Statistica Sinica **17**(2007), 945-964

A FAST EM ALGORITHM FOR QUADRATIC OPTIMIZATION SUBJECT TO CONVEX CONSTRAINTS

Ming Tan[1], Guo-Liang Tian[1], Hong-Bin Fang[1] and Kai Wang Ng[2]

[1] *University of Maryland Greenebaum Cancer Center*
and [2] *The University of Hong Kong*

Abstract: *Convex constraints* (CCs) such as box constraints and linear inequality constraints appear frequently in statistical inference and in applications. The problems of *quadratic optimization* (QO) subject to CCs occur in isotonic regression, shape-restricted non-parametric regression, variable selection (via the lasso algorithm and bridge regression), limited dependent variables models, image reconstruction, and so on. Existing packages for QO are not generally applicable to CCs. Although EM-type algorithms may be applied to such problems (Tian, Ng and Tan (2005)), the convergence rate/speed of these algorithms is painfully slow, especially for high-dimensional data. This paper develops a fast EM algorithm for QO with CCs. We construct a class of data augmentation schemes indexed by a 'working parameter' r ($r \in \mathcal{R}$), and then optimize r over $\mathcal{R}$ under several convergence criteria. In addition, we use Cholesky decomposition to reduce both the number of latent variables and the dimension, leading to further acceleration of the EM. Standard errors of the restricted estimators are calculated using a non-parametric bootstrapping procedure. Simulation and comparison are performed and a complex multinomial dataset is analyzed to illustrate the proposed methods.

Key words and phrases: Bootstrap, Cholesky decomposition, constrained optimization, convergence rate, data augmentation, EM algorithm, latent variables, working parameter.

1. Introduction

Constrained optimization problems occur in many fields including operation research, econometrics and statistics. For instance, to construct a likelihood ratio statistic for testing nested hypotheses, it is necessary to find maximum likelihood estimates (MLEs) of parameters with linear equality constraints (LECs). Specifically, consider $H_0 : A\theta = b$ against $H_1 : A\theta \neq b$, where A and b are known, and θ is the parameter of interest. To compare the null to the full model, a common procedure is to use the likelihood ratio statistic $s = -2\{\ell(\hat{\theta}_N|Y_{\text{obs}}) - \ell(\hat{\theta}_F|Y_{\text{obs}})\}$, where $\ell(\theta|Y_{\text{obs}})$ denotes the log-likelihood and $\hat{\theta}_N$ ($\hat{\theta}_F$) is the MLE under the null (full) model. Under H_0, s follows an asymptotic χ^2 distribution whose degree of freedom is the number of additional parameters in the full model relative to the null model. Thus the key is to find the restricted MLE $\hat{\theta}_N$.

By using Lagrangian multipliers, the Newton-Raphson method can be used for optimization with LECs, through reformulating the problem into one of a penalized optimization. Utilizing this idea, Nyquist (1991) proposed iteratively reweighted least squares to estimate parameters subject to LECs in generalized linear models. For the Gaussian, Lange (1999, pp.184-185) obtained a closed-form solution for quadratic optimization (QO) with LECs. Kim and Taylor (1995) developed a so-called restricted EM incorporating the LECs on model parameters. They noted that maximizing $Q(\theta|\theta^{(t)})$ (the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimate) subject to LECs is much easier than maximizing $\ell(\theta|Y_{\text{obs}})$ subject to the same LECs. Lange (1999, pp.187-188) presented a method to calculate the asymptotic covariance matrix of an estimated parameter vector in the presence of LECs for the general log-likelihood $\ell(\theta|Y_{\text{obs}})$. On the other hand, Green (1990) and Silverman, Jones, Nychka and Wilson (1990) directly solved the penalized optimization in the EM framework for a more complicated penalty than that induced by LECs.

Linear inequality constraints (LICs) of the form $\{a \leq A\theta \leq b\}$ and convex constraints (CCs) also appear frequently in statistical applications. For example, proportions in logistic regression must be confined to the unit interval, cell probabilities in multinomial models are non-negative and sum to one. In addition, ellipsoid and quadratic constraints such as $\{(\theta_1, \theta_2)^\top \in I\!R^2 : \theta_1 \geq 0,\ \theta_2 \geq \theta_1^2 - 8\}$ are typical examples of CCs. Optimization problems with LICs and/or CCs occur frequently in isotonic regressions (Robertson, Wright and Dykstra (1988)), shape-restricted non-parametric regressions (Fraser and Massam (1989) and Meyer (1999)), variable selection via the non-negative garrote (Breiman (1995)) and the lasso (Tibshirani (1996)), growth curve models in biology, limited dependent variables models in econometrics (Hajivassiliou and McFadden (1998)), image reconstruction (Titterington (1985)), and so on. Constrained QO involves minimizing a multivariate quadratic function subject to LECs and LICs. Several algorithms including Hildreth's (1954) procedure and Wolfe's (1959) simplex method have been proposed to deal with the QO problems. The optimization toolbox in MATLAB includes subroutines for constrained QO based on three methods: trust-region (Coleman and Li (1996)), preconditioned conjugate gradient, and active set (Gill, Murray and Wright (1981)). SPLUS includes `nlminb` (non-linear minimization subject to box constraints), `nlregb` (non-linear least squares subject to box constraints) and `nnls.fit` (non-negative least squares). However, none of these methods and packages is applicable to QO with arbitrary CCs. In addition, even for LICs, the existing algorithms such as Hildreth's procedure or its generalization (Dykstra (1983)) are computationally cumbersome (see Sec. 6.2 below for more details).

Alternatively, the EM is a powerful and easy-to-implement algorithm for solving optimization problems with LICs and/or CCs. For instance, Vardi and Lee (1993) proposed an EM for solving linear inverse problems with positivity restrictions. Liu (2000) developed EM algorithms for finding MLEs in discrete distributions with a class of simplex constraints. Tan, Tian and Fang (2003) studied restricted MLEs in univariate normal distribution using EM-type algorithms, but they only considered the case of box constraints. In addition, EM algorithms were developed for estimating the mean vector in multivariate normal models with known/unknown covariance matrix subject to LICs (Shi, Zheng and Guo (2005) and box constraints or arbitrary CCs (Tian et al. (2005)).

However, a key hindrance to the application of EM-type algorithms is their slow convergence, especially in high-dimensional data such as image reconstruction. Recognizing that minimizing a quadratic function subject to CCs is the core computation in non-linear programming, we develop a fast EM algorithm to find least squares estimate (LSE) of θ in the following QO problems:

$$\hat{\theta} = \arg \min_{\theta \in \mathcal{S}(\theta)} \|y - X\theta\|^2, \tag{1.1}$$

where $y_{m\times 1}$ and $X_{m\times q}$ are known, and $\theta_{q\times 1}$ is restricted to some arbitrary convex region $\mathcal{S}(\theta)$, here $\mathcal{S}(\theta)$ can include the box/rectangle $[a, b] = \prod_{j=1}^{q}[a_j, b_j]$, and LICs of the form $\{\theta : a \le A\theta \le b\}$ as special cases. Notice that (1.1) can be viewed as finding the MLE of θ in the model

$$y \sim N_m(X\theta, I) \quad \text{subject to} \quad \theta \in \mathcal{S}(\theta). \tag{1.2}$$

To derive a fast EM, we first construct a class of *data augmentation* (DA) schemes indexed by a working parameter r, ($r \in \mathcal{R}$), and then find the optimal $r^{\mathrm{c}}_{\mathrm{opt}}$ by searching r over $\mathcal{R}$ under some convergence criterion c. In addition, we show further acceleration of the EM can be achieved using Cholesky decomposition to reduce latent variables and dimension.

The rest of this paper is organized as follows. Section 2 introduces a general framework of the EM for constrained parameter problems, and briefly summarizes existing EM algorithms. In Section 3, we construct a class of DA schemes, obtain an optimal EM using 'working parameter', and present theoretical results. In Section 4, we use Cholesky decomposition to further accelerate the optimal EM. Section 5 presents a non-parametric bootstrap approach to calculate standard errors. In Section 6, we apply the proposed methods to shape-restricted non-parametric regression with and without penalty, and then compare the proposed methods with existing algorithms via simulation. The proposed method is illustrated with an example in Section 7. All computations are performed on a Pentium IV workstation. We conclude with a discussion in Section 8.

2. EM Algorithms for Constrained Parameter Problems

Let Y_{obs} denote the observed data, θ the parameters, and $\ell(\theta|Y_{\text{obs}})$ the log-likelihood. Suppose that θ is restricted to a convex region $\mathcal{S}(\theta) \subseteq I\!R^q$, and that our aim is to maximize $\ell(\theta|Y_{\text{obs}})$ subject to $\theta \in \mathcal{S}(\theta)$. Usually, directly solving

$$\hat{\theta} = \arg \max_{\theta \in \mathcal{S}(\theta)} \ell(\theta|Y_{\text{obs}}) \tag{2.1}$$

is extremely difficult. If we augment the observed data Y_{obs} with missing data Y_{mis}, we have the complete data $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$. Then $\ell(\theta|Y_{\text{obs}}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)})$, where $Q(\theta|\theta^{(t)}) = E\{\log f(Y_{\text{com}}|\theta)|Y_{\text{obs}}, \theta^{(t)}\}$ and $H(\theta|\theta^{(t)}) = E\{\log f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)|Y_{\text{obs}}, \theta^{(t)}\}$. By Jensen's inequality (Dempster, Laird and Rubin (1977)), $\ell(\theta|Y_{\text{obs}}) - Q(\theta|\theta^{(t)}) = -H(\theta|\theta^{(t)}) \geq -H(\theta^{(t)}|\theta^{(t)}) = \ell(\theta^{(t)}|Y_{\text{obs}}) - Q(\theta^{(t)}|\theta^{(t)})$, $\forall \theta, \theta^{(t)} \in \mathcal{S}(\theta)$. That is, $\ell(\theta|Y_{\text{obs}}) - Q(\theta|\theta^{(t)})$ attains its minimum at $\theta = \theta^{(t)} \in \mathcal{S}(\theta)$. Thus, the EM algorithm has the ascent property that increasing $Q(\theta|\theta^{(t)})$ forces an increase in $\ell(\theta|Y_{\text{obs}})$. This ascent property implies that solving (2.1) is equivalent to iteratively solving

$$\theta^{(t+1)} = \arg \max_{\theta \in \mathcal{S}(\theta)} Q(\theta|\theta^{(t)}). \tag{2.2}$$

In many cases maximizing $Q(\theta|\theta^{(t)})$ under constraints is much simpler than maximizing $\ell(\theta|Y_{\text{obs}})$ under the same constraints. Generally, it is more common that a closed-form solution to (2.2) exists. If such a closed-form solution does not exist, the M-step may be split into several conditional M-steps so that the ECM algorithm (Meng and Rubin (1993)) can be applied. In principle, a straightforward EM algorithm for finding the LSE (1.1) exists (Tian et al. (2005)). Given $Y_{\text{obs}} = \{y_i\}_{i=1}^m$, they augmented Y_{obs} with $m(q-1)$ independent latent variables $\{Z_{ij}\}$ to obtain the complete-data $Y_{\text{com}} = \{Z_{ij} : 1 \leq i \leq m,\ 1 \leq j \leq q\}$, where

$$Z_{ij} \overset{\text{ind}}{\sim} N(x_{ij}\theta_j, \frac{1}{q}) \quad \text{and} \quad \Sigma_{j=1}^q Z_{ij} = y_i. \tag{2.3}$$

When $\mathcal{S}(\theta) = \prod_{j=1}^q [a_j, b_j]$, given $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_q^{(t)})^\top$, the E-step calculates

$$T_j^{(t)} = \theta_j^{(t)} + (q \cdot x_j^\top x_j)^{-1} \cdot x_j^\top [y - X\theta^{(t)}], \quad 1 \leq j \leq q, \tag{2.4}$$

where x_j denotes the jth column of X. The M-step updates

$$\theta_j^{(t+1)} = \min\{\max\{a_j,\ T_j^{(t)}\},\ b_j\}, \quad 1 \leq j \leq q. \tag{2.5}$$

They also showed that the LICs of the form $\{a \leq A\theta \leq b\}$ can be reduced to the case of box constraints via a linear transformation and considered CCs.

3. Accelerating EM via Optimizing Working Parameter

Although the EM based on (2.4) and (2.5) (denoted $\mathrm{EM}^{\mathrm{TNT}}$ in the sequel) is simple and easy to implement because of explicit expressions in both the E- and M-steps, it may converge rather slowly in some applications, e.g., in image reconstruction (Titterington (1985)) where X is a very large design matrix ($m \times m$, with $m = 512$ or more). Slow convergence is due to the introduction of too many latent variables and unreasonable allocation of variances. We show ways to speed up the $\mathrm{EM}^{\mathrm{TNT}}$ while maintaining its simplicity and stability (i.e., automatic monotone convergence in log-likelihood).

3.1. Criteria for accelerating EM

The acceleration of EM is closely related to the 'working parameter' idea of Meng and van Dyk (1997). Let $Y_{\mathrm{com}}(r) = \{Y_{\mathrm{obs}}, Y_{\mathrm{mis}}(r)\}$ denote the complete (or augmented) data and r the working parameter belonging to some set $\mathcal{R}$. For each $r \in \mathcal{R}$, $Y_{\mathrm{com}}(r)$ is a legitimate DA, which induces an EM algorithm with the theoretical matrix rate of convergence, denoted by $M(r)$. As a function of the working parameter r, the matrix rate of convergence is $M(r) = I_q - I_{\mathrm{com}}^{-1}(r) I_{\mathrm{obs}}$, where I_q is the identity matrix,

$$I_{\mathrm{com}}(r) = E\left[-\frac{\partial^2 \log f(Y_{\mathrm{com}}(r)|\theta)}{\partial\theta\partial\theta^{\top}} \middle| Y_{\mathrm{obs}}, \theta \right] \Bigg|_{\theta=\hat{\theta}} \tag{3.1}$$

is the expected complete-data information matrix, and $I_{\mathrm{obs}} = -\partial^2 \log f(Y_{\mathrm{obs}}|\theta)/\partial\theta\partial\theta^{\top}|_{\theta=\hat{\theta}}$ is the observed information matrix.

Meng (1994) defined the matrix speed of convergence for an EM algorithm by $S(r) = I_q - M(r) = I_{\mathrm{com}}^{-1}(r) I_{\mathrm{obs}}$. The goal is to optimize r over $\mathcal{R}$ by maximizing $S(r)$. Since I_{obs} does not depend on r, it suffices to optimize r by minimizing $I_{\mathrm{com}}(r)$. Let $c\{I_{\mathrm{com}}(r)\}$ denote a criterion for measuring the size of the positive semidefinite matrix $I_{\mathrm{com}}(r)$, so

$$r_{\mathrm{opt}}^{\mathrm{c}} = \arg\min_{r\in\mathcal{R}} c\{I_{\mathrm{com}}(r)\}. \tag{3.2}$$

The commonly used criteria are the largest eigenvalue, determinant, or trace. Different critera c will result in different $r_{\mathrm{opt}}^{\mathrm{c}}$. The largest eigenvalue $\rho\{M(r)\}$ of $M(r)$ is known as the global rate of convergence of the EM. The smallest eigenvalue $s\{S(r)\} = 1 - \rho\{M(r)\}$ of $S(r)$ is called the global speed of the algorithm. As shown in Meng and van Dyk (1997), if $I_{\mathrm{com}}(r) \geq I_{\mathrm{com}}(r')$, then $s\{S(r)\} \leq s\{S(r')\}$.

3.2. A class of DA schemes

To accelerate the $\mathrm{EM}^{\mathrm{TNT}}$, we first erect a class of DA schemes to find the

LSE (1.1). Since $X_{m\times q} = (x_{ij})$ is known, we define two index sets

$$\mathcal{J}_i \equiv \{j : x_{ij} \neq 0\},\ 1 \le i \le m, \quad \text{and} \quad \mathcal{I}_j \equiv \{i : x_{ij} \neq 0\},\ 1 \le j \le q. \tag{3.3}$$

Further for a fixed i, we define weights

$$\lambda_{ij} \equiv \frac{|x_{ij}|^r}{\Sigma_{j'\in\mathcal{J}_i}|x_{ij'}|^r}, \qquad j \in \mathcal{J}_i, \quad r \in I\!R_+ = \{r : r \ge 0\}. \tag{3.4}$$

Obviously, $\lambda_{ij} > 0$ and $\sum_{j\in\mathcal{J}_i}\lambda_{ij} = 1$. In particular, when $r = 0$, $\lambda_{ij} = 1/n_i$, where $n_i = \#\{\mathcal{J}_i\}$ denotes the number of elements in $\mathcal{J}_i$. When $r = +\infty$, from (3.4), we have

$$\lambda_{ij} = \left(\frac{|x_{ij}|}{[\Sigma_{j'\in\mathcal{J}_i}|x_{ij'}|^r]^{1/r}}\right)^r = \lim_{r\to\infty}\left(\frac{|x_{ij}|}{\max_{j'\in\mathcal{J}_i}|x_{ij'}|}\right)^r = \begin{cases} 1, & \text{if } j = j_0, \\ 0, & \text{otherwise,} \end{cases}$$

where $j_0 \in \mathcal{J}_i$ is such that $|x_{ij_0}| = \max_{j'\in\mathcal{J}_i}|x_{ij'}|$.

We augment the observed data $Y_{\text{obs}} = \{y_i\}_{i=1}^m$ with independent latent data $\{Z_{ij}(r)\}$ to obtain a class of complete (or augmented) data $Y_{\text{com}}(r) = \{Z_{ij}(r) : 1 \le i \le m,\ j \in \mathcal{J}_i\}$ indexed by a working parameter r, where

$$Z_{ij}(r) \overset{\text{ind}}{\sim} N(x_{ij}\theta_j, \lambda_{ij}) \quad \text{and} \quad \Sigma_{j\in\mathcal{J}_i}Z_{ij}(r) = y_i. \tag{3.5}$$

Thus, the multiple DA schemes $\{Y_{\text{com}}(r)\}_{r\in I\!R_+}$ induce a class of EM algorithms when the working parameter r varies in $I\!R_+$.

Several crucial differences exist between the single (or fixed) DA scheme in (2.3) and the multiple (or flexible) DA schemes in (3.5). Firstly, the former introduces a total of $m(q-1)$ latent variables while the latter adds only $m(n_i-1)$ latent variables with $n_i = \#\{\mathcal{J}_i\} \le q$. The fewer the number of latent variables, the faster the induced EM converges. Secondly, in (3.5), λ_{ij} controls the variance of $Z_{ij}(r)$, and is more flexible than setting the variance of Z_{ij} in (2.3) to a constant. When all $n_i = q$ and $r = 0$, the latter reduces to the former. Therefore, we can identify an optimal r_{opt} so that the optimal EM has the fastest convergence among the whole EM class. Thirdly, note that in (2.3), when $x_{ij} = 0$, Z_{ij} (with mean zero and variance $1/q$) does not contribute to estimating θ_j. For such situations, we should set $Z_{ij} = 0$, which amounts to introducing $Z_{ij} \sim N(0, \lambda_{ij})$ with $\lambda_{ij} = 0$. Thus, $\lambda_{ij} \propto |x_{ij}|^r$ is a natural choice, leading intuitively to (3.4).

3.3. A class of EM algorithms

In (3.5), for a fixed r in $I\!R_+$, the complete-data log-likelihood function is given by

$$\log f(Y_{\text{com}}(r)|\theta) = -0.5\Sigma_{j=1}^q\Sigma_{i\in\mathcal{I}_j}(Z_{ij}(r) - x_{ij}\theta_j)^2(\lambda_{ij})^{-1}, \tag{3.6}$$

and the surrogate function is $Q(\theta|\theta^{(t)}) = -0.5\Sigma_{j=1}^{q}\Sigma_{i\in\mathcal{I}_j}E\{(Z_{ij}(r) - x_{ij}\theta_j)^2|Y_{\text{obs}}, \theta^{(t)}\}/\lambda_{ij}$. Define $u_j(r) \equiv v_j^2(r)\cdot\Sigma_{i\in\mathcal{I}_j}x_{ij}Z_{ij}(r)/\lambda_{ij}$, where

$$v_j^2(r) \equiv \Big\{\Sigma_{i\in\mathcal{I}_j}\Big(\frac{x_{ij}^2}{\lambda_{ij}}\Big)\Big\}^{-1}, \quad 1 \le j \le q. \tag{3.7}$$

From (3.6), the sufficient statistic for θ_j is $u_j(r)$. To calculate the conditional expectation of $u_j(r)$, we first compute $E[Z_{ij}(r)|Y_{\text{obs}}, \theta^{(t)}] = x_{ij}\theta_j^{(t)} + [y_i - x_{(i)}^{\top}\theta^{(t)}]\lambda_{ij}$, where $x_{(i)}^{\top}$ denotes the ith row of matrix X. Let $X = (x_1, \ldots, x_q)$, then the E-step of the EM is to compute $T_j^{(t)}(r) = E[u_j(r)|Y_{\text{obs}}, \theta^{(t)}] = \theta_j^{(t)} + v_j^2(r)\cdot x_j^{\top}[y - X\theta^{(t)}]$ or, equivalently in vector form,

$$T^{(t)}(r) = \theta^{(t)} + \text{diag}(v_1^2(r), \ldots, v_q^2(r))\cdot X^{\top}[y - X\theta^{(t)}]. \tag{3.8}$$

The M-step updates (for $1 \le j \le q$),

$$\theta_j^{(t+1)} = T_j^{(t)}(r), \qquad \text{if } \mathcal{S}(\theta) = I\!R^q, \tag{3.9}$$

$$\theta_j^{(t+1)} = \min\{\max\{a_j,\ T_j^{(t)}(r)\},\ b_j\}, \qquad \text{if } \mathcal{S}(\theta) = [a, b], \tag{3.10}$$

$$\theta_j^{(t+1)} = \min\{\max\{L_j(\theta_{-j}^{(t)}),\ T_j^{(t)}(r)\},\ U_j(\theta_{-j}^{(t)})\}, \text{ if } \mathcal{S}(\theta) \text{ is a convex set.} \tag{3.11}$$

In (3.11), we assume that $\mathcal{S}(\theta)$ is available with one-dimensional slices, in the sense that $\mathcal{S}_j(\theta_j|\theta_{-j}) = \{\theta_j : \theta \in \mathcal{S}(\theta)\}$ can be represented as intervals $[L_j(\theta_{-j}), U_j(\theta_{-j})]$, where θ_{-j} denotes the $(q-1)$-dimensional subvector of θ by deleting the jth component θ_j.

When $\text{rank}(X_{m\times q}) = q \le m$, the EM based on (3.8) and (3.9) is a novel fast iterative method for calculating the unconstrained LSE requiring no matrix inversion. More importantly, for instance in image reconstruction, the matrix X is often ill-conditioned and results in an unstable LSE, while our EM provides a stable solution. Furthermore, the EM converges to the unique solution (1.1). In practice, the initial values $\theta^{(0)}$ can be taken as the unconstrained LSE, or an arbitrary point in the box $[a, b]$ or $\mathcal{S}(\theta)$.

3.4. Optimal and uniformly optimal working parameters

3.4.1. Largest eigenvalue, determinant and trace criteria

The speed of convergence of the sequences $\{T^{(t)}(r)\}_{t=0}^{\infty}$ in (3.8) depends on the working parameter r. For a given criterion c, we need to determine the optimal $r_{\text{opt}}^{\text{c}}$ in (3.2). From (3.1), (3.7) and (3.2), we obtain $I_{\text{com}}(r) = \text{diag}(1/v_1^2(r), \ldots, 1/v_q^2(r))$,

$$r_{\text{opt}}^{\rho} = \arg\min_{r\in I\!R_+}\ \rho\{I_{\text{com}}(r)\} \quad = \arg\min_{r\in I\!R_+}\ \max_{1\le j\le q}\{\Sigma_{i\in\mathcal{I}_j}\frac{x_{ij}^2}{\lambda_{ij}}\}, \tag{3.12}$$

$$r_{\text{opt}}^{\text{det}} = \arg \min_{r \in I\!R_+} \det\{I_{\text{com}}(r)\} = \arg \min_{r \in I\!R_+} \Sigma_{j=1}^{q} \log[\Sigma_{i \in \mathcal{I}_j} \frac{x_{ij}^2}{\lambda_{ij}}], \tag{3.13}$$

$$r_{\text{opt}}^{\text{tr}} = \arg \min_{r \in I\!R_+} \text{tr}\{I_{\text{com}}(r)\} = \arg \min_{r \in I\!R_+} \Sigma_{i=1}^{m} \Sigma_{j \in \mathcal{J}_i} \frac{x_{ij}^2}{\lambda_{ij}}. \tag{3.14}$$

In practice, a single r_{opt} is often prefered. We suggest selecting one so that the target function $||y - X\theta||^2$ decreases (against the EM iteration) the fastest. Since (3.12), (3.13) and (3.14) depend on X, it is generally difficult to obtain a uniformly optimal working parameter for an arbitrary design matrix X. However, this may sometimes be achieved as shown in the following three simple examples.

Example 1. Let $X_{2\times 2} = (x_1, x_2)$ with $x_1 = (-1, 0)^\top$ and $x_2 = (2, -6)^\top$. Then (3.12), (3.13) and (3.14) yield $r_{\text{opt}}^{\rho} = 5.285$, $r_{\text{opt}}^{\text{det}} = 0$ and $r_{\text{opt}}^{\text{tr}} = 1$, respectively.

Example 2. Let $X_{3\times 2} = (x_1, x_2)$ with $x_1 = (0, 0, 3)^\top$ and $x_2 = (-2, 1, 6)^\top$. Then $r_{\text{opt}}^{\rho} = 2.155$, $r_{\text{opt}}^{\text{det}} = 0$ and $r_{\text{opt}}^{\text{tr}} = 1$.

Example 3. Let $X_{4\times 5} = (x_1, \ldots, x_5)$ with $x_1 = (1, 2, 3, 4)^\top$, $x_2 = (5, 6, 7, 8)^\top$, $x_3 = (9, 10, 11, 12)^\top$, $x_4 = (13, 14, 15, 16)^\top$ and $x_5 = (17, 18, 19, 20)^\top$. Then $r_{\text{opt}}^{\rho} = 2$, $r_{\text{opt}}^{\text{det}} = 0.069$ and $r_{\text{opt}}^{\text{tr}} = 1$.

3.4.2. Uniformly optimal working parameter for the trace criterion

In the three examples above, $r_{\text{opt}}^{\text{tr}} = 1$. This leads us to believe some uniformly optimal result might exist for the trace criterion. We first prove the following inequality.

Lemma 1. *Let* $\{\lambda_{ij}\}$ *be given by* (3.4), *then* $\Sigma_{j \in \mathcal{J}_i}(x_{ij}^2/\lambda_{ij}) \geq (\Sigma_{j \in \mathcal{J}_i}|x_{ij}|)^2$ *for any* $\{x_{ij}\}$ *and any* $r \in I\!R_+$, $1 \leq i \leq m$, *and equality holds if and only if* $r = 1$.

Proof. By the Cauchy-Schwartz inequality, for any two non-zero vectors ξ and η, $\xi^\top \xi \cdot \eta^\top \eta \geq (\xi^\top \eta)^2$, and equality holds if and only if there exists some non-zero constant scalar c such that $\xi = c\eta$. Now let $\xi_j = |x_{ij}|^{r/2}$ and $\eta_j = |x_{ij}|^{1-r/2}$, to get $\sum_{j \in \mathcal{J}_i} |x_{ij}|^r \cdot \sum_{j \in \mathcal{J}_i} |x_{ij}|^{2-r} \geq (\sum_{j \in \mathcal{J}_i} |x_{ij}|)^2$.

Now we consider the trace of $I_{\text{com}}(r)$. By combining (3.14) with Lemma 1, we obtain $\text{tr}\{I_{\text{com}}(r)\} \geq \sum_{i=1}^{m} (\sum_{j \in \mathcal{J}_i} |x_{ij}|)^2 = \text{tr}\{I_{\text{com}}(1)\}$, namely $r_{\text{opt}}^{\text{tr}} = 1$. Hence, we have

Theorem 1. *For the trace criterion,* $r_{\text{opt}}^{\text{tr}} = 1$ *for any covariate matrix* X.

This theorem shows that the trace criterion is invariant for any linear transformation of X. However, neither the largest eigenvalue nor the determinant criterion has this kind of property. The following example illustrates this assertion.

Example 4. Let $X = (x_1, x_2, x_3)$ with $x_1 = (79, 0, 18)^\top$, $x_2 = (90, 100, 6)^\top$ and $x_3 = (0, 3, 86)^\top$, then $r_{\text{opt}}^{\rho} = 1.06$ and $r_{\text{opt}}^{\text{det}} = 1.06$. The singular value

decomposition yields $X^\top = UDV^\top$, where $V = (v_1, v_2, v_3)$ is orthogonal with $v_1 = (-0.436, -0.895, -0.098)^\top$, $v_2 = (0.165, -0.186, 0.969)^\top$ and $v_3 = (0.885, -0.406, -0.228)^\top$. Thus we have $||y - X\theta||^2 = ||y - V\mu||^2$ with $\mu = DU^\top\theta$. Replacing X by the linear transformation $V = X(DU^\top)^{-1}$, we obtain $r^\rho_{\text{opt}} = 0.992$ and $r^{\text{det}}_{\text{opt}} = 0.996$.

3.4.3. A special class of covariate matrices

Consider a special class of covariate matrices where all absolute values of non-zero elements in each row are equal and there is at least one non-zero element in each row and each column. We denote it by $\mathcal{M}_{m\times q}$. For example, let $X^\top = (x_{(1)}, \ldots, x_{(4)})$, where $x_{(1)} = (0, 2, -2, 0)^\top$, $x_{(2)} = (-1, 0, -1, 1)^\top$, $x_{(3)} = (3, 3, -3, 3)^\top$ and $x_{(4)} = (4, 0, -4, 0)^\top$, then $X \in \mathcal{M}_{4\times 4}$. We have the following result.

Theorem 2. *If $X \in \mathcal{M}_{m\times q}$, then any non-negative real number can serve as a uniformly optimal working parameter for all three criteria, i.e., $r^\rho_{\text{opt}} = r^{\text{det}}_{\text{opt}} = r^{\text{tr}}_{\text{opt}} = r$.*

Proof. From the definition of $\mathcal{M}_{m\times q}$, we have $\mathcal{J}_i \neq \emptyset$, $\mathcal{I}_j \neq \emptyset$ and $n_i = \#\{\mathcal{J}_i\} \geq 1$. In addition, (3.4) and (3.7) yield $\lambda_{ij} = 1/n_i$ for any $j \in \mathcal{J}_i$ and $1/v_j^2(r) = \sum_{i\in\mathcal{I}_j} n_i\, x_i^{*2}$, where x_i^* denotes the non-zero element at the ith row of X. Since $1/v_j^2(r)$ does not depend on r, $I_{\text{com}}(r)$ does not depend on r.

3.4.4. A sub-optimal working parameter for all three criteria

For any given X, when $r = 0$, (3.4) yields $\lambda_{ij} = 1/n_i$ with $n_i = \#\{\mathcal{J}_i\}$. It follows from (3.7) that $1/v_j^2(0) = x_j^\top \mathbf{N} x_j$ with $\mathbf{N} \equiv \text{diag}(n_1, \ldots, n_m)$. Therefore (3.8) becomes

$$T_j^{(t)}(0) = \theta_j^{(t)} + x_j^\top[y - X\theta^{(t)}]/x_j^\top\mathbf{N}x_j, \quad j = 1, \ldots, q. \tag{3.15}$$

We denote the EM algorithms based on (3.8) and (3.10) by $\text{EM}^{\text{c}}(r)$. By comparing (2.4) with (3.15), since $q\cdot x_j^\top x_j \geq x_j^\top\mathbf{N}x_j$, we have $|T_j^{(t)} - \theta_j^{(t)}| \leq |T_j^{(t)}(0) - \theta_j^{(t)}|$. This implies the convergence speed of $\text{EM}^{\text{c}}(0)$ is faster than that of EM^{TNT}. Based on this fact, we call $r = 0$ the sub-optimal working parameter for all three criteria. We summarize these results in the following theorem.

Theorem 3. *By the criteria of largest eigenvalue, determinant and trace, Speed of $EM^{\text{c}}(0) \geq$ Speed of EM^{TNT}.*

4. Further Acceleration via Cholesky Decomposition

Let $\text{rank}(X_{m\times q}) = q \leq m$ and suppose both q and m are large. By the Cholesky decomposition, there exists a unique upper triangular matrix $B_{q\times q}$

with positive diagonal elements such that $X^\top X = B^\top B$. Thus, on the one hand, (1.1) is equivalent to

$$\hat{\theta} = \arg\ \min_{\theta \in \mathcal{S}(\theta)} ||\xi - B\theta||^2, \tag{4.1}$$

where $\xi_{q\times 1} = (B^\top)^{-1} X^\top y$ and there are at least $q(q-1)/2$ zero entries in B. On the other hand, the constrained quadratic optimization with high-dimensional design matrix (m observations, q variables) is already reduced to an equivalent optimization but with lower-dimensional design matrix (q observations, q variables). In addition, Theorem 1 shows that both (1.1) and (4.1) have the same optimal working parameter ($r^{tr}_{\rm opt} = 1$) under the criterion of trace.

Example 5. *High-dimensional simulations and comparisons.* We simulated several large sets of data to compare the convergence rates of the three EMs. The ε_i and x_{ij} were i.i.d. $N(0,1)$, the θ_j were i.i.d. $U[-2,2]$, and $y_i = x^\top_{(i)}\theta + \varepsilon_i$ for $i = 1,\ldots,m$ and $j = 1,\ldots,q$. We minimized $||y - X\theta||^2$ subject to $\theta \geq 0$ by using three algorithms: $\mathrm{EM}^{\rm TNT}_X$ based on (2.4) and (2.5) and design matrix X, $\mathrm{EM}^{\rm tr}_X(1)$ based on (3.8) and (3.10) and X with $r^{\rm tr}_{\rm opt} = 1$, and $\mathrm{EM}^{\rm tr}_B(1)$ based on B in (4.1). For instance, the second row of Table 1 shows that 545 (318) iterations are needed for $\mathrm{EM}^{\rm TNT}_X$ ($\mathrm{EM}^{\rm tr}_X(1)$) to have the same precision as $\mathrm{EM}^{\rm tr}_B(1)$, which converged at 100th iteration in 0.65 second. We obtained results for 100 replicates. For $m = 1{,}000$ and $q = 500$, Table 1 shows that $\mathrm{EM}^{\rm tr}_B(1)$ is about 7 (or 5) times faster than $\mathrm{EM}^{\rm TNT}_X$ (or $\mathrm{EM}^{\rm tr}_X(1)$).

Table 1. Convergence speed for different EM algorithms for large simulated data.

# Obs m	# Variables q	# Iteration for $\mathrm{EM}^{\rm TNT}_X$	# Iteration for $\mathrm{EM}^{\rm tr}_X(1)$	# Iteration and time for $\mathrm{EM}^{\rm tr}_B(1)$	# Replicate for simulation
100	50	545	318	100 (0.65 sec)	100
500	100	922	468	100 (1.14 sec)	100
1,000	500	3,457	2,586	500 (41.1 sec)	100
2,000	1,000	7,480	5,662	1,000 (8.12 min)	50
3,000	1,500	12,426	10,485	1,500 (32.5 min)	20

5. Standard Errors

Utilizing the fast EM algorithm, the standard errors of $\hat{\theta}$ defined in (1.1) can be obtained with a non-parametric bootstrap approach. Let $y_i = x^\top_{(i)}\theta + \varepsilon_i$, $i = 1,\ldots,m$, where $x^\top_{(i)}$ denotes the ith row of the covariate matrix $X_{m\times q}$, and the error terms $\{\varepsilon_i\}$ are assumed to be a random sample from an unknown distribution F having expectation zero. Since $\hat{\theta}$ is available, e.g., via some $\mathrm{EM}^{\rm c}(r^{\rm c}_{\rm opt})$, we can calculate $\hat{\varepsilon}_i = y_i - x^\top_{(i)}\hat{\theta}$ for each i. The obvious estimate of F is the empirical distribution of $\{\hat{\varepsilon}_i\}$, denoted by $\hat{F}$. Thus we can generate a random sample of bootstrap error terms, denoted by $\{\varepsilon^*_i\}^m_{i=1}$, where each ε^*_i equals any

one of the m values $\hat{\varepsilon}_i$ with probability $1/m$. Then the bootstrap responses are generated by

$$y_i^* = x_{(i)}^\top \hat{\theta} + \varepsilon_i^*, \quad i = 1, \ldots, m, \quad \text{or equivalently,} \quad y^* = X\hat{\theta} + \varepsilon^*, \tag{5.1}$$

where $y^* = (y_1^*, \ldots, y_m^*)^\top$ and $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_m^*)^\top$. Notice that the covariate matrix X is the same for the bootstrap data as for the actual data, and $\hat{\theta}$ is a fixed quantity in (5.1). Having obtained y^* the bootstrap replication is given by $\hat{\theta}^* = \arg\min_{\theta \in \mathcal{S}(\theta)} ||y^* - X\theta||^2$. Independently repeating the above process G times, we obtain G bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$ with $\hat{\theta}^*(g) = (\hat{\theta}_1^*(g), \ldots, \hat{\theta}_q^*(g))^\top$ and the standard error $\mathrm{se}\,(\hat{\theta}_j)$ of $\hat{\theta}_j$ can be estimated by the sample standard deviation of the G replications.

6. Application, Simulation and Comparison

We apply the proposed methods to shape-restricted non-parametric regression with and without penalty and compare them with the existing EM algorithm and Dykstra's (1983) algorithm via simulation.

6.1. Shape-restricted non-parametric regression

Given $z_0 < z_1 < \cdots < z_m < z_{m+1}$ and the observed data $Y_{\text{obs}} = \{\xi_i\}_{i=1}^m$, we consider the non-parametric regression model

$$\xi_i = f(z_i) + e_i, \qquad i = 1, \ldots, m, \tag{6.1}$$

where $\{e_i\}_{i=1}^m$ are random errors with mean zero. The goal is to estimate f subject to shape constraints (e.g., monotonicity, convexity or concavity). These constraints can be expressed as a set of LICs and written in the form $A\mu \geq 0$, where $A_{p\times m}$ is a matrix depending on z_i and $\mu = (\mu_1, \ldots, \mu_m)^\top = (f(z_1), \ldots, f(z_m))^\top$. Therefore, the problem is reduced to minimizing $\Sigma_{i=1}^m w_i(y_i - \mu_i)^2$ with known weights $\{w_i\}$ subject to LICs, i.e.,

$$\hat{\mu} = \arg\min_{A\mu \geq 0} (\xi - \mu)^\top W (\xi - \mu), \tag{6.2}$$

where $\xi = (\xi_1, \ldots, \xi_m)^\top$ and $W = \mathrm{diag}(w_1, \ldots, w_m)$.

We consider three cases and make the following linear transformations

$$\mu = \begin{cases} A^{-1}\theta & \text{with } \theta_{p\times 1} \in I\!R_+^p, & \text{if } p = m \text{ and } A^{-1} \text{ exists,} \\ (A^\top A)^{-1}A^\top\theta & \text{with } \theta_{p\times 1} \in I\!R_+^p, & \text{if } \mathrm{rank}(A) = m < p, \\ (A^*)^{-1}\theta & \text{with } \theta_{m\times 1} \in I\!R_+^p \times I\!R^{m-p}, & \text{if } \mathrm{rank}(A) = p < m, \end{cases} \tag{6.3}$$

where $A^* = \begin{pmatrix} A_1 & A_2 \\ O & I_{m-p} \end{pmatrix}$, $A = (A_1, A_2)$ with $A_1 : p \times p$ and $A_2 : p \times (m-p)$. Thus (6.3) can be treated as $\mu = B_{m\times q}\theta_{q\times 1}$ with $\theta \in [a, b]$, in a unified manner. Finding $\hat{\mu}$ in (6.2) is equivalent to computing

$$\hat{\theta} = \arg \min_{\theta\in[a,b]} ||y - X\theta||^2, \tag{6.4}$$

where $y = W^{1/2}\xi$ and $X = W^{1/2}B$.

Usually, the shape-restricted regression function estimated via (6.2) is not very smooth. This difficulty can be overcome by adding a roughness penalty, and the goal is to find

$$\hat{\mu} = \arg \min_{A\mu\geq 0} \left\{ (\xi - \mu)^\top W(\xi - \mu) + \gamma \int_{z_0}^{z_{m+1}} f''(z)^2 \, d\, z \right\}, \tag{6.5}$$

where $\gamma > 0$ is a smoothing parameter. When the LICs are absent, (6.5) is the natural cubic smoothing spline (Lange (1999, Chap.9)). Suppose that f is a natural cubic spline, i.e., $f''(z)$ is piecewise linear and continuous on $[z_0, z_{m+1}]$, and $f'' = 0$ on $[z_0, z_1]$ and $[z_m, z_{m+1}]$. Meyer (1999) showed that the integral term in (6.5) can be written as $\mu^\top P\mu$, where $P \geq 0$. Therefore, (6.5) becomes $\hat{\mu} = \arg\min_{A\mu\geq 0} (Q^{-1}W\xi - \mu)^\top Q(Q^{-1}W\xi - \mu)$ with $Q = W + \gamma P$, which is of the form (6.2).

6.2. Dykstra's algorithm

By using the concept of duality and the Gauss-Seidel computational algorithm, Hildreth (1954) studied QO with LICs like (6.2). His procedure rests on the duality theorem (Kuhn and Tucker (1951, p.487, pp.491-492) and Hildreth (1954, p.604)). Wolfe (1959) proposed a simplex method to solve the QO problem with LECs/LICs. Dykstra (1983) extended Hildreth's procedure to find the projection of a point onto a finite intersection of closed convex cones.

To solve (6.2), we define $y = W^{1/2}\xi$, $\theta = W^{1/2}\mu$ and $X^\top = AW^{-1/2}$, then $\hat{\mu} = W^{-1/2}\hat{\theta}$ with $\hat{\theta} = \arg\min_{X^\top\theta\geq 0} ||y-\theta||^2$. Following the notations of Dykstra (1983), let $X_{m\times p} = (x_1, \ldots, x_p)$, $\mathcal{C}_j = \{\theta \in I\!R^m : x_j^\top\theta \geq 0\}$ $(j = 1, \ldots, p)$, and $\mathcal{C} = \cap_{j=1}^p \mathcal{C}_j$. Then

$$\hat{\theta} = P(y|\mathcal{C}) = \arg \min_{\theta\in\mathcal{C}} ||y - \theta||^2, \tag{6.6}$$

where $P(y|\mathcal{C})$ is called the projection of y onto $\mathcal{C}$ for any given $y \in I\!R^m$. Let $P(y|\mathcal{C}_j)$ denote the projection of y onto $\mathcal{C}_j$, then $P(y|\mathcal{C}_j) = y$ if $x_j^\top y \geq 0$ and is $y - x_j \cdot (x_j^\top y / x_j^\top x_j)$ otherwise. Dykstra (1983) proposed the following algorithm to find (6.6), while Wollan and Dykstra (1987) gave its Fortran implementation.

Dykstra's algorithm:

1. Let $y_{1,1}$ denote the projection of y onto $\mathcal{C}_1$ and let $I_{1,1} = y_{1,1} - y$ be the incremental change incurred by the projection so that $y_{1,1} = y + I_{1,1}$.
2. Let $y_{1,2}$ denote the projection of $y + I_{1,1}$ onto $\mathcal{C}_2$. The incremental change is $I_{1,2} = y_{1,2} - (y + I_{1,1})$ so that $y_{1,2} = y + I_{1,1} + I_{1,2}$.
3. Let $y_{1,3}$ denote the projection of $y+I_{1,1}+I_{1,2}$ onto $\mathcal{C}_3$. The incremental change is $I_{1,3} = y_{1,3} - (y + I_{1,1} + I_{1,2})$ so that $y_{1,3} = y + I_{1,1} + I_{1,2} + I_{1,3}$.
4. After $y_{1,p}$ and $I_{1,p} = y_{1,p} - (y + I_{1,1} + \cdots + I_{1,p-1})$ are found, let $y_{2,1}$ denote the projection of $y + I_{1,2} + \cdots + I_{1,p}$ onto $\mathcal{C}_1$. Note that we have removed the increment $I_{1,1}$ before the projection. Our new increment is $I_{2,1} = y_{2,1} - (y + I_{1,2} + \cdots + I_{1,p})$ so that $y_{2,1} = y + I_{2,1} + I_{1,2} + \cdots + I_{1,p}$.
5. Let $y_{t,j}$ is the projection onto the jth cone $\mathcal{C}_j$ during the tth cycle. Thus, in general, $y_{t,j}$ is the projection of $y + I_{t,1} + \cdots + I_{t,j-1} + I_{t-1,j+1} + \cdots + I_{t-1,p}$ onto $\mathcal{C}_j$ and $I_{t,j} = y_{t,j} - (y + I_{t,1} + \cdots + I_{t,j-1} + I_{t-1,j+1} + \cdots + I_{t-1,p})$.
6. Continuing the process until $|y_{t-1,p} - y_{t,p}| < 10^{-\varepsilon_0}$ for some positive number ε_0, we have $P(y|\mathcal{C}) = y_{t,p}$.

In general, Dykstra's algorithm is computationally intensive, especially for high-dimensional cases. For example, when $p = 1,000$, each cycle in the algorithm includes 1,000 iterations. Since the next iteration depends on the output of the previous iteration, the convergence of the Dykstra's algorithm is slow. In contrary, the proposed fast EM algorithm (e.g., (3.8) and (3.10)) is simple and easy to code (e.g., with a matrix language, such as SPLUS and MATLAB).

6.3. Simulation and comparison

In (6.1), let $m = 41$, $z_0 = -\infty$, $z_{m+1} = +\infty$, $z_i = -2 + 0.1(i-1)$ for $i = 1, \ldots, m$, and

$$f(z) = 1 + \frac{3}{[1 + \exp(3 + 4z)]^{0.65}}, \qquad z \in I\!R. \tag{6.7}$$

We generate $\{e_i\}_{i=1}^m$ from $N(0, 0.2^2)$ and obtain m observations $\{\xi_i\}_{i=1}^m$ via (6.1). The simulated data are displayed in Table 2. Since f is monotone and decreasing, we restrict μ by $\mu_1 \geq \cdots \geq \mu_m$, i.e., $A\mu \geq 0$, where $A = (a_{ij})$ is an $(m-1) \times m$ matrix with $a_{ii} = 1$, $a_{i,i+1} = -1$ and $a_{ij} = 0$ otherwise. Noting that $\text{rank}(A) = m - 1$, from the third transformation of (6.3) we have $\mu = X\theta$, where $\theta_{m\times 1} \in I\!R_+^{m-1} \times I\!R$ and

$$X_{m\times m} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \equiv \Delta_m. \tag{6.8}$$

Let $W = I_m$, then (6.4) becomes $\hat{\theta} = \arg\min\ ||\xi - X\theta||^2$ subject to $\theta \in I\!R_+^{m-1} \times I\!R$.

From Theorem 2, without loss of generality, let $r_{\text{opt}} = r_{\text{opt}}^{\rho} = r_{\text{opt}}^{\det} = r_{\text{opt}}^{\text{tr}} = 0$. Hence, the sub-optimal EM$^{\text{c}}(0)$ (i.e., the EM (3.15) and (3.10)) can be applied to obtain $\hat{\theta}$ and $\hat{\mu} = X\hat{\theta}$. The EM$^{\text{c}}(0)$ converged at $t = 200$ iterations and the CPU time was 0.67 seconds. The corresponding standard errors were obtained by the non-parametric bootstrap with 500 replications (see Table 2). Figure 1(a) shows that the curve with EM$^{\text{c}}(0)$ fits well.

Table 2. Simulated data and estimated results.

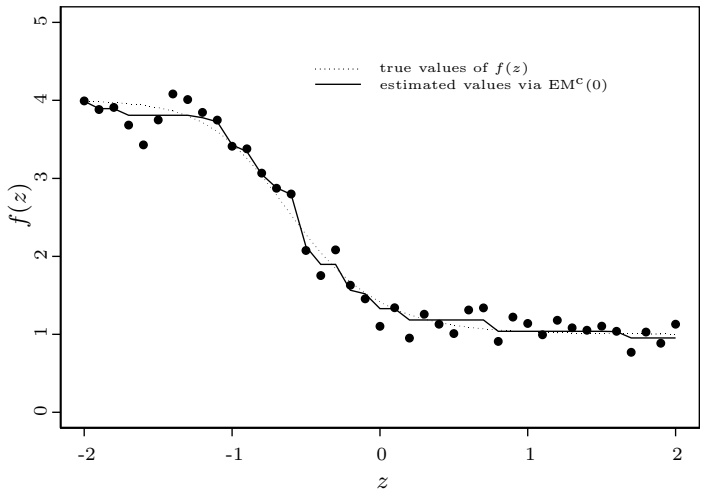
i	z_i	ξ_i	$\hat{\mu}_i^{\dagger}$	std†	$\hat{\mu}_i^{\ddagger}$	std‡	i	z_i	ξ_i	$\hat{\mu}_i^{\dagger}$	std†	$\hat{\mu}_i^{\ddagger}$	std‡
1	-2.0	3.994	3.989	0.663	3.994	0.543	21	0.0	1.102	1.330	0.305	1.222	0.312
2	-1.9	3.883	3.894	0.470	3.896	0.481	22	0.1	1.341	1.330	0.296	1.222	0.310
3	-1.8	3.910	3.894	0.419	3.896	0.412	23	0.2	0.952	1.184	0.266	1.166	0.265
4	-1.7	3.683	3.808	0.376	3.799	0.357	24	0.3	1.258	1.184	0.261	1.166	0.267
5	-1.6	3.428	3.808	0.354	3.799	0.345	25	0.4	1.129	1.184	0.262	1.166	0.258
6	-1.5	3.748	3.808	0.346	3.799	0.345	26	0.5	1.009	1.184	0.259	1.166	0.259
7	-1.4	4.082	3.808	0.340	3.799	0.341	27	0.6	1.311	1.184	0.255	1.166	0.263
8	-1.3	4.010	3.808	0.343	3.799	0.346	28	0.7	1.338	1.184	0.251	1.166	0.260
9	-1.2	3.845	3.778	0.344	3.799	0.354	29	0.8	0.908	1.038	0.250	1.089	0.254
10	-1.1	3.747	3.725	0.343	3.747	0.352	30	0.9	1.220	1.038	0.256	1.089	0.257
11	-1.0	3.411	3.428	0.358	3.411	0.351	31	1.0	1.139	1.038	0.255	1.089	0.251
12	-0.9	3.378	3.344	0.368	3.378	0.360	32	1.1	0.994	1.038	0.257	1.086	0.256
13	-0.8	3.066	3.051	0.382	3.066	0.360	33	1.2	1.179	1.038	0.255	1.086	0.254
14	-0.7	2.872	2.878	0.387	2.872	0.378	34	1.3	1.082	1.038	0.256	1.082	0.254
15	-0.6	2.799	2.793	0.387	2.799	0.376	35	1.4	1.051	1.038	0.255	1.077	0.251
16	-0.5	2.075	2.125	0.367	2.075	0.365	36	1.5	1.104	1.038	0.255	1.077	0.253
17	-0.4	1.752	1.896	0.356	1.917	0.367	37	1.6	1.039	1.034	0.259	1.039	0.258
18	-0.3	2.083	1.896	0.346	1.917	0.352	38	1.7	0.768	0.952	0.255	0.952	0.252
19	-0.2	1.629	1.565	0.332	1.629	0.341	38	1.8	1.028	0.952	0.255	0.952	0.249
20	-0.1	1.455	1.519	0.327	1.455	0.332	40	1.9	0.885	0.952	0.240	0.952	0.243
							41	2.0	1.128	0.952	0.222	0.952	0.231

$^{\dagger}\hat{\mu}_i$ and its standard error calculated by the sub-optimal EM$^{\text{c}}(0)$ algorithm.
$^{\ddagger}\hat{\mu}_i$ and its standard error calculated by Dykstra's algorithm.

To compare EM$^{\text{TNT}}$ (i.e., the EM (2.4) and (2.5)) with the sub-optimal EM$^{\text{c}}(0)$, we plot the target function values $||\xi - X\theta^{(t)}||^2$ against the EM iteration t. Figure 2 shows that the former converges much more slowly than does the latter, as is suggested by Theorem 3.

By applying Dykstra's algorithm, we obtained $\hat{\mu}$ with a CPU time of 3.58 seconds, and the corresponding standard errors with 500 bootstrap replications (see Table 2). Thus EM$^{\text{c}}(0)$ is at least 5 times as fast as Dykstra's algorithm. Figure 1(b) shows that the solutions with the sub-optimal EM$^{\text{c}}(0)$ and Dykstra's algorithm are quite close.

(a)



(b)

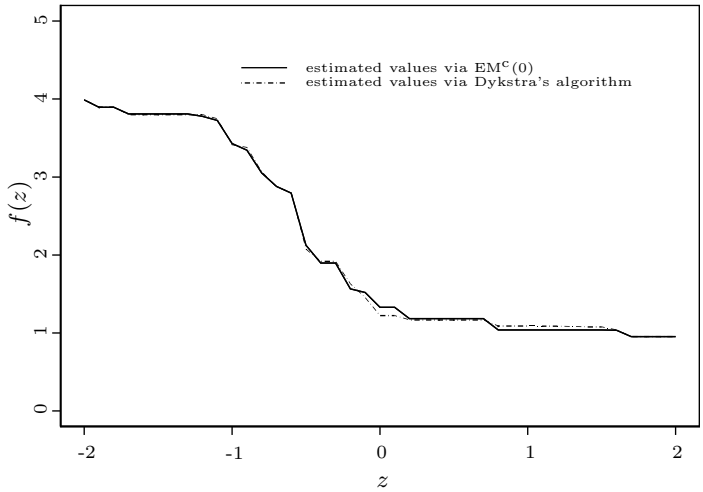


Figure 1. (a) Comparison among the true values (denoted by "· · · · · ·") of $f(z_i)$ defined by (6.7) , the simulated values (denoted by "o") of ξ_i based on (6.1) with $e_i \overset{\text{i.i.d.}}{\sim} N(0, 0.2^2)$, and the estimated values (denoted by "——") of μ_i subject to monotone and decreasing constraints via the sub-optimal EM^c(0) algorithm based on (3.15) and (3.10). (b) Comparison of the estimated values (denoted by "——") of μ_i via the sub-optimal EM^c(0) algorithm with the estimated values (denoted by "·–·–·–") of μ_i via Dykstra's algorithm.

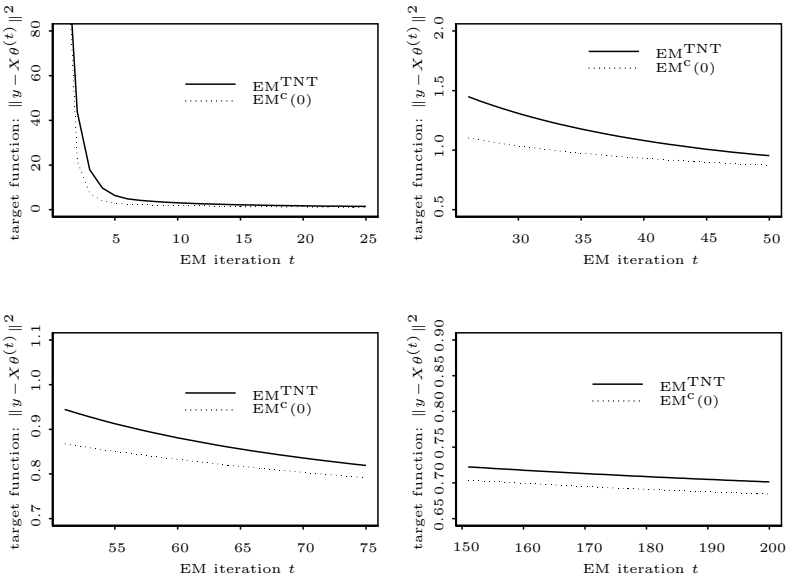


Figure 2. Comparison of the convergence speed between $\mathrm{EM}^{\mathrm{TNT}}$ and the sub-optimal $\mathrm{EM}^{\mathrm{c}}(0)$ based on (3.15) and (3.10). The y-axis is the target function $||\xi - X\theta^{(t)}||^2$.

7. Example — Freshmen's GPA Data

For illustrative purpose, we analyze the data of the first-year *grade point averages* (GPA) of 2,397 students at the University of Iowa in the fall of 1978 (Table 4 in Dykstra and Robertson (1982)). Let y_{ik} and w_{ik} denote the observed value of GPA score and the number of students in the (i,k)-cell, where the row index $i\,(1 \le i \le m$ and $m = 9)$ represents the group of student with *high-school ranks* (HSR) 91–99, 81–90, 71–80, 61–70, 51–60, 41–50, 31–40, 21–30 and 1–20, respectively, and the column index $k\,(1 \le k \le n$ and $n = 9)$ represents the group of student with ACT scores 1–12, 13–15, 16–18, 19–21, 22–24, 25–27, 28–30, 31–33 and 34–36, respectively. Let μ_{ik} denote the average GPA score for category (i,k) and $\mu_{m\times n} = (\mu_{ik})$. A natural restriction is that μ is non-decreasing in rows and non-increasing in columns, that is, $\mu \in \mathcal{S}(\mu) = \mathcal{S}_{\mathrm{row}}(\mu) \cap \mathcal{S}_{\mathrm{col}}(\mu)$, where $\mathcal{S}_{\mathrm{row}}(\mu) = \{\mu \in I\!R^{mn} : 0 \le \mu_{i1} \le \cdots \le \mu_{in},\ 1 \le i \le m\}$ and $\mathcal{S}_{\mathrm{col}}(\mu) = \{\mu \in I\!R^{mn} : \mu_{1k} \ge \cdots \ge \mu_{mk} \ge 0,\ 1 \le k \le n\}$. Dykstra (1983) proposed an algorithm to compute $\hat{\mu} = \arg\min_{\mu\in\mathcal{S}(\mu)} \Sigma_{i=1}^{m}\Sigma_{k=1}^{n} w_{ik}(y_{ik} - \mu_{ik})^2$.

Instead of calculating this $\hat{\mu}$, we are interested in estimating the effects of HSR and ACT. Specifically, the average GPA score μ_{ik} can be decomposed into two parts, one is from the effect of HSR at the ith level (denoted by α_i) and the other from the effect of ACT at the kth level (denoted by β_k), i.e., $\mu_{ik} = \alpha_i + \beta_k$. We call $\{\alpha_i\}$ the row effects and $\{\beta_k\}$ the column effects. Obviously,

with $\alpha = (\alpha_1, \ldots, \alpha_m)^\top$, $\alpha_1 \geq \cdots \geq \alpha_m \geq 0$, and with $\beta = (\beta_1, \ldots, \beta_n)^\top$, $0 \leq \beta_1 \leq \cdots \leq \beta_n$. The constrained LSEs $(\hat{\alpha}, \hat{\beta})$ are identical to their constrained MLEs in the normal model

$$y_{ik} = \alpha_i + \beta_k + e_{ik}, \quad e_{ik} \overset{\text{ind}}{\sim} N(0, \frac{1}{w_{ik}}), \quad i = 1, \ldots, m, \quad k = 1, \ldots, n. \tag{7.1}$$

Let $Y_{m\times n} = (y_{ik})$, $W_{m\times n} = (w_{ik})$, $\alpha = \Delta_m \theta^{(1)}$ and $\beta = \Delta_n^\top \theta^{(2)}$, where Δ_m is defined in (6.8), $\theta^{(1)} = (\theta_1, \ldots, \theta_m)^\top \in I\!R_+^m$ and $\theta^{(2)} = (\theta_{m+1}, \ldots, \theta_{m+n})^\top \in I\!R_+^n$. Then (7.1) becomes

$$\vec{Y} = X\theta + \vec{E}, \qquad \theta \in I\!R_+^{m+n}, \qquad \vec{E} \sim N_{mn}(0, \Omega^{-1}),$$

where $\Omega \equiv \text{diag}\,(\vec{W})$, $X \equiv (1_n \otimes \Delta_m \vdots [I_n \otimes 1_m]\Delta_n^\top)$ and $(A \vdots B)$ denotes the column-merged matrix of A and B. Obviously, the constrained MLE of $\theta = (\theta_1, \ldots, \theta_{m+n})^\top$ is

$$\hat{\theta} = \arg \min_{\theta \in I\!R_+^{m+n}} ||\Omega^{\frac{1}{2}} \vec{Y} - \Omega^{\frac{1}{2}} X\theta||^2.$$

Notice that all non-zero elements at each row of $\Omega^{1/2}X$ are equal, i.e., $\Omega^{1/2}X \in \mathcal{M}_{mn\times(m+n)}$. Thus from Theorem 2, we have $r_{\text{opt}} = r_{\text{opt}}^{\rho} = r_{\text{opt}}^{\text{det}} = r_{\text{opt}}^{\text{tr}} = 0$. Hence, the sub-optimal $\text{EM}^{\text{c}}(0)$ based on (3.15) and (3.10) can be applied to obtain $\hat{\theta}$. Using the initial values $\theta^{(0)} = (0.1, \ldots, 0.1)^\top$, the $\text{EM}^{\text{c}}(0)$ become stable at $t = 1,000$ iterations and the CPU time was 5.09 seconds. We list $\hat{\alpha} = \Delta_m \hat{\theta}^{(1)}$ and $\hat{\beta} = \Delta_n^\top \hat{\theta}^{(2)}$ in the 4th and the 9th column of Table 3. The standard errors were obtained with 25.5 minutes of CPU time by the non-parametric bootstrap with 500 replications.

Table 3. Estimations of the row and column effects for freshmen's GPA data.

i	HSR Level	α_i	$\hat{\alpha}_i$	std	k	ACT Level	β_k	$\hat{\beta}_k$	std
1	$91 \leq$ HSR ≤ 99	α_1	1.775	0.1029	1	1-12	β_1	0.699	0.0603
2	$81 \leq$ HSR ≤ 90	α_2	1.362	0.1026	2	13-15	β_2	1.001	0.1054
3	$71 \leq$ HSR ≤ 80	α_3	1.229	0.1024	3	16-18	β_3	1.160	0.1065
4	$61 \leq$ HSR ≤ 70	α_4	1.110	0.1109	4	19-21	β_4	1.237	0.0984
5	$51 \leq$ HSR ≤ 60	α_5	0.912	0.0975	5	22-24	β_5	1.321	0.0959
6	$41 \leq$ HSR ≤ 50	α_6	0.912	0.0984	6	25-27	β_6	1.369	0.0996
7	$31 \leq$ HSR ≤ 40	α_7	0.764	0.1084	7	28-30	β_7	1.520	0.1099
8	$21 \leq$ HSR ≤ 30	α_8	0.699	0.0982	8	31-33	β_8	1.578	0.1279
9	HSR ≤ 20	α_9	0.699	0.0603	9	34-36	β_9	1.578	0.1856

8. Discussion

Box constraints and LICs are two special cases of the general convex constraints (CCs). Quadratic optimization with CCs, the core computation in nonlinear programming with CCs, is a fundamental and difficult problem. Existing

packages are not applicable to optimization problems beyond the box constraints, the LECs and the LICs. Although EM-type algorithms may be applicable, slow convergence has hindered their applications, especially, to high-dimensional data. The method we developed accelerates an EM for solving the QO problem with CCs by utilizing the 'working parameter' scheme. The techniques of latent-variable reduction and dimension reduction were proposed to further speed up the EM. Theoretical and simulated results showed that the new algorithm outperforms that of Tian et al. (2005) substantially. The proposed methods can be applied to QO with CCs such as ellipsoid, simplex, or quadratic constraints. For example, we used the $\mathrm{EM}^{\mathrm{tr}}(1)$ algorithm to re-analyze the bituminous concrete data (Sec. 7.4 of Tian et al. (2005)) and obtain the same results, but with only 20 iterations as opposed to 30 iterations.

Meng and van Dyk's acceleration scheme is basically an art, in which one finds the optimal value for the working parameter before the EM starts, by minimizing the fraction of missing information. Another closely related acceleration scheme is that of the Liu, Rubin and Wu (1998) parameter expansion method, in which the expanded parameters are estimated during the EM process, together with original parameters of interest. For example, instead of defining $\{\lambda_{ij}\}$ explicitly according to (3.4), if we treat $\{\lambda_{ij}\}$ in (3.5) as expanded parameters and estimate them from the imputed data, subject to $\lambda_{ij} > 0$ and $\Sigma_{j\in\mathcal{J}_i}\lambda_{ij} = 1$, then we can derive a PX-EM algorithm. Theoretically, if the M-step could be done in closed form, then the PX-EM would outperform the optimal EM algorithm (from Theorem 1, we have $r^{\mathrm{tr}}_{\mathrm{opt}} = 1$). However, the explicit expressions for the complete-data MLEs of $\{\lambda_{ij}\}$ are unavailable for the present situation. In fact, let $\Theta = (\theta, \lambda)$, where $\lambda = \{\lambda_{ij}\}$ denote the auxiliary parameters. Similar to (3.6), the log-likelihood function of Θ for the complete-data is proportional to $-0.5\Sigma_{i=1}^{m}\Sigma_{j\in\mathcal{J}_i}\log\lambda_{ij} - 0.5\Sigma_{i=1}^{m}\Sigma_{j\in\mathcal{J}_i}(Z_{ij} - x_{ij}\theta_j)^2/\lambda_{ij}$. For ease of presentation, we assume that $\mathcal{J}_i = \{1, \ldots, q-1, q\}$. For given θ, the complete-data MLEs of $\{\lambda_{ij}\}$ with constraints $\Sigma_{j=1}^{q}\lambda_{ij} = 1$ are determined by the system of non-linear equations:

$$\lambda_{ij}^{-1} - \delta_{ij}\lambda_{ij}^{-2} = \lambda_{iq}^{-1} - \delta_{iq}\lambda_{iq}^{-2}, \qquad j = 1, \ldots, q-1, \quad i = 1, \ldots, m,$$

where $\delta_{ij} \equiv (Z_{ij} - x_{ij}\theta_j)^2$. Although explicit solution to the equations is not available, it is worthwhile to explore to what extent the PX-EM can be best utilized in the future.

In the context of working parameter, Meng (1994) and Meng and van Dyk (1997) suggested using the largest eigenvalue as the criterion to speed up EM. We found that the trace criterion works much better (at least in our examples of both simulations and data). Therefore it is worthwhile to theoretically investigate this issue further in other models such as logistic regression for binary

data and Poisson regression for counting data. Furthermore, the fast EM algorithm designed for the QO with constraints may also improve the lasso-type algorithms in variables selection and other EM-type algorithms in medical image reconstruction. We plan to address these issues in a separate report.

Acknowledgements

We are grateful to the Editor, an associate editor and two referees for their constructive comments. M Tan and GL Tian's research was supported in part by U.S. National Cancer Institute grants CA119758. The research of K. W. Ng was partially supported by a research grant of the University of Hong Kong.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Coleman, T. F. and Li, Y. (1996). A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM J. Optim.* **6**, 1040-1058.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78**, 837-842.

Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Amer. Statist.* **10**, 708-716.

Fraser, D. A. S. and Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints with application to concave regression. *Scand. J. Statist.* **16**, 65-74.

Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization.* Academic Press, London.

Green, P. J. (1990). On the use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52**, 443-452.

Hajivassiliou, V. A. and McFadden, D. L. (1998). The method of simulated scores for the estimation of LDV models. *Econometrica* **66**, 863-896.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49**, 598-619.

Kim, D. K. and Taylor, J. M. G. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *J. Amer. Statist. Assoc.* **90**, 708-716.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Edited by J. Neyman). University of California Press, Berkeley.

Lange, K. (1999). *Numerical Analysis for Statisticians.* Springer, New York.

Liu, C. H. (2000). Estimation of discrete distributions with a class of simplex constraints. *J. Amer. Statist. Assoc.* **95**, 109-120.

Liu, C. H., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755-770.

Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22**, 326-339.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267-278.

Meng, X. L. and van Dyk, D. (1997). The EM algorithm — an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meyer, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *J. Statist. Plann. Inference* **81**, 13-31.

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Appl. Statist.* **40**, 133-141.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Shi, N. Z., Zheng, S. R. and Guo, J. H. (2005). The restricted EM algorithm under inequality restrictions on the parameters. *J. Multivariate Anal.* **92**(1), 53-76.

Silverman, B. W., Jones, M. C., Nychka, D. W. and Wilson, J. D. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 271-324.

Tan, M., Tian, G. L. and Fang, H. B. (2003). Estimating restricted normal means using the EM-type algorithms and IBF sampling. In *Development of Modern Statistics and Related Topics — In Celebration of Prof. Yaoting Zhang's 70th Birthday* (Edited by J. Huang and H. Zhang), 53-73. World Scientific, New Jersey.

Tian, G. L., Ng, K. W. and Tan, M. (2005). Likelihood-based approaches for constrained parameter problems in multiple regression models. *Ann. Inst. Statist. Math.*, in revision.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Titterington, D. M. (1985). General structure of regularization procedures in image reconstruction. *Astronomy and Astrophysics* **144**, 381-387.

Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 569-612.

Wolfe, P. (1959). The simplex method for quadratic programming. *Econometrica* **27**, 382-398.

Wollan, P. C. and Dykstra, R. L. (1987). Minimizing linear inequality constrained Mahalanobis distances. *Appl. Statist.* **36**, 234-240.

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, Maryland 21201, U.S.A.

E-mail: mtan@umm.edu

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, Maryland 21201, U.S.A.

E-mail: gtian2@umm.edu

Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, Maryland 21201, U.S.A.

E-mail: hfang@umm.edu

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P.R. China.

E-mail: kw.ng@hkuspace.hku.hk

(Received June 2005; accepted March 2006)