Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Efficient Robbins–Monro procedure for multivariate binary data

Cui Xiong & Jin Xu

To cite this article: Cui Xiong & Jin Xu (2018) Efficient Robbins–Monro procedure for multivariate binary data, Statistical Theory and Related Fields, 2:2, 172-180, DOI: 10.1080/24754269.2018.1507384

To link to this article: https://doi.org/10.1080/24754269.2018.1507384

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 172–180
https://doi.org/10.1080/24754269.2018.1507384

Taylor & Francis
Taylor & Francis Group

Efficient Robbins–Monro procedure for multivariate binary data

Cui Xiong and Jin Xu

School of Statistics, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

This paper considers the problem of jointly estimating marginal quantiles of a multivariate distribution. A sufficient condition for an estimator that converges in probability under a multivariate version of Robbins–Monro procedure is provided. We propose an efficient procedure which incorporates the correlation structure of the multivariate distribution to improve the estimation especially for cases involving extreme marginal quantiles. Estimation efficiency of the proposed method is demonstrated by simulation in comparison with a general multivariate Robbins–Monro procedure and an efficient Robbins–Monro procedure that estimates the marginal quantiles separately.

ARTICLE HISTORY
Received 6 January 2018
Revised 18 July 2018
Accepted 31 July 2018

KEYWORDS
Binary response; quantile estimation; Robbins–Monro procedure; sequential design

1. Introduction

Let $M(x)$ be the distribution function of a random variable X. Robbins and Monro (1951) introduced a stochastic approximation method to find the α-quantile $\theta = M^{-1}(\alpha)$ (assuming it is unique) through a sequential search given by

$$x_{n+1} = x_n - a_n(y_n - \alpha), \tag{1}$$

where x_1 is an arbitrary initial guess of θ, y_n is the binary response with expected value $M(x_n)$ and a_n is a prespecified sequence of positive constants. They showed that when a_n satisfies $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n^2 < \infty$, x_n converges to θ in probability. Applications of this procedure include quantal response curve estimation in sensitivity experiments, dose-finding in clinical trials and sequential learning, to just name a few (Cheung, 2010; Duflo, 1997; Wu, 1985).

It is known that the procedure is asymptotically efficient when $a_n = \{n\dot{M}(\theta)\}^{-1}$, where $\dot{M}$ is the first derivative of M (Chung, 1954; Sacks, 1958). Various variants of Robbins–Monro procedure of (1) and other model-based approaches were proposed to improve the finite sample performance (Chaloner & Larntz, 1989; Chaudhuri & Mykland, 1993; Dror & Steinberg, 2006, 2008; Hung & Joseph, 2014; Lai & Robbins, 1979; Neyer, 1994; Ruppert, 1988; Wu, 1985, 1986; Wu & Tian, 2014). It is also known that the Robbins–Monro procedure does not perform well for extreme values of α (Wetherill, 1963; Young & Easterling, 1994). To improve the convergence performance in this case, Joseph (2004) proposed an efficient Robbins–Monro procedure which modifies (1) by

$$x_{n+1} = x_n - a_n(y_n - b_n), \tag{2}$$

where b_n is a sequence of constants in $(0, 1)$ converging to α. The sequences a_n and b_n are chosen in a way such that the conditional mean square error is minimised under a Bayesian framework. The explicit forms of a_n and b_n under normal approximation are derived and showed to work for general M as well.

In this paper, we consider a multivariate extension of this estimation problem. Let $M(\mathbf{x})$ be the distribution function of a p-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$ with finite second moments. Denote its jth marginal distribution by M_j. Given a constant vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ in $(0, 1)^p$, we are interested in jointly estimating the marginal quantiles $\boldsymbol{\theta} = (M_1^{-1}(\alpha_1), \ldots, M_p^{-1}(\alpha_p))$, assuming that $M_j^{-1}(\alpha_j)$ is unique for each j. We also assume that $\dot{M}_j(\theta_j) > 0$ for all $j = 1, \ldots, p$. Suppose that given each factor x_j, we observe an independent binary response y_j with $\mathrm{E}(y_j \mid x_j) = M_j(x_j)$. Such situation arises in different fields of researches. For instance, in sensitivity experiment study, several sensitivity experiments are conducted in parallel. In each experiment, stimulus level of one factor is tested with dichotomous outcome, response or non-response, and the factors considered across experiments are highly correlated. It is of interest to coordinate the designs for individual factor for more efficiency. In oncology dose-finding clinical trials, several agents are considered at the same time. For each agent, a trial is conducted to search for the level of maximum tolerated dose, which corresponds to the 25% or 30% quantile of a unknown distribution. Across the agents, various binary responses representing different types of adverse events are observed. The joint marginal dose levels are used for evaluation of possible combination agent trials. Apparently, $\boldsymbol{\theta}$ can be estimated by applying the Robbins–Monro procedure or

CONTACT Jin Xu jxu@stat.ecnu.edu.cn

the efficient version to each component of $\mathbf{x}$. However since $x_1, \ldots, x_p$ are correlated, these methods may lose efficiency when estimating the marginal quantiles separately. This motivates us to seek a sequential procedure that estimates $\boldsymbol{\theta}$ jointly.

Multivariate Robbins–Monro procedures that aim to find the root of a multivariate continuous function $\mathbf{f} : \mathbb{R}^p \to \mathbb{R}^p$ through a regression $\mathrm{E}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$ have been studied (Duflo, 1997). For example, Ruppert (1985) proposed a multivariate Newton–Raphson version which is in a way similar to multivariate Kiefer–Wolfowitz procedure to minimise $\| \mathbf{f}(\mathbf{x}) \|^2$. Wei (1987) proposed a multivariate Robbins–Monro procedure which employs a Venter-type estimate of the Jacobian of $\mathbf{f}$. The method we propose here is primarily for binary responses.

The remainder of the paper is organised as follows. In Section 2, we first present a general multivariate Robbins–Monro procedure under which the sequential estimator converges in probability. Second, we develop an efficient version which is optimal under a criterion that is naturally extended from the univariate case. Section 3 contains simulation studies to demonstrate the superiority of the proposed method over a general multivariate Robbins–Monro procedure and an efficient Robbins–Monro procedure that estimates the marginal quantiles separately. All proofs are gathered in Appendix.

2. Main results

First, we extend (1) to a multivariate version as follows:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - A_n(\mathbf{y}_n - \boldsymbol{\alpha}), \tag{3}$$

where $\mathbf{y}_n = (y_{1n}, \ldots, y_{pn})^\top$ contains p binary responses observed at $\mathbf{x}_n$ and each y_{jn} has the expected value $M_j(x_{jn})$, A_n is a sequence of $p \times p$ constant matrices whose (j, k)th element is denoted by $a_{jk,n}$. The following theorem gives a sufficient condition for $\mathbf{x}_n$ to converge to $\boldsymbol{\theta}$ in probability.

Theorem 2.1: *Suppose that A_n satisfies the following conditions:*

$$\begin{aligned} &a_{jj,n} > 0, \quad \sum_{n=1}^{\infty} a_{jj,n} = \infty, \quad \sum_{n=1}^{\infty} a_{jj,n}^2 < \infty, \\ &\text{for } j = 1, \ldots, p, \\ &|a_{jk,n}/a_{jj,n}| \to 0, \quad \text{for } k \neq j. \end{aligned} \tag{4}$$

Then, $\mathbf{x}_n$ in (3) converges to $\boldsymbol{\theta}$ in probability as $n \to \infty$.

When $a_{jk,n} = 0$ for all $j \neq k$, Theorem 2.1 reduces to p univariate Robbins–Monro procedures. It indicates that when the diagonal elements of A_n are of $\mathrm{O}(n^{-1})$ and dominate the off-diagonal elements in magnitude, the sequential estimator converges regardless of the exact values of $a_{jk,n}$. Clearly, this arbitrariness of A_n can lead to inefficiency in estimation as showed by simulation in Section 3.

In light of Joseph (2004) to improve the convergence in case of extreme quantile, we propose an efficient version of (3) by replacing $\boldsymbol{\alpha}$ by a vector sequence $\mathbf{b}_n$ in $(0, 1)^p$, i.e.

$$\mathbf{x}_{n+1} = \mathbf{x}_n - A_n(\mathbf{y}_n - \mathbf{b}_n). \tag{5}$$

Additional condition on $\mathbf{b}_n$ to guarantee the convergence is provided in the following theorem.

Theorem 2.2: *Suppose that A_n satisfies* (4) *and $\mathbf{b}_n$ satisfies*

$$\sum_{n=2}^{\infty} \sum_{s=1}^{n-1} \mathbf{1}_p^\top \mid A_s^\top A_n(\boldsymbol{\alpha} - \mathbf{b}_n) \mid_{\mathrm{e}} < \infty, \tag{6}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of ones, $|\cdot|_{\mathrm{e}}$ stands for the operator that takes element-wise absolute value of a vector or a matrix. Then, $\mathbf{x}_n$ in (5) *converges to $\boldsymbol{\theta}$ in probability as $n \to \infty$.*

Here the idea of introducing a varying sequence $\mathbf{b}_n$ is to create a balanced step length in early stage when $\boldsymbol{\alpha}$ contains extreme values. As n gets large, $\mathbf{b}_n$ converges to $\boldsymbol{\alpha}$ and its effect diminishes. Again, there are many sequences of A_n and $\mathbf{b}_n$ satisfying the conditions of Theorem 2.2. We now seek a pair of them that is optimal in a way that is naturally extended from the univariate case.

Assume that M is from a location family with parameter $\boldsymbol{\theta}$. Hereafter, we denote $M(\mathbf{x})$ by $M(\mathbf{x} - \boldsymbol{\theta})$. Suppose the initial guess $\mathbf{x}_1$ is obtained with some prior information of $\boldsymbol{\theta}$, to be specific, a prior distribution of $\boldsymbol{\theta}$ with $\mathrm{E}(\boldsymbol{\theta}) = \mathbf{x}_1$ and $\mathrm{cov}(\boldsymbol{\theta}) = \Sigma_1$.

Let $\mathbf{z}_n = \mathbf{x}_n - \boldsymbol{\theta}$. Then, (5) becomes

$$\mathbf{z}_{n+1} = \mathbf{z}_n - A_n(\mathbf{y}_n - \mathbf{b}_n), \tag{7}$$

where y_{jn} is a binary variable with expected value $M_j(z_{jn})$. Denote $\mathbf{m}_n = (m_{1n}, \ldots, m_{pn})^\top = (M_1(z_{1n}), \ldots, M_p(z_{pn}))^\top$ and $\Sigma_n = \mathrm{cov}(\mathbf{z}_n) = (\sigma_{jk,n})$. As a natural extension of the univariate case in Joseph (2004), we propose to choose A_n and $\mathbf{b}_n$ such that $\mathrm{tr}\{\mathrm{cov}(\mathbf{z}_{n+1})\}$ is minimised subject to the condition that $\mathrm{E}(\mathbf{z}_{n+1}) = \mathbf{0}$. By (7), this condition implies

$$\mathrm{E}(\mathbf{z}_n) - A_n\{\mathrm{E}(\mathbf{m}_n) - \mathbf{b}_n\} = 0.$$

Since $(A_1, \mathbf{b}_1), \ldots, (A_{n-1}, \mathbf{b}_{n-1})$ are chosen such that $\mathrm{E}(\mathbf{z}_2) = \cdots = \mathrm{E}(\mathbf{z}_n) = \mathbf{0}$, we have $\mathbf{b}_n = \mathrm{E}(\mathbf{m}_n)$, which,

together with (7), leads to

$$\Sigma_{n+1} = \Sigma_n - A_n \mathrm{E}(\mathbf{m}_n \mathbf{z}_n^\top) - \mathrm{E}(\mathbf{z}_n \mathbf{m}_n^\top) A_n^\top + A_n \mathrm{cov}(\mathbf{y}_n) A_n^\top.$$

Minimising $\mathrm{tr}\{\mathrm{cov}(\mathbf{z}_{n+1})\}$ with respect to A_n by solving $\partial \mathrm{tr}\Sigma_{n+1}/\partial A_n = \mathbf{0}$, we obtain

$$A_n = \mathrm{E}(\mathbf{z}_n \mathbf{m}_n^\top)\mathrm{cov}^{-1}(\mathbf{y}_n),$$

and

$$\Sigma_{n+1} = \Sigma_n - \mathrm{E}(\mathbf{z}_n \mathbf{m}_n^\top)\mathrm{cov}^{-1}(\mathbf{y}_n)\mathrm{E}(\mathbf{m}_n \mathbf{z}_n^\top). \quad (8)$$

The components of $\mathrm{cov}(\mathbf{y}_n)$ can be expressed as

$$\begin{aligned} &\mathrm{var}(y_{jn}) = \mathrm{E}(m_{jn})\{1 - \mathrm{E}(m_{jn})\}, \quad j = 1, \ldots, p, \\ &\mathrm{cov}(y_{jn}, y_{kn}) = \mathrm{E}(m_{jn} m_{kn}) - \mathrm{E}(m_{jn})\mathrm{E}(m_{kn}), \quad j \neq k. \end{aligned} \quad (9)$$

The expectations of $\mathrm{E}(m_{jn})$, $\mathrm{E}(z_{jn} m_{kn})$ and $\mathrm{E}(m_{jn} m_{kn})$ in (8) and (9) depend on the unknown distribution M and the distribution of $\mathbf{z}_n$. (Note that $\mathrm{cov}(\mathbf{y}_n)$ is invertible unless some x_j and x_k are identical.)

To facilitate the evaluation of these expectations, we first approximate $M(\mathbf{z})$ by

$$G(\mathbf{z}) = \Phi_p\left(\mathbf{a} + B\mathbf{z}; \mathbf{0}, R\right), \quad (10)$$

where $\Phi_p(.; \boldsymbol{\mu}, \Sigma)$ is the distribution function of p-dimensional normal vector with mean $\boldsymbol{\mu}$ and covariance Σ, $\mathbf{a} = (\Phi^{-1}(\alpha_1), \ldots, \Phi^{-1}(\alpha_p))^\top$, $B = \mathrm{diag}(\beta_1, \ldots, \beta_p)$ with $\beta_j = \dot{M}_j(0)/\phi\{\Phi^{-1}(\alpha_j)\}$, Φ and ϕ are respectively the distribution function and density of the standard normal variable, $R = (\rho_{jk})$ with $\rho_{jk} = \mathrm{corr}(z_j, z_k)$. Then, the marginal distribution of G, given by $G_j(z) = \Phi(\Phi^{-1}(\alpha_j) + \beta_j z)$, coincides with $M_j(z)$ in both the value and the derivative at 0, i.e. $G_j(0) = M_j(0)$ and $\dot{G}_j(0) = \dot{M}_j(0)$. In this way, G captures the local behaviour of M at the point of interest. Further let $\mathbf{g}_n = (g_{1n}, \ldots, g_{pn})^\top = (G_1(z_{1n}), \ldots, G_p(z_{pn}))^\top$.

Second, denote the density of $\mathbf{z}_n$ by f_n. Observe that f_{n+1} can be obtained recursively by

$$\begin{aligned} f_{n+1}(\mathbf{z}) &= \sum_{s_1, \ldots, s_p = 0,1} \mathrm{P}(y_{1n} = s_1, \ldots, y_{pn} \\ &\quad = s_p \mid \mathbf{z}_n) f_n[\mathbf{z} + A_n\{(s_1, \ldots, s_p)^\top - \mathbf{b}_n\}], \end{aligned}$$

which is rather complicated. Again, we choose to approximate the distribution of $\mathbf{z}_n$ by another multivariate normal distribution $\Phi_p(\mathbf{z}; \mathbf{0}, \Sigma_n)$ as their first two moments agree.

It is worth pointing out that neither the first-order approximation by G nor the moment agreement by $\Phi_p(\mathbf{z}; \mathbf{0}, \Sigma_n)$ guarantees the overall closeness to the distribution. Such approximations are not unique or optimal. The main advantages of using multivariate normal approximations are computational ease and sufficiency for the desired convergence, as we show next.

Now, based on these two approximations, we can estimate $\mathrm{E}(m_{jn})$, $\mathrm{E}(z_{jn} m_{kn})$ and $\mathrm{E}(m_{jn} m_{kn})$ respectively by $\mathrm{E}(g_{jn})$, $\mathrm{E}(z_{jn} g_{kn})$ and $\mathrm{E}(g_{jn} g_{kn})$, where the expectations are taken with respect to $\Phi_p(\mathbf{z}; \mathbf{0}, \Sigma_n)$. Their expressions are obtained as follows:

$$\begin{aligned} \mathrm{E}(g_{jn}) &= \mathrm{E}\{\Phi(\Phi^{-1}(\alpha_j) + \beta_j z_{jn})\} \\ &= \Phi\left\{\frac{\Phi^{-1}(\alpha_j)}{(1 + \beta_j^2 \sigma_{jj,n})^{1/2}}\right\}, \end{aligned} \quad (11)$$

$$\begin{aligned} \mathrm{E}(z_{jn} g_{kn}) &= \mathrm{E}\{z_{jn} \Phi(\Phi^{-1}(\alpha_k) + \beta_k z_{kn})\} \\ &= \frac{\beta_k \sigma_{jk,n}}{(1 + \beta_k^2 \sigma_{kk,n})^{1/2}} \phi\left\{\frac{\Phi^{-1}(\alpha_k)}{(1 + \beta_k^2 \sigma_{kk,n})^{1/2}}\right\}, \end{aligned} \quad (12)$$

$$\begin{aligned} \mathrm{E}(g_{jn} g_{kn}) &= \mathrm{E}\{\Phi(\Phi^{-1}(\alpha_j) + \beta_j z_{jn}) \Phi(\Phi^{-1}(\alpha_k) \\ &\quad + \beta_k z_{kn})\} \\ &= \Phi_2\left\{(\Phi^{-1}(\alpha_j), \Phi^{-1}(\alpha_k))^\top; \mathbf{0}, \mathrm{I}_2 + \tilde{\Sigma}_{jk,n}\right\}, \end{aligned} \quad (13)$$

where

$$\tilde{\Sigma}_{jk,n} = \begin{pmatrix} \beta_j^2 \sigma_{jj,n} & \beta_j \beta_k \sigma_{jk,n} \\ \beta_j \beta_k \sigma_{jk,n} & \beta_k^2 \sigma_{kk,n} \end{pmatrix}.$$

At last, we obtain the sequences of A_n and $\mathbf{b}_n$ in the efficient procedure of (5) or (7) as

$$A_n = \mathrm{E}(\mathbf{z}_n \mathbf{g}_n^\top)\mathrm{cov}^{-1}(\mathbf{y}_n), \quad \mathbf{b}_n = \mathrm{E}(\mathbf{g}_n), \quad (14)$$

with

$$\Sigma_{n+1} = \Sigma_n - \mathrm{E}(\mathbf{z}_n \mathbf{g}_n^\top)\mathrm{cov}^{-1}(\mathbf{y}_n)\mathrm{E}(\mathbf{g}_n \mathbf{z}_n^\top), \quad (15)$$

where the components of $\mathrm{E}(\mathbf{z}_n \mathbf{g}_n^\top)$, $\mathrm{cov}(\mathbf{y}_n)$ and $\mathbf{b}_n$ are given in (11)–(13). Note that A_n and $\mathbf{b}_n$ are sequences that can be specified before the experiment once Σ_1 is provided. When Σ_1 is diagonal, i.e. the components of $\mathbf{x}_1$ are uncorrelated, the component-wise coefficients of A_n and $\mathbf{b}_n$ in (14) reduce to a_n and b_n respectively in (2) given by Joseph (2004).

The following theorem gives the convergence property of the proposed sequential design when M is multivariate normal.

Theorem 2.3: *Suppose that the distribution of* $\mathbf{z}$ *is given by* (10)*. Then, for the procedure in* (7) *with coefficient sequences in* (14), $\Sigma_n \to \mathbf{0}$, $\mathbf{b}_n \to \boldsymbol{\alpha}$, *as* $n \to \infty$.

Theorem 2.3 implies that $\mathbf{z}_n$ converges to $\mathbf{0}$ and hence $\mathbf{x}_n$ converges to $\boldsymbol{\theta}$ in probability. The next theorem shows that the result in fact holds for general M.

Theorem 2.4: *For the procedure in* (7) *with coefficient sequences in* (14), $\mathbf{z}_n \to \mathbf{0}$ *in probability, as* $n \to \infty$.

We call the sequential design in (5) with coefficients in (14) efficient multivariate Robbins–Monro procedure. A couple of points are worthy to be noted. (i) The most important impact of the procedure lies on the sequence $\mathbf{b}_n$ for which each of its component is between α_j and 0.5 for the early stage to avoid unnecessary large-scale oscillation of the search steps as pointed out by Joseph (2004). (ii) The contribution of the correlation structure of $\mathbf{x}$ is implemented through the procedure in (14) to minimise $\mathrm{tr}\{\mathrm{cov}(\mathbf{z}_n)\}$.

In the end, we would like to comment on two practical issues in carrying out the procedure. First, as seen from (5), the starting value of Σ_1 plays a key role in the construction of A_n and $\mathbf{b}_n$. In reality it is usually unknown to the experimenter. A plausible solution is to estimate it from a moderate sample of $\mathbf{x}$. It is possible and less expensive since no response of $\mathbf{y}$ is needed. Second, for the unknown coefficients β_j which also depends on the unknown M_j, it can be estimated adaptively from the data by fitting a parametric model like in Anbar (1978) and Lai and Robbins (1979). We will demonstrate the effect of these approximations through simulation in the next section.

3. Simulations

In this section, we conduct simulations to compare the performance of the following four procedures for jointly estimating the marginal quantiles: (i) the multivariate Robbins–Monro procedure in (3), denoted by MRM; (ii) the proposed efficient version in (5), denoted by eMRM; (iii) the procedure that estimates the marginal quantiles separately by (1), denoted by RM; (iv) the efficient procedure that estimates the marginal quantiles separately by (2), denoted by eRM.

3.1. Set up

Consider three bivariate distributions given in the first column of Table 1, where (i) $\mathrm{MVN}(\mathbf{0}, \Sigma)$ is bivariate normal distribution with mean $\mathbf{0}$ and $\Sigma = \left(\begin{smallmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{smallmatrix}\right) = \left(\begin{smallmatrix}1 & 0.9\\ 0.9 & 1\end{smallmatrix}\right)$, (ii) $\mathrm{MVT}(\mathbf{0}, \Sigma, 4)$ is bivariate t distribution with location parameter $\mathbf{0}$, scale matrix Σ and degrees of freedom four (Kotz & Nadarajah, 2004), (iii) $\mathrm{MSN}(\mathbf{0}, \Sigma, \boldsymbol{s})$ is a bivariate skew-normal distribution with location parameter $\mathbf{0}$, scale matrix Σ, shape parameter $\boldsymbol{s} = (1, 1)^\top$ (Azzalini, 1998). The covariance matrices of these three distributions are respectively

$$\Sigma, \quad 2\Sigma, \quad \text{and} \quad \Sigma - \boldsymbol{\mu}_s\boldsymbol{\mu}_s^\top, \tag{16}$$

where $\boldsymbol{\mu}_s = (2/\pi)^{1/2}(1 + \boldsymbol{s}^\top\Sigma\boldsymbol{s})^{-1/2}\Sigma\boldsymbol{s} = (0.692, 0.692)^\top$, all indicating a strong positive correlation. The marginal distribution of $M_j(z_j)$ under these models are given in the second column of Table 1, where (i) $t(.,f)$, $t^{-1}(.,f)$ and $d_t(.,f)$ are respectively the distribution function, quantile and density of a t random variable with degrees of freedom f, (ii) $\mathrm{sn}(.;\mu,\omega,s)$, $\mathrm{sn}^{-1}(.;\mu,\omega,s)$ and $d_{\mathrm{sn}}(.;\mu,\omega,s)$ are the distribution function, quantile and density of a (univariate) skew-normal random variable with location parameter μ, scale parameter ω and shape parameter s. In addition, the corresponding β_j under these models are given in the third column of Table 1, where $\bar{s}_1 = \bar{s}_2 = (s_1 + \sigma_{11}^{-1}\sigma_{12}s_2)/(1 + s_2^\top\sigma_{22.1}s_2)^{1/2} = 1.742, \sigma_{22.1} = \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} = 0.19$ (Azzalini, 2014).

Table 1. Three bivariate distributions and their marginal distributions.

$M(\mathbf{x})$	$M_j(z_j)$	β_j
$\mathrm{MVN}(\mathbf{0}, \Sigma)$	$\Phi(\Phi^{-1}\{\alpha_j\} + z_j\}$	1
$\mathrm{MVT}(\mathbf{0}, \Sigma, 4)$	$t\{t^{-1}(\alpha_j, 4) + z_j, 4\}$	$\frac{d_t\{t^{-1}(\alpha_j,4),4\}}{\phi\{\Phi^{-1}(\alpha_j)\}}$
$\mathrm{MSN}(\mathbf{0}, \Sigma, \boldsymbol{s})$	$\mathrm{sn}\{\mathrm{sn}^{-1}(\alpha_j; 0, 1, \bar{s}_j) + z_j; 0, 1, \bar{s}_j\}$	$\frac{d_{\mathrm{sn}}\{\mathrm{sn}^{-1}(\alpha_j;0,1,\bar{s}_j);0,1,\bar{s}_j\}}{\phi\{\Phi^{-1}(\alpha_j)\}}$

Since we are mainly concerned with the estimation performance under moderate sample size, we compare the efficiency of these estimations by the sum of marginal square root of the mean square error (RMSE) of $\mathbf{z}_n$ after 20, 30 and 50 iterations, respectively.

3.2. Comparison with true Σ_1

First, we consider the comparison with $\mathbf{x}_1$ to be the true value or equivalently $\mathbf{z}_1 = \mathbf{0}$ and Σ_1 to be its corresponding true value given in (16).

For MRM, we let

$$A_n = \begin{pmatrix} \{n\dot{M}_1(0)\}^{-1} & \\ (n+1)^{-2}\{\dot{M}_1(0)\dot{M}_2(0)\}^{-1/2} & \\ (n+1)^{-2}\{\dot{M}_1(0)\dot{M}_2(0)\}^{-1/2} & \\ \{n\dot{M}_2(0)\}^{-1} & \end{pmatrix}, \tag{17}$$

which satisfies the conditions in (4). For eMRM, A_n is given in (14) with Σ_1 to be the true value. For RM, we set a_n to be the optimal value $\{n\dot{M}_j(0)\}^{-1}$ for the jth margin. And for eRM, the sequences a_n and b_n are obtained by (5) with Σ_1 replaced by $\mathrm{diag}(\Sigma_1)$. Then, we use (3) and (5) to obtain sequences for MRM and eMRM respectively and use (1) and (2) to obtain RM and eRM respectively. For all procedures, the binary responses y_{jn}, $j = 1, \ldots, p$, are obtained as Bernoulli variables with success probabilities $M_j(x_{jn})$, respectively.

Tables 2–4 respectively report the sum of marginal RMSEs of $\mathbf{z}_{21}$, $\mathbf{z}_{31}$ and $\mathbf{z}_{51}$ obtained by the four procedures under various values of α_1 and α_2 and the three models in Table 1. (The simulation size is 1000 throughout.) The sequential design of $\mathbf{z}_n$ corresponds to simultaneous estimates of two (α_1 and α_2) marginal quantiles of an unknown bivariate distribution. For example, one wants to simultaneously estimate the 30th percentiles of two dose response curves based on two possibly correlated agents.

Table 2. Sum of marginal RMSEs of $\mathbf{z}_{21}$ obtained by the four procedures.

		With $\mathbf{x}_1 = \boldsymbol{\theta}$ and true Σ_1				With estimated $\mathbf{x}_1$ and Σ_1			
(α_1, α_2)	Model	eMRM	eRM	MRM	RM	eMRM	eRM	MRM	RM
(0.1,0.1)	MVN	0.563	0.693	2.517	2.499	0.500	0.692	2.655	2.711
	MVT	0.793	0.860	2.670	2.790	0.751	0.883	2.719	2.672
	MSN	0.386	0.473	2.574	2.557	0.340	0.485	2.702	2.760
(0.3,0.3)	MVN	0.464	0.568	0.690	0.658	0.357	0.550	0.578	0.581
	MVT	0.575	0.621	0.852	0.874	0.440	0.630	0.834	0.819
	MSN	0.349	0.407	0.465	0.471	0.263	0.420	0.456	0.456
(0.5,0.5)	MVN	0.453	0.548	0.578	0.577	0.357	0.550	0.578	0.581
	MVT	0.522	0.581	0.632	0.650	0.389	0.599	0.632	0.640
	MSN	0.342	0.413	0.423	0.420	0.273	0.425	0.430	0.433
(0.1,0.5)	MVN	0.502	0.625	1.543	1.569	0.463	0.603	1.461	1.689
	MVT	0.661	0.708	1.667	1.684	0.604	0.735	1.585	1.700
	MSN	0.362	0.457	1.297	1.395	0.310	0.460	1.448	1.595
(0.1,0.9)	MVN	0.557	0.698	2.318	2.576	0.558	0.694	2.336	2.663
	MVT	0.779	0.865	2.457	2.632	0.811	0.907	2.503	2.730
	MSN	0.416	0.541	2.214	2.534	0.406	0.541	2.425	2.699
(0.25,0.75)	MVN	0.468	0.583	0.775	0.809	0.436	0.575	0.761	0.797
	MVT	0.589	0.653	0.967	1.015	0.508	0.665	0.984	1.033
	MSN	0.370	0.461	0.605	0.635	0.317	0.460	0.599	0.640

Table 3. Sum of marginal RMSEs of $\mathbf{z}_{31}$ obtained by the four procedures.

		With $\mathbf{x}_1 = \boldsymbol{\theta}$ and true Σ_1				With estimated $\mathbf{x}_1$ and Σ_1			
(α_1, α_2)	Model	eMRM	eRM	MRM	RM	eMRM	eRM	MRM	RM
(0.1,0.1)	MVN	0.488	0.562	2.491	2.201	0.458	0.577	2.449	2.485
	MVT	0.694	0.768	2.371	2.473	0.660	0.744	2.676	2.534
	MSN	0.341	0.403	2.226	2.371	0.313	0.398	2.567	2.667
(0.3,0.3)	MVN	0.405	0.471	0.557	0.558	0.353	0.473	0.558	0.547
	MVT	0.496	0.531	0.681	0.689	0.380	0.520	0.678	0.671
	MSN	0.291	0.338	0.355	0.351	0.240	0.345	0.357	0.357
(0.5,0.5)	MVN	0.394	0.459	0.478	0.468	0.328	0.458	0.463	0.473
	MVT	0.452	0.480	0.540	0.516	0.353	0.482	0.520	0.527
	MSN	0.304	0.339	0.355	0.352	0.246	0.342	0.349	0.344
(0.1,0.5)	MVN	0.436	0.520	1.434	1.407	0.432	0.516	1.402	1.445
	MVT	0.569	0.619	1.450	1.555	0.547	0.617	1.566	1.510
	MSN	0.322	0.375	1.244	1.346	0.277	0.374	1.351	1.462
(0.1,0.9)	MVN	0.474	0.582	1.909	2.424	0.508	0.590	2.274	2.444
	MVT	0.697	0.755	2.252	2.530	0.730	0.773	2.184	2.481
	MSN	0.376	0.461	1.996	2.260	0.369	0.460	2.181	2.399
(0.25,0.75)	MVN	0.420	0.481	0.617	0.642	0.379	0.480	0.629	0.640
	MVT	0.512	0.542	0.800	0.840	0.453	0.562	0.818	0.863
	MSN	0.321	0.373	0.460	0.479	0.278	0.374	0.455	0.477

Table 4. Sum of marginal RMSEs of $\mathbf{z}_{51}$ obtained by the four procedures.

		With $\mathbf{x}_1 = \boldsymbol{\theta}$ and true Σ_1				With estimated $\mathbf{x}_1$ and Σ_1			
(α_1, α_2)	Model	eMRM	eRM	MRM	RM	eMRM	eRM	MRM	RM
(0.1,0.1)	MVN	0.390	0.465	1.994	2.148	0.401	0.473	2.136	2.181
	MVT	0.573	0.629	2.162	2.214	0.582	0.644	2.329	2.223
	MSN	0.288	0.329	2.043	1.899	0.259	0.318	2.209	2.217
(0.3,0.3)	MVN	0.339	0.375	0.409	0.419	0.292	0.369	0.408	0.420
	MVT	0.405	0.421	0.530	0.522	0.332	0.410	0.508	0.510
	MSN	0.240	0.265	0.272	0.272	0.209	0.261	0.268	0.268
(0.5,0.5)	MVN	0.312	0.350	0.362	0.363	0.275	0.348	0.360	0.367
	MVT	0.359	0.381	0.394	0.383	0.296	0.382	0.389	0.402
	MSN	0.239	0.266	0.264	0.262	0.210	0.270	0.270	0.262
(0.1,0.5)	MVN	0.352	0.401	1.193	1.270	0.361	0.402	1.270	1.303
	MVT	0.480	0.501	1.301	1.252	0.467	0.499	1.242	1.351
	MSN	0.261	0.283	1.179	1.212	0.237	0.290	1.214	1.269
(0.1,0.9)	MVN	0.407	0.459	1.809	2.026	0.439	0.454	1.960	2.227
	MVT	0.556	0.626	2.052	2.253	0.620	0.643	2.088	2.400
	MSN	0.316	0.360	1.762	2.092	0.319	0.354	1.927	2.240
(0.25,0.75)	MVN	0.342	0.377	0.458	0.473	0.325	0.375	0.455	0.471
	MVT	0.413	0.450	0.631	0.652	0.386	0.440	0.586	0.634
	MSN	0.266	0.291	0.331	0.336	0.239	0.288	0.325	0.339

We summarise the finding as follows. (i) The proposed efficient multivariate Robbins–Monro procedure (5) has significant improvement over the general multivariate version of (3). The reduction in terms of the sum of marginal RMSEs is $\sim$78.7% when the joint estimators are concern with extreme marginal quantile (the first three cases in Tables 2–4). The reduction is still remarkable (by $\sim$15.5%) for the median.

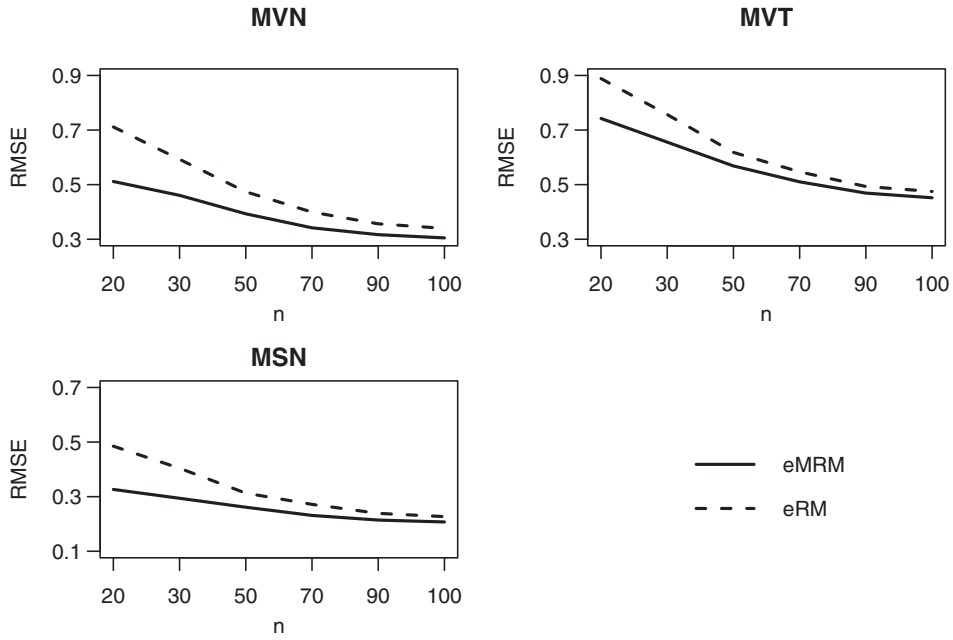

Figure 1. Sum of marginal RMSEs by eMRM and eRM under the three models.

The averaged reduction is 49.2% across all 18 cases. This type of improvement is also seen in comparison between eRM and RM, as reported by Joseph (2004). (ii) The proposed eMRM procedure uniformly outperforms the eRM in terms of reduction in the sum of marginal RMSEs by an average (over 18 cases) of 15.5% for $\mathbf{x}_{21}$, 12.1% for $\mathbf{x}_{31}$ and 9.6% for $\mathbf{x}_{51}$. This exactly shows the efficiency gained by the joint estimation with incorporation of the correlation. (The reduction percentage gets smaller as n increases since both procedures converge.) (iii) The results of MRM and RM are comparable since the A_n we chose in (17) only guarantees the convergence in large sample. This also reflects the importance of an appropriate selection of coefficient matrix A_n in the sequential design when the sample size is limited.

We continue the sequential experiments up to 100 steps to examine the convergence behaviour of the procedures. We use the first combination of $(\alpha_1, \alpha_2) = (0.1, 0.1)$ as an example to illustrate. Figure 1 shows the declining trend of the sum of marginal RMSEs by eMRM and eRM under the three models. (Those by MRM and RM are significantly larger, thus not included.) It is seen that the superiority of the joint estimation prevails with a remarkable difference.

3.3. Comparison with estimated Σ_1

The previous simulations are carried out under the perfect initial guess and the true M. Now we consider the situation that these values are unknown. We propose using a pilot sample of $\mathbf{x}$ to estimate them. To be specific, we (i) estimate the initial value of x_{j1} by the sample α_j-quantile; (ii) estimate Σ_1 based on a bootstrap sample (of size 500) of $\mathbf{x}_1$; and (iii) approximate $\dot{M}_j(0)$ by its normal counterpart, i.e. $\phi(\Phi^{-1}(\alpha_j))$, hence $\beta_j = 1$. The size of the pilot sample depends on the dimension of $\mathbf{x}$. For the bivariate models considered in Section 3.1, we set the size to be 20. Noted that this pilot sample does not need responses of ys so that in reality it is feasible, as observing responses $\mathbf{y}$ can be expensive or time consuming. When the dimension increases, the size of the pilot sample should increase as well. After obtaining $\mathbf{x}_1$ and A_n, we proceed the four competing procedures in the same way as outlined in Section 3.2.

We carry out the similar comparison under the same models and obtain results in the last four columns of Tables 2–4. It is seen that the reduction in the sum of marginal RMSEs are even larger for eMRM both from MRM (by 58.9% for $\mathbf{x}_{21}$, by 56.3% for $\mathbf{x}_{31}$ and by 51.9% for $\mathbf{x}_{51}$) and from eRM (by 26.9% for $\mathbf{x}_{21}$, by 21.1% for $\mathbf{x}_{31}$ and by 14.6% for $\mathbf{x}_{51}$) in average across all 18 cases.

Certainly larger sample size of the pilot study can yield more accurate estimation of Σ_1 and hence better result in sequential estimation. Here our simulation shows that a pilot study of a moderate sample (without responses) serves the purpose for providing initial information of $\mathbf{x}_1$ and Σ_1 in practice. On the other hand, if Σ_1 is mis-specified, e.g. use negative values for positive correlations, we found the performance of the sequential estimation of eMRM is worse than those of the separate eRM procedures, though the sequence still converges (result not shown). This indicates that correct estimation of the sign of the correlation is important.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [grant number 11271134] and the 111 Project (B14019) of Ministry of Education of China.

Notes on contributors

Cui XIong is a PhD student at the school of statistics, East China Normal University. She is currently a biostatistician at GlaxoSmithKline.

Jin Xu is a professor at the school of statistics, East China Normal University.

References

Anbar, D. (1978). A stochastic Newton–Raphson method. *Journal of Statistical Planning and Inference*, *2*, 153–163.

Azzalini, A. (2014). *The skew-normal and related families*. Cambridge: Cambridge University Press.

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B*, *61*, 579–602.

Chaloner, K., & Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, *21*, 191–208.

Chaudhuri, P., & Mykland, P. A. (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, *88*, 538–546.

Cheung, Y. K. (2010). Stochastic approximation and modern model-based designs for dose-finding clinical trials. *Statistical Science*, *25*, 191–201.

Chung, K. L. (1954). On a stochastic approximation method. *Annals of Mathematical Statistics*, *25*, 463–483.

Dror, H. A., & Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, *48*, 520–529.

Dror, H. A., & Steinberg, D. M. (2008). Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association*, *103*, 288–298.

Duflo, M. (1997). *Random iterative models*. Berlin: Springer-Verlag.

Hung, Y., & Joseph, V. R. (2014). Discussion of "Three-phase optimal design of sensitivity experiments" by Wu and Tian. *Journal of Statistical Planning and Inference*, *149*, 16–19.

Joseph, V. R. (2004). Efficient Robbins–Monro procedure for binary data. *Biometrika*, *91*, 461–470.

Kotz, S., & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge: Cambridge University Press.

Lai, T. L., & Robbins, H. (1979). Adaptive design and stochastic approximation. *Annals of Statistics*, *7*, 1196–1221.

Neyer, B. T. (1994). A D-optimality-based sensitivity test. *Technometrics*, *36*, 61–70.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407.

Ruppert, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Annals of Statistics*, *13*, 236–245.

Ruppert, D. (1988). *Efficient estimators from a slowly convergent Robbins–Monro process*. School of Operations Research and Industrial Engineering Technical Report, 781. Cornell University, Ithaca, NY.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Annals of Mathematical Statistics*, *29*, 373–405.

Wei, C. Z. (1987). Multivariate adaptive stochastic approximation. *Annals of Statistics*, *15*, 1115–1130.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society, Series B*, *25*, 1–48.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association*, *80*, 974–984.

Wu, C. F. J. (1986). Maximum likelihood recursion and stochastic approximation in sequential designs. In J. V. Ryzin (Ed.), *IMS monograph series: Vol. 8. Adaptive statistical procedures and related topics* (pp. 298–314). Hayward, CA: Institute of Mathematical Statistics.

Wu, C. F. J., & Tian, Y. (2014). Three-phase optimal design of sensitivity experiments. *Journal of Statistical Planning and Inference*, *149*, 1–15.

Young, L. J., & Easterling, R. G. (1994). Estimation of extreme quantiles based on sensitivity tests: A comparative study. *Technometrics*, *36*, 48–60.

Appendices

Appendix 1. Proof of Theorem 2.1

First, by (3), we have

$$\begin{aligned}
&\mathrm{E}\{(\mathbf{x}_{n+1}-\boldsymbol{\theta})^{\top}(\mathbf{x}_{n+1}-\boldsymbol{\theta})\}\\
&\quad=\mathrm{E}\Big(\mathrm{E}\Big[\{\mathbf{x}_n-\boldsymbol{\theta}-A_n(\mathbf{y}_n-\boldsymbol{\alpha})\}^{\top}\\
&\quad\times\{\mathbf{x}_n-\boldsymbol{\theta}-A_n(\mathbf{y}_n-\boldsymbol{\alpha})\}\mid\mathbf{x}_n\Big]\Big)\\
&\quad=\mathrm{E}(\mathbf{x}_n-\boldsymbol{\theta})^{\top}(\mathbf{x}_n-\boldsymbol{\theta})-2\mathrm{E}(\mathbf{x}_n-\boldsymbol{\theta})^{\top}A_n(\mathbf{m}_n^x-\boldsymbol{\alpha})\\
&\quad+\mathrm{E}\left\{(\mathbf{y}_n-\boldsymbol{\alpha})^{\top}A_n^{\top}A_n(\mathbf{y}_n-\boldsymbol{\alpha})\right\}\\
&\quad=\mathrm{E}(\mathbf{x}_1-\boldsymbol{\theta})^{\top}(\mathbf{x}_1-\boldsymbol{\theta})-2\sum_{i=1}^{n}\mathrm{E}(\mathbf{x}_i-\boldsymbol{\theta})^{\top}A_i(\mathbf{m}_i^x-\boldsymbol{\alpha})\\
&\quad+\sum_{i=1}^{n}\mathrm{E}\left\{(\mathbf{y}_i-\boldsymbol{\alpha})^{\top}A_i^{\top}A_i(\mathbf{y}_i-\boldsymbol{\alpha})\right\}\\
&\quad\geq 0,
\end{aligned}$$

where the expectations are taken with respect to $\mathbf{x}_n$, $\mathbf{m}_n^x=\mathrm{E}(\mathbf{y}_n\mid\mathbf{x}_n)=(M_1(x_{1n}),\ldots,M_p(x_{pn}))^{\top}$. Let $\mathbf{e}_i=(e_{1i},\ldots,e_{pi})^{\top}=\mathbf{y}_i-\boldsymbol{\alpha}$. Clearly, e_{ji} is bounded. Observe that $\sum_{i=1}^{n}\mathrm{E}\{(\mathbf{y}_i-\boldsymbol{\alpha})^{\top}A_i^{\top}A_i(\mathbf{y}_i-\boldsymbol{\alpha})\}$ can be expressed as

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\left\{\mathrm{E}(e_{ji}^2)\sum_{s=1}^{p}a_{sj,i}^2\right\}+\sum_{j\neq k}\sum_{i=1}^{n}\left\{\mathrm{E}(e_{ji}e_{ki})\sum_{s=1}^{p}a_{sj,i}a_{sk,i}\right\}.$$

By assumption (4), both series $\sum_{i=1}^{n}\{\mathrm{E}(e_{ji}^2)\sum_{s=1}^{p}a_{sj,i}^2\}$ and $\sum_{i=1}^{n}\{\mathrm{E}(e_{ji}e_{ki})\sum_{s=1}^{p}a_{sj,i}a_{sk,i}\}$ converge absolutely and hence converge. Thus, the series $\sum_{i=1}^{n}\mathrm{E}\{(\mathbf{y}_i-\boldsymbol{\alpha})^{\top}A_i^{\top}A_i(\mathbf{y}_i-\boldsymbol{\alpha})\}$ converges (to a non-negative value). Therefore, $\sum_{n=1}^{\infty}\mathrm{E}\{(\mathbf{x}_n-\boldsymbol{\theta})^{\top}A_n(\mathbf{m}_n^x-\boldsymbol{\alpha})\}<\infty$.

Second, for $j=1,\ldots,p$, since M_j is a distribution function and $\dot{M}_j(\theta_j)>0$, there exist positive constants ℓ_j and u_j such that

$$0<\ell_j\leq\frac{M_j(x_j)-\alpha_j}{x_j-\theta_j}\leq u_j<\infty,$$

for all x_j. Let $\ell = \min\{\ell_j : j = 1, \ldots, p\}$ and $u = \max\{u_j : j = 1, \ldots, p\}$. Let $\mathbf{a}_{j,n}$ be the jth column of A_n. Then,

$$\begin{aligned}
&(\mathbf{x}_n - \boldsymbol{\theta})^\top A_n(\mathbf{m}_n^x - \boldsymbol{\alpha}) \\
&= \sum_{j=1}^{p} (\mathbf{x}_n - \boldsymbol{\theta})^\top \mathbf{a}_{j,n}(x_{jn} - \theta_j)\frac{M_j(x_{jn}) - \alpha_j}{x_{jn} - \theta_j} \\
&\geq \sum_{j=1}^{p} (\mathbf{x}_n - \boldsymbol{\theta})^\top \mathbf{a}_{j,n}(x_{jn} - \theta_j)\delta_{jn} \\
&= (\mathbf{x}_n - \theta)^\top A_n \Delta_n (\mathbf{x}_n - \theta),
\end{aligned}$$

where $\Delta_n = \operatorname{diag}(\delta_{1n}, \ldots, \delta_{pn})$ with

$$\delta_{jn} = \begin{cases} \ell, & \text{if } (\mathbf{x}_n - \boldsymbol{\theta})^\top \mathbf{a}_{j,n}(x_{jn} - \theta_j) \geq 0, \\ u, & \text{otherwise.} \end{cases} \tag{A1}$$

Then, we have $\sum_{n=1}^{\infty} \mathrm{E}(\mathbf{x}_n - \theta)^\top A_n \Delta_n (\mathbf{x}_n - \theta) < \infty$. Since this quantity can be expressed as

$$\begin{aligned}
&\sum_{j=1}^{p}\sum_{n=1}^{\infty} a_{jj,n}\delta_{jn}\mathrm{E}(x_{jn} - \theta_j)^2 \\
&\quad + \sum_{j\neq k}\sum_{n=1}^{\infty} a_{jk,n}\delta_{kn}\mathrm{E}(x_{jn} - \theta_j)(x_{kn} - \theta_k).
\end{aligned} \tag{A2}$$

Observe that the second term of (A2) is no greater than

$$\begin{aligned}
&\sum_{j\neq k}\sum_{n=1}^{\infty} \frac{1}{2}|a_{jk,n}\delta_{kn}| \left\{\mathrm{E}(x_{jn} - \theta_j)^2 + \mathrm{E}(x_{kn} - \theta_k)^2\right\} \\
&= \frac{1}{2}\sum_{j=1}^{p}\sum_{n=1}^{\infty} \mathrm{E}(x_{jn} - \theta_j)^2 \sum_{k\neq j}\{|a_{jk,n}\delta_{kn}| + |a_{kj,n}\delta_{jn}|\}.
\end{aligned}$$

The assumptions $|a_{jk,n}/a_{jj,n}| \to 0$ when $n \to \infty$ for $j \neq k$ and $a_{jj,n} > 0$ imply that there exists an integer N such that for $n > N$, $|a_{jk,n}| < a_{jj,n}/\{2c(p-1)\}$ for all $k \neq j$, where c is some constant great than $u/(2\ell)$ and u and ℓ are defined in (A1). Hence, for fixed j

$$\begin{aligned}
&\sum_{k\neq j} |a_{jk,n}\delta_{kn}| + |a_{kj,n}\delta_{jn}| < \sum_{k\neq j}\left\{\frac{a_{jj,n}\delta_{kn}}{2c(p-1)} + \frac{a_{jj,n}\delta_{jn}}{2c(p-1)}\right\} \\
&\quad < \frac{a_{jj,n}u}{c}.
\end{aligned}$$

Then, the infinite sum

$$\begin{aligned}
&\sum_{j=1}^{p}\sum_{n>N}^{\infty} a_{jj,n}\delta_{jn}\mathrm{E}(x_{jn} - \theta_j)^2 \\
&\quad + \sum_{j\neq k}\sum_{n>N}^{\infty} a_{jk,n}\delta_{kn}\mathrm{E}(x_{jn} - \theta_j)(x_{kn} - \theta_k) \\
&\geq \sum_{j=1}^{p}\sum_{n>N}^{\infty} a_{jj,n}\delta_{jn}\mathrm{E}(x_{jn} - \theta_j)^2 - \frac{1}{2}\sum_{j=1}^{p}\sum_{n>N}^{\infty} \mathrm{E}(x_{jn} - \theta_j)^2 \\
&\quad \times \sum_{k\neq j}\{|a_{jk,n}\delta_{kn}| + |a_{kj,n}\delta_{jn}|\} \\
&> \sum_{j=1}^{p}\sum_{n>N}^{\infty} \left(\delta_{jn} - \frac{u}{2c}\right) a_{jj,n}\mathrm{E}(x_{jn} - \theta_j)^2,
\end{aligned}$$

which is positive by the choice of c. This implies that (A2) is also bounded from below. Thus, the first term of (A2) is finite. And by (4), $\mathrm{E}(x_{jn} - \theta_j)^2$ must converge to 0 for all $j = 1, \ldots, p$, which implies $\mathbf{x}_n$ converges to $\boldsymbol{\theta}$ in probability.

Appendix 2. Proof of Theorem 2.2

The proof is similar to the proof of Theorem 2.1. First, by (3), we have

$$\begin{aligned}
&\mathrm{E}\{(\mathbf{x}_{n+1} - \boldsymbol{\theta})^\top (\mathbf{x}_{n+1} - \boldsymbol{\theta})\} \\
&\quad = \mathrm{E}\{(\mathbf{x}_1 - \boldsymbol{\theta})^\top (\mathbf{x}_1 - \boldsymbol{\theta})\} \\
&\qquad - 2\sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\mathbf{m}_i^x - \mathbf{b}_i)\} \\
&\qquad + \sum_{i=1}^{n} \mathrm{E}\left\{(\mathbf{y}_i - \mathbf{b}_i)^\top A_i^\top A_i(\mathbf{y}_i - \mathbf{b}_i)\right\} \\
&\quad \geq 0,
\end{aligned}$$

where $\mathbf{m}_n^x$ is as defined before. The finiteness of $\sum_{n=1}^{\infty} \mathrm{E}[(\mathbf{y}_n - \mathbf{b}_n)^\top A_n^\top A_n(\mathbf{y}_n - \mathbf{b}_n)]$ (showed by the same way as for the case $\mathbf{b}_n = \boldsymbol{\alpha}$ in the proof of Theorem 2.1) implies that $\sum_{n=1}^{\infty} \mathrm{E}\{(\mathbf{x}_n - \boldsymbol{\theta})^\top A_n(\mathbf{m}_n^x - \mathbf{b}_n)\} < \infty$.

Second, similar to the proof in Theorem 2.1, we can show that

$$(\mathbf{x}_n - \boldsymbol{\theta})^\top A_n(\mathbf{m}_n^x - \boldsymbol{\alpha}) \geq (\mathbf{x}_n - \boldsymbol{\theta})^\top A_n \Delta_n (\mathbf{x}_n - \boldsymbol{\theta}),$$

where $\Delta_n = \operatorname{diag}(\delta_{1n}, \ldots, \delta_{pn})$ and δ_{jn} are some positive constants depending on A_n. Then,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathrm{E}(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\mathbf{m}_i^x - \mathbf{b}_i) \\
&\quad = \sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\mathbf{m}_i^x - \boldsymbol{\alpha})\} \\
&\qquad + \sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\boldsymbol{\alpha} - \mathbf{b}_i)\} \\
&\quad \geq \sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i \Delta_i(\mathbf{x}_i - \boldsymbol{\theta})\} \\
&\qquad + \sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\boldsymbol{\alpha} - \mathbf{b}_i)\}.
\end{aligned}$$

Since $\mathbf{x}_n = \mathbf{x}_1 + \sum_{s=1}^{n-1} A_s(\mathbf{y}_s - \mathbf{b}_s)$ and $\mathrm{E}(\mathbf{x}_1) = \boldsymbol{\theta}$ (by assumption), we have

$$\begin{aligned}
&\left|\sum_{i=1}^{n} \mathrm{E}\{(\mathbf{x}_i - \boldsymbol{\theta})^\top A_i(\boldsymbol{\alpha} - \mathbf{b}_i)\}\right| \\
&= \left|\sum_{i=2}^{n}\sum_{s=1}^{i-1} \mathrm{E}\{(\mathbf{y}_s - \mathbf{b}_s)^\top A_s^\top A_i(\boldsymbol{\alpha} - \mathbf{b}_i)\}\right| \\
&\leq \sum_{i=2}^{n}\sum_{s=1}^{i-1} \mathbf{1}^\top \mid A_s^\top A_i(\boldsymbol{\alpha} - \mathbf{b}_i) \mid_{\mathrm{e}},
\end{aligned}$$

which is finite by the assumption in (6). Thus, $\sum_{n=1}^{\infty} \mathrm{E}\{(\mathbf{x}_n - \boldsymbol{\theta})^\top A_n \Delta_n (\mathbf{x}_n - \boldsymbol{\theta})\} < \infty$. Finally, using the same argument in the proof of Theorem 2.1, $\mathbf{x}_n$ converges to $\boldsymbol{\theta}$ in probability.

Appendix 3. Proof of Theorem 2.3

First, under the assumption that M is given by (10), (15) is the covariance of $\mathbf{z}_{n+1}$. We have

$$\sigma_{jj,n+1} = \sigma_{jj,n} - \mathrm{E}(z_{jn}\mathbf{g}_n^\top)\mathrm{cov}^{-1}(\mathbf{y}_n)\mathrm{E}(\mathbf{g}_n z_{jn}).$$

Recall that the expectations are taken with respect to $\mathbf{z}_n$ with distribution $\Phi_p(\mathbf{z}; \mathbf{0}, \Sigma_n)$.

Let

$$D_n = \operatorname{diag}\left[\frac{\beta_1\sigma_{11,n}^{1/2}}{(1+\beta_1^2\sigma_{11,n})^{1/2}}\phi\left\{\frac{\Phi^{-1}(\alpha_1)}{(1+\beta_1^2\sigma_{11,n})^{1/2}}\right\},\ldots,\right.$$
$$\left.\frac{\beta_p\sigma_{pp,n}^{1/2}}{(1+\beta_p^2\sigma_{pp,n})^{1/2}}\phi\left\{\frac{\Phi^{-1}(\alpha_p)}{(1+\beta_p^2\sigma_{pp,n})^{1/2}}\right\}\right],$$

$\rho_{jk,n} = \sigma_{jk,n}/(\sigma_{jj,n}\sigma_{kk,n})^{1/2}$, $\boldsymbol{\rho}_{j,n} = (\rho_{j1,n},\ldots,\rho_{jp,n})^\top$. Let

$$\boldsymbol{\gamma}_{j,n} = \boldsymbol{\gamma}_{j,n}(\sigma_{11,n},\ldots,\sigma_{pp,n}) = D_n\boldsymbol{\rho}_{j,n} \tag{A3}$$

be a function of $\sigma_{11,n},\ldots,\sigma_{pp,n}$. Since $\boldsymbol{\gamma}_{j,n} \neq \mathbf{0}$ as its jth element is not zero, we have $0 < 1 - \boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n} < 1$ and

$$0 < \sigma_{jj,n+1} = \sigma_{jj,n}\{1 - \boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n}\}$$
$$= \sigma_{jj,1}\prod_{i=1}^{n}\left\{1 - \boldsymbol{\gamma}_{j,i}^\top \operatorname{cov}^{-1}(\mathbf{y}_i)\boldsymbol{\gamma}_{j,i}\right\} < \sigma_{jj,n}.$$

Thus, $\sigma_{jj,n}$ is strictly decreasing to a limit σ_{jj}. Since σ_{jj} satisfies the equation $\sigma_{jj} = \sigma_{jj}\{1 - q(\sigma_{jj})\}$ where q is a non-negative function of $\sigma_{11},\ldots,\sigma_{pp}$. The unique solution is $\sigma_{jj} = 0$ for all $j = 1,\ldots,p$. This implies $\Sigma_n \to \mathbf{0}$ and by (11) $\mathbf{b}_n \to \boldsymbol{\alpha}$.

Appendix 4. Proof of Theorem 2.4

(i) First, we show that $\sigma_{jj,n} = \mathrm{O}(n^{-1})$ for $j = 1,\ldots,p$. By Theorem 2.3, we have $\sigma_{jj,n} \to 0$ and $\operatorname{cov}(\mathbf{y}_n) \to \operatorname{diag}\{\alpha_1(1-\alpha_1),\ldots,\alpha_p(1-\alpha_p)\}$ hence $\operatorname{cov}^{-1}(\mathbf{y}_n) \to \operatorname{diag}[\{\alpha_1(1-\alpha_1)\}^{-1},\ldots,\{\alpha_p(1-\alpha_p)\}^{-1}]$. Then, for sufficiently large n, there exist positive constants $c_1,\ldots,c_p$ such that

$$\sigma_{jj,n} = \frac{c_j}{n} \quad \text{for } j = 1,\ldots,p, \quad \text{and}$$
$$\boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n} = \frac{1}{n+1},$$

where $\boldsymbol{\gamma}_{j,n}$ is defined in (A3). It is clear that

$$\boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n} = \boldsymbol{\gamma}_{j,n}\left(\frac{c_1}{n},\ldots,\frac{c_p}{n}\right)^\top$$
$$\times \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n}\left(\frac{c_1}{n},\ldots,\frac{c_p}{n}\right)$$
$$< \boldsymbol{\gamma}_{j,n}\left(\frac{c_1}{n},\ldots,\frac{c_{j-1}}{n},\frac{2c_j}{n},\frac{c_{j+1}}{n},\ldots,\frac{c_p}{n}\right)^\top$$
$$\times \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n}\left(\frac{c_1}{n},\ldots,\frac{c_{j-1}}{n},\frac{2c_j}{n},\frac{c_{j+1}}{n},\ldots,\frac{c_p}{n}\right). \tag{A4}$$

Now, treat

$$\sigma_{jj,n+1} = \sigma_{jj,n}\{1 - \boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{j,n}\} = h_j(\sigma_{jj,n})$$

as a function of $\sigma_{jj,n}$ given $\sigma_{11,n},\ldots,\sigma_{j-1,j-1,n},\sigma_{j+1,j+1,n},\ldots,\sigma_{pp,n}$. Observe that $\partial h_j(\sigma_{jj,n})/\partial\sigma_{jj,n} > 0$ when $c_j/n \le \sigma_{jj,n} \le 2c_j/n$. Then, we get

$$\frac{c_j}{n+1} = h_j\left(\frac{c_j}{n}\right) = h_j\left(\sigma_{jj,n}\right) \le \sigma_{jj,n+1} < h_j\left(\frac{2c_j}{n}\right) < \frac{2c_j}{n+1},$$

where the last inequality is obtained from (A4). By mathematical induction, $\sigma_{jj,n}$ is$\mathrm{O}(n^{-1})$.

Further, by (15) and (A3), we have for $j \neq k$

$$\sigma_{jk,n+1} = (\sigma_{jj,n}\sigma_{kk,n})^{1/2}\left\{\rho_{jk,n} - \boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{k,n}\right\}.$$

Using the fact that $\operatorname{cov}^{-1}(\mathbf{y}_n) = \operatorname{diag}[\{\alpha_1(1-\alpha_1)\}^{-1},\ldots,\{\alpha_p(1-\alpha_p)\}^{-1}] + \mathrm{O}(n^{-1})$, we express

$$\boldsymbol{\gamma}_{j,n}^\top \operatorname{cov}^{-1}(\mathbf{y}_n)\boldsymbol{\gamma}_{k,n}$$
$$= \sum_{s=1}^{p}\rho_{js,n}\rho_{ks,n}\frac{\beta_s^2\sigma_{ss,n}\phi^2\{\Phi^{-1}(\alpha_s)\}}{(1+\beta_s^2\sigma_{ss,n})\alpha_s(1-\alpha_s)} + \mathrm{O}(n^{-1}),$$

which is $\mathrm{O}(n^{-1})$ from the previous result. Thus, if $\sigma_{jk,n}$ is $\mathrm{o}(n^{-1})$, $\sigma_{jk,n+1}$ is $\mathrm{o}(n^{-1})$. This can be achieved by properly choosing the starting value of Σ_1.

(ii) Second, we show that A_n in (14) satisfies (4). By the construction in (14), we have

$$a_{jk,n} = \mathrm{E}(z_{jn}g_{kn})\left[\{\alpha_j(1-\alpha_j)\}^{-1} + \mathrm{O}(n^{-1})\right].$$

Then, the results in (i) imply that $a_{jk,n}$ is $\mathrm{O}(n^{-1})$ for $j = k$ and $\mathrm{o}(n^{-1})$ for $j \neq k$.

(iii) Third, we show A_n together with $\mathbf{b}_n$ in (14) satisfies (6). By (14) and the results in (i), we have each component of $\boldsymbol{\alpha} - \mathbf{b}_n$ is $\mathrm{O}(n^{-1})$ and each component of $A_n(\boldsymbol{\alpha} - \mathbf{b}_n) \triangleq \boldsymbol{\ell}_n$ is $\mathrm{O}(n^{-2})$. Observe that

$$\sum_{s=1}^{n-1}\mathbf{1}_p^\top\left|A_s^\top A_n(\boldsymbol{\alpha} - \mathbf{b}_n)\right|_{\mathrm{e}} = \sum_{s=1}^{n-1}\sum_{k=1}^{p}\left|\sum_{j=1}^{p}a_{jk,s}\ell_{j,n}\right|$$
$$\le \sum_{k=1}^{p}\sum_{j=1}^{p}\sum_{s=1}^{n-1}|\,a_{jk,s}|.|\ell_{j,n}|,$$

which is $\mathrm{O}(\log(n)/n^2)$ after the result in (ii). Then (6) follows.

Combining (ii) and (iii), the result follows after Theorem 2.2.