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}_i} k w_{ik} \\ \text{s.t.}\ & \sum_{k\in\mathcal{K}_i} z_{ik} = \mu_i^2 + \sigma_i^2 \quad \forall\, i\in\mathcal{N} \\ & \sum_{k\in\mathcal{K}_i} w_{ik} = \mu_i \quad \forall\, i\in\mathcal{N} \\ & \sum_{k\in\mathcal{K}_i} y_{ik} = 1 \quad \forall\, i\in\mathcal{N} \\ & z_{ik} y_{ik} \ge w_{ik}^2 \quad \forall\, k\in\mathcal{K}_i\ \forall\, i\in\mathcal{N} \\ & (\underline{c}_i + \bar{c}_i) w_{ik} \ge \underline{c}_i \bar{c}_i y_{ik} + z_{ik} \quad \forall\, k\in\mathcal{K}_i\ \forall\, i\in\mathcal{N} \\ & \mathbf{y} \in \mathcal{CH}(\mathcal{Y}), \end{aligned} \tag{16}$$

where the region $\mathcal{Y}$ is given as

$$\mathcal{Y} = \left\{ \mathbf{y} \,\middle|\, \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}_i} a_i k y_{ik} \le b,\ \sum_{k\in\mathcal{K}_i} y_{ik} = 1\ \forall\, i\in\mathcal{N},\ y_{ik}\in\{0,1\}\ \forall\, k\in\mathcal{K}_i\ \forall\, i\in\mathcal{N} \right\}.$$

The difficulty in solving the second-order cone program (16) lies in the last constraint, which needs a complete characterization of the convex hull of the set $\mathcal{Y}$. In our computations, we used a straightforward vertex-based formulation to represent $\mathcal{CH}(\mathcal{Y})$:

$$\mathcal{CH}(\mathcal{Y}) = \left\{ \mathbf{y} \,\middle|\, y_{ik} = \sum_{p\in\mathcal{P}} \lambda_p y_{ik}[p]\ \forall\, k\in\mathcal{K}_i\ \forall\, i\in\mathcal{N},\ \sum_{p\in\mathcal{P}} \lambda_p = 1\ \forall\, p\in\mathcal{P},\ \lambda_p \ge 0\ \forall\, p\in\mathcal{P} \right\},$$

where $y_{ik}[p]$ denotes the value of the y_{ik} variable at the pth extreme point. This formulation is useful if the number of vertices $|\mathcal{P}|$ is not too large. The second-order conic program (16) was solved with the conic programming solver SeDuMi, which can be downloaded from http://sedumi.mcmaster.ca/.

Example 6. We consider a stochastic integer knapsack problem taken from Steinberg and Parks (1979) with the size of the knapsack to be 30. Table 4 shows the weight and mean and variance information of the utilities. While the original problem assumes $\tilde{c}_i$ to be normally distributed, we set the support to be $[\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$ so that the probability for a normally generated random variable to be in the support is 99.7%. There are exactly 70 vertices in the convex hull of $\mathcal{Y}$.

To evaluate the persistency values from (16), we compared the results with simulation tests based on three canonical families of probability distribution, chosen for their differences in behavior around the mean. We generate the random coefficients of each item in the objective function using normal distribution (centered around the mean), uniform distribution (evenly distributed around the mean), and a two-point distribution (dispersed from mean) with $\tilde{c}_i$ set to be either $\mu_i - \sigma_i$ or $\mu_i + \sigma_i$. The parameters for these distributions are chosen to fit the prescribed first- and second-moment conditions. A total of 10,000 cases were generated for each distribution. Table 5 shows the persistence values obtained from both the MMM and the simulations.

Table 4 **The Steinberg-Parks Numerical Example**

i	1	2	3	4	5	6	7	8	9	10
a_i	5	7	11	9	8	4	12	10	3	6
μ_i	7	12	14	13	12	5	16	11	4	7
σ_i^2	15	20	15	10	8	20	8	15	20	25

Eleven of the persistence values equal zero in both the MMM and the simulation and hence are not included in the table. Generally, the values from MMM agree with the simulation. The average of the differences between the persistency from MMM and the simulations is 0.0346, and the MMM accurately predicts that x_3, x_4, x_5, x_7, and x_8 are very likely to be zero. Compared with the results for any one of the distributions, 29 persistence values computed by the MMM have a difference of less than 0.05, and there are only 6 persistence values with a difference greater than 0.1. Even for these 6 values, the discrepancy is mainly caused by the difference in the simulation results using various distributions. Based on these results, we conclude that the MMM gives valuable insights for the stochastic knapsack problem.

Another interesting observation is that x_5 is more likely to take value zero than x_6 in both the MMM persistence values and the simulation results. However, if we consider the average value/weight ratio using μ_i in place of $\tilde{c}_i$, the ratios for x_5 and x_6 are 1.50 and 1.25, respectively. Therefore, a naive decision maker would assign higher choice probability to x_5 than to x_6. It is indeed interesting that the MMM, assuming only minimal conditions on the random parameters, is able to pick this anomaly out.

4.2.1. Permutation Experiments. We further investigated the robustness of the persistency solution obtained from the MMM. We considered 10 stochastic knapsack problems with the values for a_i and b to be the same as the Steinberg-Parks example, and $\tilde{c}_i$, $i = 1, \ldots, 10$ assumed to be independent truncated normal distributions ($\tilde{c}_i \ge 0$). For each problem, the mean and variance of $\tilde{c}_i$ were generated uniformly from $[4, 16]$ and $[8, 12]$, respectively. Given the distributions of $\tilde{c}_i$, we first generated 100 vectors of the

Table 5 MMM Persistence Values and Simulation Results

Persistency	MMM	Normal	Uniform	Two-point	Persistency	MMM	Normal	Uniform	Two-point
$P(x_1^*(\tilde{\mathbf{c}})=0)$	0.8719	0.8196	0.8206	0.8118	$P(x_6^*(\tilde{\mathbf{c}})=0)$	0.8620	0.7920	0.7770	0.7537
$P(x_1^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0195	0.0171	0.0661	$P(x_6^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0125	0.0118	0.0628
$P(x_1^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0168	0.0216	0.0599	$P(x_6^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0038	0.0015	0.0000
$P(x_1^*(\tilde{\mathbf{c}})=6)$	0.1281	0.1441	0.1407	0.0622	$P(x_6^*(\tilde{\mathbf{c}})=3)$	0.0000	0.0064	0.0042	0.0153
					$P(x_6^*(\tilde{\mathbf{c}})=4)$	0.0000	0.0290	0.0382	0.0646
					$P(x_6^*(\tilde{\mathbf{c}})=5)$	0.0000	0.0169	0.0212	0.0599
					$P(x_6^*(\tilde{\mathbf{c}})=6)$	0.0000	0.0546	0.0713	0.0287
					$P(x_6^*(\tilde{\mathbf{c}})=7)$	0.1380	0.0848	0.0748	0.0150
$P(x_2^*(\tilde{\mathbf{c}})=0)$	0.8715	0.7720	0.7793	0.7299	$P(x_7^*(\tilde{\mathbf{c}})=0)$	0.9523	0.9864	0.9893	0.9889
$P(x_2^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0041	0.0029	0.0017	$P(x_7^*(\tilde{\mathbf{c}})=1)$	0.0477	0.0078	0.0072	0.0111
$P(x_2^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0546	0.0560	0.0722	$P(x_7^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0058	0.0035	0.0000
$P(x_2^*(\tilde{\mathbf{c}})=3)$	0.0000	0.0618	0.0582	0.1328					
$P(x_2^*(\tilde{\mathbf{c}})=4)$	0.1285	0.1075	0.1036	0.0634					
$P(x_3^*(\tilde{\mathbf{c}})=0)$	0.9000	0.9776	0.9854	0.9973	$P(x_8^*(\tilde{\mathbf{c}})=0)$	0.9296	0.9856	0.9909	0.9992
$P(x_3^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0033	0.0038	0.0017	$P(x_8^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0029	0.0024	0.0008
$P(x_3^*(\tilde{\mathbf{c}})=2)$	0.1000	0.0191	0.0108	0.0010	$P(x_8^*(\tilde{\mathbf{c}})=3)$	0.0703	0.0115	0.0067	0.0000
$P(x_4^*(\tilde{\mathbf{c}})=0)$	0.9000	0.9306	0.9445	0.8920	$P(x_9^*(\tilde{\mathbf{c}})=0)$	0.7148	0.6422	0.6187	0.4961
$P(x_4^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0287	0.0243	0.0733	$P(x_9^*(\tilde{\mathbf{c}})=1)$	0.0523	0.0228	0.0159	0.0037
$P(x_4^*(\tilde{\mathbf{c}})=2)$	0.0477	0.0200	0.0193	0.0310	$P(x_9^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0449	0.0402	0.0000
$P(x_4^*(\tilde{\mathbf{c}})=3)$	0.0523	0.0207	0.0119	0.0037	$P(x_9^*(\tilde{\mathbf{c}})=3)$	0.0000	0.0273	0.0252	0.0000
					$P(x_9^*(\tilde{\mathbf{c}})=4)$	0.0000	0.0001	0.0000	0.0000
					$P(x_9^*(\tilde{\mathbf{c}})=10)$	0.2328	0.2627	0.3000	0.5002
$P(x_5^*(\tilde{\mathbf{c}})=0)$	0.9000	0.9208	0.9474	0.9703	$P(x_{10}^*(\tilde{\mathbf{c}})=0)$	0.8979	0.8589	0.8454	0.9401
$P(x_5^*(\tilde{\mathbf{c}})=1)$	0.1000	0.0158	0.0095	0.0018	$P(x_{10}^*(\tilde{\mathbf{c}})=1)$	0.0000	0.0437	0.0578	0.0457
$P(x_5^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0270	0.0180	0.0109	$P(x_{10}^*(\tilde{\mathbf{c}})=2)$	0.0000	0.0037	0.0040	0.0071
$P(x_5^*(\tilde{\mathbf{c}})=3)$	0.0000	0.0364	0.0251	0.0170	$P(x_{10}^*(\tilde{\mathbf{c}})=5)$	0.1021	0.0937	0.0928	0.0071

utility values, $\mathbf{c}^l$, $l = 1, \ldots, 100$. The MMM in (16) was solved using the marginal moments information from the samples.

Note that each row of the 100×10 matrix (c_i^l) represents an objective function for the knapsack problem with 10 items. No matter how we permute the entries in a column, the sample marginal moment values for each item remain unchanged. Hence, the prediction from MMM remains unchanged. However, the rows, which represent the objective functions used in the knapsack problem, will be affected by the column permutation. The experimental choice probabilities will thus be affected. We tested also whether the column permutation affects the accuracy of the persistency values obtained from the MMM by comparing the persistency values with simulation results (after permuting the entries in each column of the matrix). We repeated the experiment 10 times.

Note that we have to solve 100 deterministic knapsack problem for each experiment. Because we repeated the experiment 10 times, we have 10 empirical persistency values of $P(x_i^*(\tilde{\mathbf{c}}) = k)$. From the persistency values obtained, we considered the three statistics—the maximum, the minimum, and the average—and compared these three values with the MMM prediction. If the MMM prediction lies between the maximum and minimum, it is straightforward that the MMM model gives a reasonable prediction for the persistency. The difference between the MMM prediction and the average of the 10 empirical persistency values also measures how well the MMM approximates the true distribution. If the difference is less than the difference between the maximum and minimum empirical persistence values, the MMM prediction should be acceptable, as it may result in larger errors by picking the empirical persistency value of one permutation to predict the empirical persistency value of another.

The average performance of MMM is summarized in Table 6. Note that numbers in the column "x_i" represent the average performance across different $k \in \mathscr{X}_i$ and the 10 knapsack problems, and the column "Total" is the average across i, k, and the 10 problems. The first row gives the percentage of MMM predictions falling in the interval of the maximum and minimum empirical persistency values. In general, around 84% of the predictions are within the upper and lower bounds defined by the empirical values. The second row shows the difference between the MMM predictions and the average of empirical values, and the last row shows the difference between the maximum and minimum empirical values. Note that for any variable x_i, the difference between MMM and the average is less that that between the maximum and the minimum. Taking the average for all i, the difference between MMM and the empirical average is 75% of

Table 6 **MMM Performance in Permutation Experiments**

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Total
in(min, max) (%)	87.14	86.00	100.00	100.00	80.00	65.00	100.00	95.00	82.73	78.33	84.36
$\|\text{MMM} - \text{avg}\|$	0.0158	0.0121	0.0005	0.0015	0.0125	0.0355	0.0004	0.0045	0.0186	0.0202	0.0156
$\|\max - \min\|$	0.0339	0.0166	0.0020	0.0053	0.0140	0.0371	0.0027	0.0058	0.0215	0.0285	0.0207

that between the maximum and minimum empirical values. As a result, for this problem instance, we can conclude that the MMM gives reasonably accurate prediction for stochastic knapsack problems.

4.2.2. Comparison with Marginal Distribution Model. We next compare the numerical results for the MDM with the MMM for the stochastic knapsack problem. The feasible region is assumed to be the same as that in the Steinberg-Parks example, with the marginal distribution of $\tilde{c}_i$ set to be exponential with parameter $\lambda_i = 1/\mu_i$, where μ_i is displayed in Table 4. For the MMM, we let the mean and variance be $1/\lambda_i$ and $1/\lambda_i^2$, respectively, and the support be $[0, 10/\lambda_i]$ for each i. To evaluate these estimates, we also compared the results with simulation tests, where the random coefficients for the items are generated with independent exponential distributions with parameter λ_i.

Table 7 shows the persistence values for MDM, MMM, and the simulation results over 10,000 runs. Note that we exclude the persistence values, which are equal to zero in MDM, MMM, and the simulations. The average difference between the MDM and the simulation results is 0.0083, and that for the MMM is 0.0119 (close to 30% improvement). We also compared the maximum difference between our predictions and the simulation results, which are 0.0301 for the MDM model and 0.0516 for the MMM model. Thus, the MDM model performs better in both average and worst case performance. These results suggest that the additional information of the marginal distributions, if available, can add to the accuracy of the predictions of the persistency values.

5. Conclusion

In this paper, we study the problem of evaluating the probability distribution for the decision variable of a general stochastic discrete optimization problem. Under a set of known marginal distributions or marginal moments for the objective coefficients, we propose a concave maximization technique to

Table 7 **Performance of MDM and MMM Persistence Values**

Persistency	MDM	MMM	Exponential	Persistency	MDM	MMM	Exponential
$P(x_1^*(\tilde{\mathbf{c}}) = 0)$	0.8436	0.8611	0.8258	$P(x_6^*(\tilde{\mathbf{c}}) = 0)$	0.8831	0.9075	0.8559
$P(x_1^*(\tilde{\mathbf{c}}) = 1)$	0.0379	0.0121	0.0602	$P(x_6^*(\tilde{\mathbf{c}}) = 1)$	0.0000	0.0052	0.0290
$P(x_1^*(\tilde{\mathbf{c}}) = 2)$	0.0011	0.0007	0.0044	$P(x_6^*(\tilde{\mathbf{c}}) = 2)$	0.0397	0.0063	0.0197
$P(x_1^*(\tilde{\mathbf{c}}) = 6)$	0.1174	0.1262	0.1096	$P(x_6^*(\tilde{\mathbf{c}}) = 3)$	0.0000	0.0004	0.0069
				$P(x_6^*(\tilde{\mathbf{c}}) = 4)$	0.0014	0.0003	0.0078
				$P(x_6^*(\tilde{\mathbf{c}}) = 5)$	0.0146	0.0005	0.0067
				$P(x_6^*(\tilde{\mathbf{c}}) = 6)$	0.0000	0.0005	0.0144
				$P(x_6^*(\tilde{\mathbf{c}}) = 7)$	0.0612	0.0793	0.0596
$P(x_2^*(\tilde{\mathbf{c}}) = 0)$	0.8127	0.8215	0.8094	$P(x_7^*(\tilde{\mathbf{c}}) = 0)$	0.8898	0.9186	0.8992
$P(x_2^*(\tilde{\mathbf{c}}) = 1)$	0.0094	0.0024	0.0176	$P(x_7^*(\tilde{\mathbf{c}}) = 1)$	0.0262	0.0009	0.0145
$P(x_2^*(\tilde{\mathbf{c}}) = 2)$	0.0214	0.0009	0.0236	$P(x_7^*(\tilde{\mathbf{c}}) = 2)$	0.0840	0.0806	0.0863
$P(x_2^*(\tilde{\mathbf{c}}) = 3)$	0.0302	0.0008	0.0191				
$P(x_2^*(\tilde{\mathbf{c}}) = 4)$	0.1263	0.1744	0.1303				
$P(x_3^*(\tilde{\mathbf{c}}) = 0)$	0.9020	0.9273	0.9097	$P(x_8^*(\tilde{\mathbf{c}}) = 0)$	0.9211	0.9292	0.9290
$P(x_3^*(\tilde{\mathbf{c}}) = 1)$	0.0094	0.0010	0.0065	$P(x_8^*(\tilde{\mathbf{c}}) = 1)$	0.0134	0.0008	0.0041
$P(x_3^*(\tilde{\mathbf{c}}) = 2)$	0.0887	0.0717	0.0838	$P(x_8^*(\tilde{\mathbf{c}}) = 3)$	0.0655	0.0701	0.0669
$P(x_4^*(\tilde{\mathbf{c}}) = 0)$	0.8519	0.8823	0.8756	$P(x_9^*(\tilde{\mathbf{c}}) = 0)$	0.7459	0.7012	0.7158
$P(x_4^*(\tilde{\mathbf{c}}) = 1)$	0.0302	0.0007	0.0123	$P(x_9^*(\tilde{\mathbf{c}}) = 1)$	0.1010	0.1168	0.0993
$P(x_4^*(\tilde{\mathbf{c}}) = 2)$	0.0169	0.0010	0.0235	$P(x_9^*(\tilde{\mathbf{c}}) = 2)$	0.0569	0.0790	0.0753
$P(x_4^*(\tilde{\mathbf{c}}) = 3)$	0.1010	0.1160	0.0886	$P(x_9^*(\tilde{\mathbf{c}}) = 3)$	0.0000	0.0005	0.0068
				$P(x_9^*(\tilde{\mathbf{c}}) = 4)$	0.0000	0.0002	0.0000
				$P(x_9^*(\tilde{\mathbf{c}}) = 10)$	0.0962	0.1023	0.1028
$P(x_5^*(\tilde{\mathbf{c}}) = 0)$	0.8328	0.8381	0.0260	$P(x_{10}^*(\tilde{\mathbf{c}}) = 0)$	0.8436	0.8392	0.8532
$P(x_5^*(\tilde{\mathbf{c}}) = 1)$	0.0490	0.0626	0.0189	$P(x_{10}^*(\tilde{\mathbf{c}}) = 1)$	0.0876	0.0848	0.0710
$P(x_5^*(\tilde{\mathbf{c}}) = 2)$	0.0199	0.0009	0.1047	$P(x_{10}^*(\tilde{\mathbf{c}}) = 2)$	0.0000	0.0004	0.0027
$P(x_5^*(\tilde{\mathbf{c}}) = 3)$	0.0983	0.0984	0.8559	$P(x_{10}^*(\tilde{\mathbf{c}}) = 5)$	0.0689	0.0757	0.0731

construct approximate solutions to the probability distribution of the optimal solution values. We show that the problem can be solved efficiently if we can characterize the convex hull of a mapping of the original polytope to a higher dimension 0-1 polytope. For simpler problems, the above methodology can be implemented in an efficient manner. For instance, with only the first two moments, the persistency model for a 0-1 problem reduces to a simple second-order conic program. For more general discrete optimization problems, we can often use the valid inequalities in the binary reformulation to generate cuts for the original problem. The computational results on discrete choice modeling problems with and without budget constraints provide encouraging evidence that the models can be used effectively to study discrete optimization problems under data uncertainty.

We believe that the approach opens new research avenues for using structural properties of convex optimization models to study general discrete optimization problems under data uncertainty. In particular, is there a way to efficiently compute $CH(\mathscr{Y})$ for special classes of discrete optimization problems? The connection between the persistency model and the discrete choice model also opens the possibility that a simple convex optimization model can be used to calibrate empirical data obtained in more complicated choice decision problems. We leave these issues to future research.

Acknowledgments
The authors thank Moshe E. Ben-Akiva for sharing the data on the transport choice model studied in Bierlaire et al. (2001) and the associate editor and two referees for their constructive comments. The research was supported in part by the Singapore-MIT Alliance, and part of the work was done when the third author was at SKK Graduate School of Business, Sungkyunkwan University, Seoul, Korea.

Appendix
Proof of Theorem 1. We know that formulation (6) provides an upper bound on Z^*. To show tightness, we construct an extremal distribution that attains the bound from the optimal $\mathbf{y}^*$ in formulation (6). Expressing $\mathbf{y}^*$ as a convex combination of the extreme points of the polytope $\mathscr{Y}$ (denoted as $\mathscr{P}$) implies that there exists a set of numbers λ_p^* such that

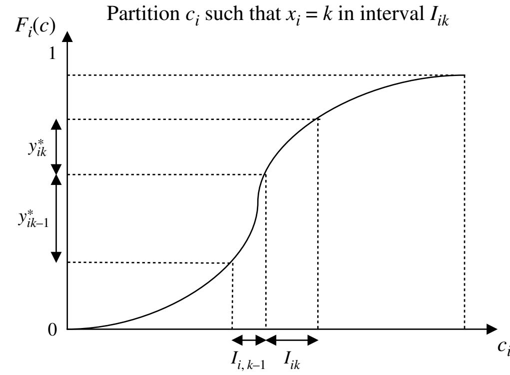
(i) $\lambda_p^* \geq 0$ for all $p \in \mathscr{P}$
(ii) $\sum_{p\in\mathscr{P}} \lambda_p^* = 1$
(iii) $y_{ik}^* = \sum_{p:\, y_{ik}[p]=1} \lambda_p^*$ for all $k \in \mathscr{X}_i$, $i \in \mathscr{N}$,

where $y_{ik}[p]$ denotes the value of the y_{ik} variable at the pth extreme point. We define the intervals:

$$I_{ik} = \left\{ c \,\middle|\, F_i^{-1}\left(\sum_{j\leq k-1} y_{ij}^*\right) \leq c \leq F_i^{-1}\left(\sum_{j\leq k} y_{ij}^*\right)\right\}. \tag{17}$$

We now generate the multivariate distribution θ^* as follows:
(i) Choose an extreme point p^* in $\mathscr{P}$ with probability λ_p^*
(ii) For each $i \in \mathscr{N}$ with $y_{ik}[p^*] = 1$, generate $\tilde{c}_i \sim f_i(c) \cdot \mathbb{I}(c \in I_{ik})/y_{ik}^*$ (see Figure A.1).

Figure A.1 **Constructing the Extremal Distribution for $\tilde{c}_i$**



Note that the cross dependency between the variables is not important here; hence, we can assume that for each fixed p, the distribution for $\tilde{c}_i$s are generated independently. Under this construction, if $f_i'(c)$ denotes the density function of $\tilde{c}_i$, then we have

$$\begin{aligned} f_i'(c) &= \sum_{k\in\mathscr{X}_i} \sum_{p:\, y_{ik}[p]=1} \lambda_p^* \frac{f_i(c)\mathbb{I}(c\in I_{ik})}{y_{ik}^*} \\ &= \sum_{k\in\mathscr{X}_i} f_i(c)\mathbb{I}(c \in I_{ik}) = f_i(c). \end{aligned}$$

Thus, the distribution θ^* satisfies the marginal distribution conditions and lies in Θ. Furthermore, under $\tilde{\mathbf{c}}$, if we simply pick the pth solution with probability λ_p^* instead of solving for $Z(\tilde{\mathbf{c}})$, we have

$$\begin{aligned} E_{\theta^*}(Z(\tilde{\mathbf{c}})) &\geq \sum_{p\in\mathscr{P}} \lambda_p^* \left(\sum_{i\in\mathscr{N}}\left(\sum_{k:\, y_{ik}[p]=1}\left(k\frac{\int c f_i(c)\mathbb{I}(c\in I_{ik})\,dc}{y_{ik}^*}\right)\right)\right) \\ &= \sum_{i\in\mathscr{N}}\left(\sum_{k\in\mathscr{X}_i}\left(\sum_{p:\, y_{ik}[p]=1}\lambda_p^*\right)\left(k\frac{\int c f_i(c)\mathbb{I}(c\in I_{ik})\,dc}{y_{ik}^*}\right)\right) \\ &= \sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i} k\int c f_i(c)\mathbb{I}(c\in I_{ik})\,dc \\ &= \sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i} k\int_{\sum_{j\leq k-1} y_{ij}^*}^{\sum_{j\leq k} y_{ij}^*} F_i^{-1}(t)\,dt. \end{aligned}$$

Because θ^* generates an expected optimal objective value that is greater than or equal to the optimal solution from formulation (6) and satisfies the marginal distributions, it attains Z^*.

To check the concavity of the objective function, we define a new variable Y_{ik} by the affine transformation $Y_{ik} = \sum_{j\leq k} y_{ij}$. The objective function can then be expressed as

$$\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{X}_i} k\int_{Y_{i,k-1}}^{Y_{ik}} F_i^{-1}(t)\,dt.$$

The first derivative of the objective function with respect to Y_{ik} is then given as

$$\begin{aligned} &\frac{\partial\left(k\int_{Y_{i,k-1}}^{Y_{ik}} F_i^{-1}(t)\,dt + (k+1)\int_{Y_{ik}}^{Y_{i,k+1}} F_i^{-1}(t)\,dt\right)}{\partial Y_{ik}} \\ &\quad = kF_i^{-1}(Y_{ik}) - (k+1)F_i^{-1}(Y_{ik}) = -F_i^{-1}(Y_{ik}). \end{aligned}$$

This is a decreasing function in Y_{ik} because F_i^{-1} is the inverse cumulative distribution function. Hence, the objective is concave in the Y_{ik} variables and therefore in the y_{ik} variables. □

Proof of Proposition 1. Clearly, formulation (8) is a restricted version of formulation (7) at the two extreme values α_i and β_i for each $i \in \mathcal{N}$. To show tightness, consider an optimal solution to formulation (7) denoted as y_{ik}^*. Define the variables for formulation (8) as

$$y_{i\alpha_i} = \sum_{k\in\mathcal{K}_i}\left(\frac{\beta_i - k}{\beta_i-\alpha_i}\right)y_{ik}^* \quad \text{and} \quad y_{i\beta_i} = \sum_{k\in\mathcal{K}_i}\left(\frac{k-\alpha_i}{\beta_i-\alpha_i}\right)y_{ik}^*.$$

This defines a feasible solution because

$$y_{i\alpha_i} + y_{i\beta_i} = \sum_{k\in\mathcal{K}_i} y_{ik}^* = 1 \quad \text{and} \quad \alpha_i y_{i\alpha_i} + \beta_i y_{i\beta_i} = \sum_{k\in\mathcal{K}_i} k y_{ik}^*.$$

Furthermore, we set

$$y_{i\alpha_i}(c) = \sum_{k\in\mathcal{K}_i}\left(\frac{\beta_i - k}{\beta_i-\alpha_i}\right)y_{ik}^*(c) \quad \text{and}$$

$$y_{i\beta_i}(c) = \sum_{k\in\mathcal{K}_i}\left(\frac{k-\alpha_i}{\beta_i-\alpha_i}\right)y_{ik}^*(c),$$

where $y_{ik}^*(c) = \mathbb{I}(c \in I_{ik})$ in (17). The objective function to the restricted formulation can be expressed as

$$\begin{aligned}
&\sum_{i\in\mathcal{N}}\left(\alpha_i \int c y_{i\alpha_i}(c) f_i(c)\,dc + \beta_i \int c y_{i\beta_i}(c) f_i(c)\,dc\right)\\
&= \sum_{i\in\mathcal{N}}\left(\alpha_i \sum_{k\in\mathcal{K}_i}\left(\frac{\beta_i-k}{\beta_i-\alpha_i}\right)\int c y_{ik}^*(c)\,dc\right.\\
&\qquad\left. + \beta_i \sum_{k\in\mathcal{K}_i}\left(\frac{k-\alpha_i}{\beta_i-\alpha_i}\right)\int c y_{ik}^*(c) f(c)\,dc\right)\\
&= \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}_i} k\int c y_{ik}^*(c) f(c)\,dc = \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}_i}\left(k\int_{\sum_{j\le k-1} y_{ij}^*}^{\sum_{j\le k} y_{ij}^*} F_i^{-1}(t)\,dt\right),
\end{aligned}$$

which proves the desired result. □

Proof of Proposition 2. For given values of $\mathbf{y} \in \mathcal{C}H(Y)$, computing Z^* in formulation (9) reduces to solving subproblems of the type

$$\max_{\mathbf{z}_i,\mathbf{w}_i}\left\{\sum_{k\in\mathcal{K}_i} k w_{ik}\colon \sum_{k\in\mathcal{K}_i} z_{ik} = \mu_i^2 + \sigma_i^2,\ \sum_{k\in\mathcal{K}_i} w_{ik} = \mu_i,\ z_{ik}y_{ik} \ge w_{ik}^2\right\}.$$

A relaxation to this subproblem is obtained from aggregating the last set of constraints:[2]

$$\max_{\mathbf{z}_i,\mathbf{w}_i}\left\{\sum_{k\in\mathcal{K}_i} k w_{ik}\colon \sum_{k\in\mathcal{K}_i} w_{ik} = \mu_i,\ \sum_{k\in\mathcal{K}_i}\frac{w_{ik}^2}{y_{ik}} \le \mu_i^2 + \sigma_i^2\right\}.$$

This relaxation is tight because we can generate an optimal solution to the original subproblem by setting $z_{ik} = w_{ik}^2/y_{ik} + \epsilon_{ik}$ with appropriate perturbations $\epsilon_{ik} \ge 0$ to ensure that the second moment constraint is met. Furthermore, this does not change the objective, which is independent of z_{ik}. Thus, the ith subproblem reduces to maximizing a linear objective over a linear equality constraint and a convex quadratic constraint. Introducing multipliers λ_i and ν_i for the constraints, the Karush-Kuhn-Tucker feasibility and optimality conditions for this problem are as follows:

(i) $\sum_{k\in\mathcal{K}_i} w_{ik} = \mu_i$ and $\sum_{k\in\mathcal{K}_i}(w_{ik}^2/y_{ik}) \le \mu_i^2+\sigma_i^2$
(ii) $-k + \lambda_i + 2\nu_i(w_{ik}/y_{ik}) = 0$ for all $k \in \mathcal{K}_i$
(iii) $\nu_i(\mu_i^2+\sigma_i^2 - \sum_{k\in\mathcal{K}_i}(w_{ik}^2/y_{ik})) = 0$
(iv) $\nu_i \ge 0$.

From condition (ii), we can assume that $\nu_i > 0$, or else the problem is trivial with only a single value for $k \in \mathcal{K}_i$. Substituting $w_{ik} = (k-\lambda)y_{ik}/2\nu_i$ into the feasibility constraints yields the optimal values:

$$\nu_i = \frac{\sqrt{\sum_{k\in\mathcal{K}_i} k^2 y_{ik} - \left(\sum_{k\in\mathcal{K}_i} k y_{ik}\right)^2}}{2\sigma_i},$$

$$\lambda_i = \sum_{k\in\mathcal{K}_i} k y_{ik} - 2\nu_i\mu_i,$$

$$w_{ik} = \mu_i y_{ik} + \sigma_i y_{ik}\left(\frac{k - \sum_{k\in\mathcal{K}_i} k y_{ik}}{\sqrt{\sum_{k\in\mathcal{K}_i} k^2 y_{ik} - \left(\sum_{k\in\mathcal{K}_i} k y_{ik}\right)^2}}\right).$$

The corresponding optimal objective value for the ith subproblem is

$$\mu_i \sum_{k\in\mathcal{K}_i} k y_{ik} + \sigma_i\sqrt{\sum_{k\in\mathcal{K}_i} k^2 y_{ik} - \left(\sum_{k\in\mathcal{K}_i} k y_{ik}\right)^2},$$

which proves the desired result. □

[2] We assume that all $y_{ik} > 0$. It is possible to extend this to allow for some of the $y_{ik} = 0$.

References

Adams, W. P., J. B. Lassiter, H. D. Sherali. 1998. Persistency in 0-1 polynomial programming. *Math. Oper. Res.* **23**(2) 359–389.

Ben-Akiva, M., M. Bierlaire. 1999. Discrete choice methods and their applications to short-term travel decisions. R. Hall, ed. *Handbook of Transportation Science. International Series in Operations Research and Management Science*, Vol. 23. Kluwer Academic, Boston, 5–34.

Ben-Akiva, M., S. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Bertsimas, D., K. Natarajan, C. P. Teo. 2004. Probabilistic combinatorial optimization: Moments, semidefinite programming and asymptotic bounds. *SIAM J. Optim.* **15**(1) 185–209.

Bertsimas, D., K. Natarajan, C. P. Teo. 2006. Persistence in discrete optimization under data uncertainty. *Math. Prog.* **108**(2) 251–274.

Bierlaire, M. 2003. BIOGEME: A free package for the estimation of discrete choice models. T. Chevroulet, A. Sevestre, eds. *Proc. 3rd Swiss Transportation Res. Conf., Ascona, Switzerland.*

Bierlaire, M., K. Axhausen, G. Abay. 2001. The acceptance of modal innovation: The case of Swissmetro. *Proc. 1st Swiss Transportation Res. Conf., Ascona, Switzerland.*

Lovász, L. 2001. Energy of convex sets, shortest paths, and resistance. *J. Comb. Theory, Ser. A* **94**(2) 363–382.

Lyons, R., R. Pemantle, Y. Peres. 1999. Resistance bounds for first-passage percolation and maximum flow. *J. Comb. Theory, Ser. A* **86**(1) 158–168.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. P. Zarembka, ed. *Frontiers in Econmetrics*. Academic Press, New York, 105–142.

Meilijson, I., A. Nadas. 1979. Convex majorization with an application to the length of critical path. *J. Appl. Probab.* **16** 671–677.

Nadas, A. 1979. Probabilistic PERT. *IBM J. Res. Dev.* **23** 339–347.

Sherali, H. D., W. P. Adams. 1999. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems.* Kluwer Academic Publishers, Dordrecht/Boston/London.

Spellucci, P. 1993. *DONLP2 Users Guide.* Department of Mathematics, Technical University at Darmstadt, Darmstadt, Germany.

Steele, J. M. 1997. *Probability Theory and Combinatorial Optimization*. *CBMS-NSF Reg. Conf. Ser. in Appl. Math.*, Vol 69. SIAM, Philadephia.

Steinberg, E., M. S. Parks. 1979. A preference order dynamic program for a knapsack problem with stochastic rewards. *J. Oper. Res. Soc.* **30**(2) 141–147.

Weiss, G. 1986. Stochastic bounds on distributions of optimal value functions with applications to PERT, network flows and reliability. *Oper. Res.* **34**(4) 595–605.

Vanderbei, R. 2006. LOQO user's manual—Version 4.05. Technical report, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, http://www.princeton.edu/~rvdb/tex/loqo/loqo405.pdf.