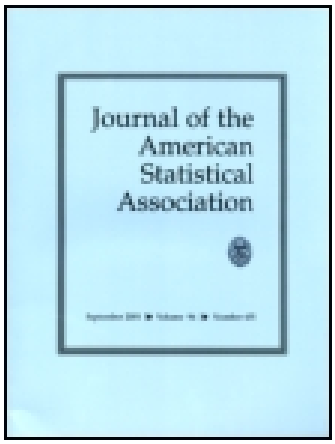

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Two-Stage Importance Sampling With Mixture Proposals

Wentao Li [a], Zhiqiang Tan [b] & Rong Chen [b]

[a] Department of Mathematics and Statistics, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4YF, United Kingdom

[b] Department of Statistics, Rutgers University, Piscataway, NJ, 08854

Accepted author version posted online: 24 Aug 2013.Published online: 19 Dec 2013.

To cite this article: Wentao Li, Zhiqiang Tan & Rong Chen (2013) Two-Stage Importance Sampling With Mixture Proposals, Journal of the American Statistical Association, 108:504, 1350-1365, DOI: 10.1080/01621459.2013.831980

To link to this article: http://dx.doi.org/10.1080/01621459.2013.831980

Two-Stage Importance Sampling With Mixture Proposals

Wentao Li, Zhiqiang Tan, and Rong Chen

For importance sampling (IS), multiple proposals can be combined to address different aspects of a target distribution. There are various methods for IS with multiple proposals, including Hesterberg's stratified IS estimator, Owen and Zhou's regression estimator, and Tan's maximum likelihood estimator. For the problem of efficiently allocating samples to different proposals, it is natural to use a pilot sample to select the mixture proportions before the actual sampling and estimation. However, most current discussions are in an empirical sense for such a two-stage procedure. In this article, we establish a theoretical framework of applying the two-stage procedure for various methods, including the asymptotic properties and the choice of the pilot sample size. By our simulation studies, these two-stage estimators can outperform estimators with naive choices of mixture proportions. Furthermore, while Owen and Zhou's and Tan's estimators are designed for estimating normalizing constants, we extend their usage and the two-stage procedure to estimating expectations and show that the improvement is still preserved in this extension.

KEY WORDS: Control variates; Normalizing constant; Pilot samples.

1. INTRODUCTION

Importance sampling (IS) is a very useful Monte Carlo method for approximating analytically intractable integrals. Normally IS technique is used to approximate two types of integrals:

1. $Z = \int \pi(x)dx$, where $\pi(x)$ is a nonnegative integrable function.
2. $\mu = \int h(x)\pi^*(x)dx$, where $\pi^*(x)$ is a probability density and $h(x)$ is a real function.

The integral Z can be treated as the normalizing constant of some probability density $\pi^*(x)$. This type of integration arises in many areas, including missing data analysis, marginal likelihood calculation, estimation of free energies in physics (Gelman and Meng 1998), and communication system (Smith, Shafi, and Gao 1997). The second type of integral can be treated as the expectation of $h(x)$ under $\pi^*(x)$. Applications of IS to this type of integral are also popular in many areas including rare event simulation (Denny 2001), reliability (Hesterberg 1995), computational finance (Owen and Zhou 1999), and computer graphics (Veach and Guibas 1995).

The idea of IS for approximating Z is based on the identity

$$\int \pi(x)dx = \int \frac{\pi(x)}{q(x)} q(x)dx,$$

where $q(x)$ is a probability density, called the proposal or trial density. Then Z can be treated as the expectation under density $q(x)$. With a sample $(x_1, \ldots, x_n)$ from $q(x)$, Z can be approximated by the sample average $n^{-1}\sum \pi(x_i)/q(x_i)$. For approximating μ, similar ideas hold. In this case, even when it is not difficult to generate observations from $\pi^*(x)$, IS estimator can have higher efficiency than using observations from $\pi^*(x)$, by appropriately choosing $q(x)$ so that it takes the shape of both $h(x)$ and $\pi^*(x)$ into account. In some applications, the improvement can be of several orders of magnitude. One example is given in Hesterberg (1995). In rare event simulation, $h(x)$ is an indicator function with support in the tail of $\pi^*(x)$. A trial distribution $q(x)$ that focuses on the "important" part of the integrand $h(x)\pi^*(x)$ will be more efficient than generating samples directly from $\pi^*(x)$.

However, in practice it is usually challenging to implement IS efficiently. One reason is that the complexity of the integrand makes it difficult to find $q(x)$ that covers all its important parts. For example, when the integrand is multimodal, a unimodal $q(x)$ will not be efficient. Some of such multimodal integrand can be found in Owen and Zhou (1999). Another well-known reason is that when $q(x)$ has a "lighter" tail than the integrand, that is, $\pi(x)/q(x)$ or $h(x)\pi^*(x)/q(x)$ are unbounded, IS estimator can have infinite variance. When there is a lack of knowledge of the integrand in some regions, unexpected large values of integrand may result in inaccurate results. See Ford and Gregory (2007) for an example.

For both problems, a general remedy is to consider multiple proposal distributions to address different aspects of the integrand. For multimodal integrands, Oh and Berger (1993) used a family of Student's t distributions and Owen and Zhou (1999) used a family of beta distributions to model each mode of integrand individually. West (1993) and Givens and Raftery (1996) used a kernel estimate of the integrand as the proposal, which is a mixture of normal or t distributions. Even for a unimodal target distribution, one can construct a mixture of two proposals where one mimics the center of target and the other dominates the tail. Such a construction was used in Giordani and Kohn (2010), although in a different scenario. The requirement that the tail of integrand needs to be dominated by the proposal distribution can be met by including some heavy-tailed distributions in the mixture as "protection." For example, Hesterberg

Wentao Li is a Senior Research Associate, Department of Mathematics and Statistics, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4YF, United Kingdom (E-mail: *w.li@lancaster.ac.uk*). Zhiqiang Tan is Associate Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854 (E-mail: *ztan@stat.rutgers.edu*). Rong Chen is Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854 (E-mail: *rongchen@stat.rutgers.edu*).
Chen's research is sponsored in part by NSF grants DMS 0800183, DMS 0905763, and DMS0915139. Tan's research is sponsored in part by NSF grant DMS 0749418. The authors thank the two editors, an associate editor, and two anonymous referees for their helpful comments.

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.831980

(1995) included the target distribution itself as one of the components to provide an upper bound for the estimation variance, and Owen and Zhou (2000) used uniform distribution to bound the sample weights in a bounded domain case. Liang, Liu, and Carroll (2007) divided the state domain into subregions and used the mixture of truncated target distributions in all subregions as the proposal, which leads to bounded importance weights. In Bayesian analysis, the prior density of the parameters can serve as the heavy-tailed component, as used in Ford and Gregory (2007).

Given multiple potentially useful proposals, a straightforward combination method is to use their mixture as the new proposal. This method has two issues. One is that the mixture proposal may contaminate the good components in the mixture. Owen and Zhou (2000) showed that a mixture can lose efficiency by several orders of magnitude if the original proposal is nearly perfect. Another problem is that the mixture proportions need to be determined. Proper mixture proportions can increase the efficiency by an order of magnitude, as shown in Emond, Raftery, and Steele (2001).

Owen and Zhou (2000) suggested a regression method to combine multiple proposals using control variates to deal with the contamination problem. Control variate is a useful technique for variance reduction (Rubinstein and Kroese 2008). Owen and Zhou's method has the property that it will not perform worse than using the best of component proposals individually, if the sample size assigned to the best component is the same as in the mixture case. Such a lower bound lessens the contamination problem. Tan (2004) proposed to use nonparametric maximum likelihood estimation (MLE) in place of regression and showed that the MLE method is the most efficient among several classes of estimators including those in Owen and Zhou (2000), Hesterberg (1995), and Veach and Guibas (1995).

To determine appropriate proportions, Fan et al. (2006) and Hesterberg (1995) followed some heuristic rules derived from experience or interpretation of proposals. A more sophisticated approach is to use a pilot study to determine the optimal proportions via minimizing some criterion. The estimated proportions are then used to generate the sample and construct the estimators. The criterion was selected to be the asymptotic variance of IS estimator with mixture proposal in Raghavan and Cox (1998), and the variation coefficient of pilot sample in Oh and Berger (1993). However, few theoretical properties have been investigated.

In this article, we propose a similar two-stage procedure and investigate its theoretical properties. In the first stage, pilot sample is drawn from a mixture proposal with predetermined proportions. The optimal mixture proportions are then estimated by minimizing the estimated asymptotic variance of Owen's regression estimator or Tan's MLE based on control variates. In the second stage, the sample is drawn from the mixture proposal with the estimated proportions. Integral estimators are constructed using all observations, including those from the pilot stage. Then we establish a theoretical framework of such a two-stage procedure. It is shown that under very weak conditions, the integral estimators constructed by the two-stage procedure are consistent and asymptotic normal with minimum asymptotic variance over all mixture proportions. Therefore, the two-stage procedure is adaptive toward using the optimal mixture proportions. The optimal sample size used for the pilot stage is also calculated in the sense of minimizing an approximated mean square error (MSE) in higher order. Furthermore, we extend Owen's regression estimator and Tan's MLE to ratio estimators (Rubinstein and Kroese 2008). When estimating μ, if one can evaluate $\pi^*(x)$ only up to a normalizing constant, a ratio estimator is used, with the numerator being the estimated unnormalized integral and the denominator being the estimated normalizing constant. We show that the two-stage procedure for this extension also has the desirable asymptotic properties.

The remainder of this article is organized as follows. Section 2 reviews in detail some existing techniques related to IS. Section 3 proposes the new two-stage procedure and establishes its theoretical framework. Section 4 provides the extension of Owen's regression estimator and Tan's MLE to ratio estimators and their two-stage procedures. In Section 5, we demonstrate the two-stage approach with several numerical examples including estimation of a rare event probability and Value at Risk (VaR) under a Bayesian GARCH model.

2. REVIEW OF IS TECHNIQUES

In Sections 2 and 3, we assume that all functions in integrals (1) and (2) can be evaluated exactly. Since

$$\mu = \int h(x)^+ \pi^*(x)dx - \int h(x)^- \pi^*(x)dx,$$

where $h(x)^+ = h(x)1_{\{h(x)\geq 0\}}$ and $h(x)^- = -h(x)1_{\{h(x)<0\}}$, the estimation of μ can be achieved by estimation of these two integrals, both with positive integrands. With this approach, μ becomes a special case of Z and therefore we only consider estimating Z in these two sections.

2.1 Mixture Importance Sampling

Assume observations $\{x_1, \ldots, x_n\}$ are taken iid from a proposal distribution $q(x)$. The integral $Z = \int \pi(x)dx$ can be estimated by

$$\widehat{Z}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i)}{q(x_i)}. \tag{1}$$

Under mild conditions, the asymptotic variance is $\text{var}_q[\pi(x)/q(x)]$, where var_q is the variance under distribution $q(x)$ (Robert and Casella 2004). The optimal proposal is $\pi(x)/Z$, suggesting that the proposal $q(x)$ should be chosen to mimic the shape of $\pi(x)$ so that the high- and low-density regions of $q(x)$ coincide with those of $\pi(x)$. With such a proposal, the majority of Monte Carlo sample from $q(x)$ fall in the high-density region of $\pi(x)$, the importance region. In some scenarios, more than one $q(x)$ may be needed. For example, for a multimodal $\pi(x)$, it is helpful to use several proposal distributions, each targeted at one importance region. Suppose $q_1(x), \ldots, q_p(x)$ are p probability densities serving as proposals. Given a mixture proportion vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ satisfying $\sum_{k=1}^{p} \alpha_k = 1$, we can use the mixture distribution as the proposal and estimate Z by

$$\widehat{Z}_{\text{MIS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i)}{q_{\boldsymbol{\alpha}}(x_i)}, \tag{2}$$

where $q_{\boldsymbol{\alpha}} = \sum_{k=1}^{p} \alpha_k q_k(x)$ and $\{x_1, \ldots, x_n\}$ are generated from q_α. In addition, the variance of $\widehat{Z}_{\mathrm{IS}}$ also demands that the ratio $\pi(X)/q(X)$ has a finite variance. A mixture distribution certainly makes it easy to satisfy such a condition as one can simply include a proposal distribution $q_1(X)$ having $\mathrm{var}[\pi(X)/q_1(X)] < \infty$, such as a uniform distribution if the domain is bounded. Such a proposal distribution sets an upper bound to the estimating variance and therefore plays the role of "safeguard" in IS, which is the key idea of defensive IS (Hesterberg 1988).

2.2 Stratified Sampling

Instead of generating samples directly from the mixture distribution as that in (2), stratified samples $\{x_{k1}, \ldots, x_{kn_k}\}$ can be taken with deterministic size $n_k = \alpha_k n$ from the kth proposal q_k, which leads to the estimator in Hesterberg (1988)

$$\widehat{Z}_{\mathrm{SIS}}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}. \tag{3}$$

Veach and Guibas (1995) considered the estimator

$$\sum_{k=1}^{p} \frac{1}{n_k} \sum_{i=1}^{n_k} \omega_k(x_{ki}) \frac{\pi(x_{ki})}{q_k(x_{ki})}, \tag{4}$$

where $\{\omega_k(x)\}_{k=1}^{p}$ is a group of coefficient functions for the sample weights and satisfies $\sum_{k=1}^{p} \omega_k(x) = 1$. They showed that $\widehat{Z}_{\mathrm{SIS}}$ is a suboptimal choice in this large class. Raghavan and Cox (1998) proposed a two-stage algorithm to construct $\widehat{Z}_{\mathrm{SIS}}$ with estimated optimal mixture proportions in the sense of minimizing the asymptotic variance.

2.3 Importance Sampling With Control Variates

One problem of using a mixture proposal distribution is the possible loss of efficiency due to mixing of good proposal distributions with poor ones (Owen and Zhou 2000). It is a premium to pay for the insurance of valid IS, but can be reduced by combining importance sampling and control variates. Given an unbiased estimator X_n of Z, improvement can be gained by constructing a proper control variate vector Y and using $X_n - \boldsymbol{\beta}^T(Y - E[Y])$ to estimate Z. The optimal $\boldsymbol{\beta}$ can be estimated using a regression approach to minimize asymptotic variance (Cochran 1977). In Owen and Zhou (2000), combining $\widehat{Z}_{\mathrm{SIS}}$ and control variates $\boldsymbol{g}(x) = (q_2(x) - q_1(x), \ldots, q_p(x) - q_1(x))^T$ results in the estimator

$$\widehat{Z}_{\mathrm{Reg}}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} \frac{\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}, \tag{5}$$

where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \widetilde{\mathrm{var}}\left[\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right]^{-1} \widetilde{\mathrm{cov}}^T\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right],$$

and $\widetilde{\mathrm{var}}$ and $\widetilde{\mathrm{cov}}$ denote the pooled-sample variance and covariance. There are two appealing properties of $\widehat{Z}_{\mathrm{Reg}}$. First, its asymptotic variance is zero when $\pi(x)$ is a linear combination of the proposals. Second, $\widehat{Z}_{\mathrm{Reg}}$ has smaller asymptotic variance than every IS estimator constructed solely with q_k with n_k samples, $k = 1, \ldots, p$. That is, $\widehat{Z}_{\mathrm{Reg}}$ is always at least as good as the best one among the individual proposals.

2.4 Likelihood Approach

All previous integration methods directly approximate the target integrals. On the other hand, in Kong et al. (2003), Monte Carlo integration is treated as a statistical inference problem where the Monte Carlo sample serves as observations, the underlying measure in target integral, usually Lebesgue measure or counting measure, is treated as an unknown nonnegative measure, and the Monte Carlo sample is modeled using a semiparametric model. Then by nonparametric maximum likelihood, the unknown measure is estimated by a discrete measure with the Monte Carlo sample as support, and the target integral is estimated by the integration over the discrete measure. As an example, with $\{x_1, \ldots, x_n\}$ generated identically and independently from q_1 under Lebesgue measure, the model assumes that x_i is distributed as $q_1(x)d\nu / \int q_1(x)d\nu$, where ν is an unknown nonnegative measure. The nonparametric maximum likelihood estimator of ν is

$$\widehat{\nu} \propto \frac{\widehat{P}(\{x\})}{q_1(x)},$$

where $\widehat{P}$ has the support on $\{x_1, \ldots, x_n\}$ with mass n^{-1} at each point. Then $Z = \int q(x)d\nu$ can be estimated by

$$\frac{\int q_1(x)d\widehat{\nu}}{\int q(x)d\widehat{\nu}} = \frac{1}{n} \sum_{i=1}^{n} \frac{q(x_i)}{q_1(x_i)}.$$

This is the same as the IS estimator with proposal distribution $q_1(x)$.

Given multiple proposals $q_1, \ldots, q_p$ and control variates $\boldsymbol{g}(x)$, Tan (2004) proposed to restrict the measure ν in the set $\{\nu : \int q_k(x)d\nu = \int q_1(x)d\nu,\ k = 1, \ldots, p\}$. The nonparametric MLE of ν under such a restriction is

$$\widehat{\nu} \propto \frac{\widehat{P}(\{x\})}{q_{\boldsymbol{\alpha}}(x) + \widehat{\boldsymbol{\zeta}}^T \boldsymbol{g}(x)},$$

where

$$\widehat{\boldsymbol{\zeta}} = \operatorname*{argmax}_{\boldsymbol{\zeta}} \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log\left[q_{\boldsymbol{\alpha}}(x_{ki}) + \boldsymbol{\zeta}^T \boldsymbol{g}(x_{ki})\right],$$

and the integral estimator is given by

$$\widehat{Z}_{\mathrm{MLE}}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widehat{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})}. \tag{6}$$

It is shown that $\widehat{Z}_{\mathrm{Reg}}$ is a first-order approximation of $\widehat{Z}_{\mathrm{MLE}}$ and hence has the same asymptotic efficiency. The estimator $\widehat{Z}_{\mathrm{MLE}}$ also achieves the highest asymptotic efficiency among the class of estimators in the form of

$$\sum_{k=1}^{p} \frac{1}{n_k} \sum_{i=1}^{n_k} \omega_k(x_{ki}) \frac{\pi(x_{ki}) - \boldsymbol{\beta}_k^T(x_{ki}) \boldsymbol{g}(x_{ki})}{q_k(x_{ki})}, \tag{7}$$

where $\omega_1(x), \ldots, \omega_p(x)$ and $\boldsymbol{\beta}_1(x), \ldots, \boldsymbol{\beta}_p(x)$ satisfy that $\omega_k(x) = 0$ when $q_k(x) = 0$, $\sum_{i=1}^{p} \omega_k(x) = 1$ and $\sum_{i=1}^{p} \omega_k(x)\boldsymbol{\beta}_k(x) = \boldsymbol{b}$ for some constant vector $\boldsymbol{b}$, and therefore dominates the class of estimators in (4).

Because $\widehat{Z}_{\mathrm{Reg}}$ and $\widehat{Z}_{\mathrm{MLE}}$ dominate the other estimators, we will only discuss the two-stage procedure for these two estimators. Furthermore, there is another important benefit of using $\widehat{Z}_{\mathrm{Reg}}$ and $\widehat{Z}_{\mathrm{MLE}}$ in that their asymptotic variance is a

convex function of $\boldsymbol{\alpha}$ and hence can be easily minimized. See Remark 1.

3. TWO-STAGE PROCEDURE

3.1 The Algorithm

Suppose p proposal distributions $q_1, \ldots, q_p$ are given and the sample size is budgeted at n. Let $\Theta = [\delta, 1-\delta]^p$, where δ is some constant close to 0. The following algorithm is proposed to select mixture proportions $\boldsymbol{\alpha}$ and construct estimators:

1. First stage: Given a p-dimensional vector $\boldsymbol{\gamma}$ satisfying $\sum_{k=1}^p \gamma_k = 1$, generate n_0 independent stratified observations $\{x_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\gamma}}(x) = \sum_{k=1}^p \gamma_k q_k(x)$, that is, $n_0\gamma_k$ observations from $q_k(x)$, $k = 1, \ldots, p$. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing

$$\widehat{\sigma}_Z^2(\boldsymbol{\alpha}) = \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\left[\pi(x_i) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(x_i)\right]^2}{q_{\boldsymbol{\alpha}}(x_i) q_{\boldsymbol{\gamma}}(x_i)}, \tag{8}$$

where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\boldsymbol{g}(x_i)\boldsymbol{g}(x_i)^T}{q_{\boldsymbol{\alpha}}(x_i) q_{\boldsymbol{\gamma}}(x_i)}\right)^{-1} \left(\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\pi(x_i)\boldsymbol{g}(x_i)}{q_{\boldsymbol{\alpha}}(x_i) q_{\boldsymbol{\gamma}}(x_i)}\right)$$

and $\boldsymbol{g}(x) = (q_2(x) - q_1(x), \ldots, q_p(x) - q_1(x))^T$, with respect to $\boldsymbol{\alpha}$ over Θ.
2. Second stage: Generate $n - n_0$ independent stratified observations $\{x_i\}_{i=n_0+1}^{n}$ from $q_{\widehat{\boldsymbol{\alpha}}}(x) = \sum_{k=1}^p \widehat{\alpha}_k q_k(x)$. Estimate integral Z by $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ with all n observations, where

$$\widetilde{\boldsymbol{\alpha}} = \frac{n_0}{n}\boldsymbol{\gamma} + \frac{n - n_0}{n}\widehat{\boldsymbol{\alpha}} \tag{9}$$

and $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ can be either $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ or $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$.

Some rationale and implementation remarks are as follows:

(i) Criterion of selecting $\boldsymbol{\alpha}$: In the first stage, the optimal $\boldsymbol{\alpha}$ is estimated using the n_0 samples and it is desirable to select $\boldsymbol{\alpha}$ that gives the smallest asymptotic variance of the final estimator. Let $\mathrm{var}_{\boldsymbol{\alpha}}$ denotes the variance taken with respect to $q_{\boldsymbol{\alpha}}(x)$. We set the following conditions:

(C1) The union of supports of $q_k(x)$ contains the support of $\pi(x)$.
(C2) $\alpha_i > 0$ for $i = 1, \ldots, p$.
(C3) $\mathrm{var}_{\boldsymbol{\alpha}}[\pi(X)/q_{\boldsymbol{\alpha}}(X)] < \infty$ for some $\boldsymbol{\alpha} \in \Theta$.

Owen and Zhou (2000) and Tan (2004) showed that, under the above conditions, $\widehat{Z}_{\mathrm{Reg}}(\boldsymbol{\alpha})$ and $\widehat{Z}_{\mathrm{MLE}}(\boldsymbol{\alpha})$ are asymptotic normal and have the same asymptotic variance

$$\sigma_Z^2(\boldsymbol{\alpha}) = \mathrm{var}_{\boldsymbol{\alpha}}\left[\frac{\pi(X) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right] = \int \frac{\left(\pi(x) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(x)\right)^2}{q_{\boldsymbol{\alpha}}(x)}dx - Z^2, \tag{10}$$

where

$$\boldsymbol{\beta}_{\boldsymbol{\alpha}} = \mathrm{var}_{\boldsymbol{\alpha}}\left[\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right]^{-1} \mathrm{cov}_{\boldsymbol{\alpha}}^T\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right] = \left(\int \frac{\boldsymbol{g}(x)\boldsymbol{g}(x)^T}{q_{\boldsymbol{\alpha}}(x)}dx\right)^{-1}\left(\int \frac{\pi(x)\boldsymbol{g}(x)}{q_{\boldsymbol{\alpha}}(x)}dx\right).$$

Conditions (C1) to (C3) are satisfied when we have at least one proposal component dominating the tail of $\pi(x)$. With the sample $\{x_i\}_{i=1}^{n_0}$ from the pilot stage, $\sigma_Z^2(\boldsymbol{\alpha}) + Z^2$ is estimated by the IS estimator $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$ in (8) and the optimal $\boldsymbol{\alpha}$ is obtained by minimizing $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$.

(ii) Optimization range for $\boldsymbol{\alpha}$: The purpose of restricting $\boldsymbol{\alpha}$ in $[\delta, 1-\delta]^p$ for some small δ is to avoid unreliable estimators of $\sigma_Z^2(\boldsymbol{\alpha})$ or $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$. When $\alpha_i = 0$ for some i, $\int \pi(x)^2/q_{\boldsymbol{\alpha}}(x)dx$ can be infinite if q_i is the only proposal that dominates certain part of $\pi(x)$'s tail, or $\int \boldsymbol{g}(x)\boldsymbol{g}(x)^T/q_{\boldsymbol{\alpha}}(x)dx$ and $\int \pi(x)\boldsymbol{g}(x)/q_{\boldsymbol{\alpha}}(x)dx$ can be infinite if q_i is the only proposal that dominates some other proposals. In this case, if α_i is too close to 0, the estimator $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$ or $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ is unreliable. Experience shows that $\delta = 0.001$ is a reasonable choice.

(iii) Choice of the initial proportions $\boldsymbol{\gamma}$: $\boldsymbol{\gamma}$ is preferred to be close to the optimal proportion vector $\boldsymbol{\alpha}^*$. If there is no any prior knowledge about $\boldsymbol{\alpha}^*$, it is recommended to use $\boldsymbol{\gamma}$ with equal components in the first stage so that pilot sample is generated from each proposal equally.

(iv) $\widehat{Z}$ in second stage: Instead of using $n - n_0$ observations to construct the estimator $\widehat{Z}(\widehat{\boldsymbol{\alpha}})$, we use all n observations to construct the estimator $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ where the mixture proportions $\widetilde{\boldsymbol{\alpha}}$ account for the proportions of the combined sample.

3.2 Theoretical Properties

Let $\boldsymbol{\alpha}^*$ be the minimizer of $\sigma_Z^2(\boldsymbol{\alpha})$ under restriction $\boldsymbol{\alpha} \in \Theta$. We assume the following additional conditions:

(C4) $n_0 = o(n)$ and $n_0 \to \infty$ as $n \to \infty$.
(C5) $\pi(x)$ is not a linear combination of $q_1(x), \ldots, q_p(x)$.
(C6) $\boldsymbol{\alpha}^*$ is in the interior of Θ, that is, $\boldsymbol{\alpha}^* \in (\delta, 1-\delta)^p$.

Condition (C4) ensures $\widetilde{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}^*$. Condition (C5) is necessary since if $\pi(x)$ is a linear combination of $q_1(x), \ldots, q_p(x)$, $\sigma^2(\boldsymbol{\alpha})$ will be 0 for all α_1. Some discussions of condition (C6) are given in Remark 7.

3.2.1 *First-Order Properties*

Theorem 1. Under conditions (C1) to (C5), $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ are consistent and

$$\sqrt{n}\left(\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}}) - Z\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_Z^2(\boldsymbol{\alpha}^*)\right) \quad \text{and} \quad \sqrt{n}\left(\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}}) - Z\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_Z^2(\boldsymbol{\alpha}^*)\right).$$

Therefore, the two-stage procedure achieves the minimum asymptotic variance that Owen and Zhou's and Tan's estimators can achieve among all possible mixture proportions. Furthermore, since $\widehat{Z}_{\mathrm{Reg}}(\boldsymbol{\alpha})$ and $\widehat{Z}_{\mathrm{MLE}}(\boldsymbol{\alpha})$ are better than the stratified sampling estimator $\widehat{Z}_{\mathrm{SIS}}(\boldsymbol{\alpha})$, the two-stage procedure

outperforms all estimators introduced in Section 2 in asymptotic variance. The proof is given in Appendix A.

Remark 1. It is important to point out that $\sigma_Z^2(\boldsymbol{\alpha})$ and its estimator $\widehat{\sigma}^2(\boldsymbol{\alpha})$ are strictly convex by Lemma 1 in Appendix A. This guarantees a unique solution and applicability of convex optimization algorithms in the pilot stage. This property, or equivalently the strict convexity of the function $\sigma^2(\boldsymbol{\alpha}, \beta) = \mathrm{var}_{\boldsymbol{\alpha}}[(\pi(X) - \boldsymbol{\beta}^T \boldsymbol{g}(X))/q_{\boldsymbol{\alpha}}(X)]$, also ensures the consistency and asymptotic normality with convergence rate $\sqrt{n_0}$ of random proportion vector $\widehat{\boldsymbol{\alpha}}$ under mild conditions, by asymptotic theory for M-estimation with a convex criterion function (Haberman 1989). Therefore, larger n_0 gives more reliable $\widehat{\boldsymbol{\alpha}}$.

Remark 2. For $\widehat{Z}_{\mathrm{SIS}}(\boldsymbol{\alpha})$, the optimal mixture proportions are the ones that make the mixture proposal $q_{\boldsymbol{\alpha}}$ the closest to the target distribution π. Therefore, knowledge about the target density surface can help to find an approximate choice of $\boldsymbol{\alpha}$. However, the optimal mixture proportions $\boldsymbol{\alpha}^*$ for $\sigma_Z^2(\boldsymbol{\alpha})$ sometimes can be counterintuitive. For instance, in Example 1(B2) of Section 5, the target distribution is a mixture of a normal distribution and a t distribution, with mixing probability 0.8 and 0.2, respectively. When the same normal distribution is used as one of the proposal distributions, its optimal mixture proportion is only 0.1%. This is because, for $\widehat{Z}_{\mathrm{Reg}}(\boldsymbol{\alpha})$ and $\widehat{Z}_{\mathrm{MLE}}(\boldsymbol{\alpha})$, the numerator of $\sigma_Z^2(\boldsymbol{\alpha})$ involves $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$, a function of $\boldsymbol{\alpha}$, which complicates the determination of the optimal proportions. Hence, an automatic selection for mixture proportions becomes necessary for $\widehat{Z}_{\mathrm{Reg}}(\boldsymbol{\alpha})$ and $\widehat{Z}_{\mathrm{MLE}}(\boldsymbol{\alpha})$.

Remark 3. If $\widetilde{\boldsymbol{\alpha}}$ in (9) can be replaced by some other random proportion vector, as long as it is consistent to $\boldsymbol{\alpha}^*$ as $n \to \infty$, the same asymptotic results hold. For example, one can choose the mixture proportions of the second stage so that the combined sample (of both the pilot stage and second stage) is as close to the estimated optimal proportion vector $\widehat{\boldsymbol{\alpha}}$ as possible. For example, if $n_0\gamma_k < n\widehat{\alpha}_k$ for all $k = 1, \ldots, p$, one can use $(n\widehat{\boldsymbol{\alpha}} - n_0\boldsymbol{\gamma})/(n - n_0)$ in the second stage, which results in the combined sample having the exact estimated optimal proportion $\widehat{\boldsymbol{\alpha}}$. In this case, actually one should use n_0 as large as possible until it violates the above condition.

Remark 4. Similar asymptotic properties for $\widehat{Z}_{\mathrm{MIS}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{SIS}}(\widetilde{\boldsymbol{\alpha}})$ are presented in Lemma 3 in Appendix A. They are always inferior to the control variate based estimators and hence of less interest.

3.2.2 High-Order Properties. Theorem 1 shows that the selection of the pilot sample size n_0 does not affect the first-order property of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ as long as $n_0 = o(n)$ and $n_0 \to \infty$. Therefore, an optimal choice of n_0 needs to be determined by higher-order properties of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$. Consider the convergence rate of $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$, a weighted average of $\boldsymbol{\gamma} - \boldsymbol{\alpha}^*$ and $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ with weights n_0/n and $1 - n_0/n$. Since $\boldsymbol{\gamma} - \boldsymbol{\alpha}^*$ is biased, one would want to have a smaller n_0. However, a large n_0 makes $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ closer to 0, at the rate $O(1/\sqrt{n_0})$. Therefore, the optimal n_0 is chosen to balance the effects of these two rates. The following proposition gives the higher-order asymptotic expansions of $\widetilde{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widetilde{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$.

Proposition 1. Under conditions (C1)–(C6), $\widetilde{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widetilde{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ can be expanded as $\widehat{Z}^* + o(n_0/(n\sqrt{n})) + o(1/(n_0\sqrt{n}))$ and $\widehat{Z}^* = Z + g_1(\widetilde{\boldsymbol{\alpha}}) + g_2(\widetilde{\boldsymbol{\alpha}})$, where

$$g_1(\widetilde{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i) - \beta_{\boldsymbol{\alpha}^*} g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int \frac{\pi(x) - \beta_{\boldsymbol{\alpha}^*} g(x)}{q_{\boldsymbol{\alpha}^*}(x)} q_{\widetilde{\boldsymbol{\alpha}}}(x)dx,$$

and

$$g_2(\widetilde{\boldsymbol{\alpha}}) = O\left(\frac{n_0}{n\sqrt{n}}\right) + O\left(\frac{1}{n_0\sqrt{n}}\right).$$

The explicit forms of $g_2(\widetilde{\boldsymbol{\alpha}})$ are tedious and therefore presented in the Appendices. The selection of optimal n_0 is based on minimizing the MSE of $\widehat{Z}^*$, which is an approximation of the MSE of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$. Such an approximation of moments, as the criterion of second-order optimality, has been widely used in higher-order asymptotic theory, for example, Rothenberg (1984).

Theorem 2. Under conditions (C1)–(C6) and

(C7) $\int \pi(x)^4/q_{\boldsymbol{\alpha}}(x)^4 dx < \infty$ for some $\boldsymbol{\alpha} \in \Theta$,

it holds that

$$E[\widehat{Z}^* - Z] = O\left(\frac{1}{n}\right) \quad \text{and} \quad \mathrm{var}[\widehat{Z}^* - Z] = \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O\left(\frac{n_0}{n^2}\right) + O\left(\frac{1}{nn_0}\right).$$

Therefore, $\mathrm{MSE}[\widehat{Z}^*] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*) = O(\frac{n_0}{n^2}) + O(\frac{1}{nn_0})$.

The above result gives the approximate mean squared error with higher-order terms beyond the usual asymptotic variance $n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. The order can be attributed to three sources of variability. See the Appendices for details. One source of variability is due to using the pilot sample with mixture proportions $\boldsymbol{\gamma} \neq \boldsymbol{\alpha}^*$, which leads to terms of order $O(n_0/n^2)$. The second source is the variability of estimator $\widehat{\boldsymbol{\alpha}}$, which is of the order $O(1/(nn_0))$. The third source is the variability of estimating $\boldsymbol{\beta}_{\boldsymbol{\alpha}^*}$, which is the optimal coefficient of control variates, in $\sigma_Z^2(\boldsymbol{\alpha}^*)$. In $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$, the estimator of $\boldsymbol{\beta}_{\boldsymbol{\alpha}^*}$ is $\widehat{\boldsymbol{\beta}}_{\widetilde{\boldsymbol{\alpha}}}$. In $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$, a similar estimator is used, as can be seen from the proof of Theorem 1. This variability is of the order $O(1/(n\sqrt{n}))$, which is also $O(n_0/n^2) + O(1/(nn_0))$ because $2/\sqrt{n} \le n_0/n + 1/n_0$ by the inequality $2ab \le a^2 + b^2$.

Remark 5. By minimizing the order of difference, the optimal n_0 is $O(\sqrt{n})$ and hence $\mathrm{MSE}[\widehat{Z}^*] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$ is of order $O(1/(n\sqrt{n}))$. The asymptotic rate shows that how n_0 should change with the total sample size n. In practice, another consideration of selecting n_0 is the coverage of the target distribution with pilot samples. A poor coverage can lead to poorly estimated asymptotic variance and result in inaccurate $\widehat{\boldsymbol{\alpha}}$. Our experience shows one should choose n_0 at least $\sqrt{n}$ and possibly larger according to the complexity of problem and the quality of proposal distributions. On the other hand, one can assess $\widehat{\boldsymbol{\alpha}}$ by estimating its standard error after the pilot stage. If the standard error is larger than some criterion, such as 10% of $\widehat{\boldsymbol{\alpha}}$, one can add additional pilot samples. The standard error formula is given in Appendix B.

Remark 6. One essential fact leading to Theorem 2 is $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = O(1/\sqrt{n_0}) + O(n_0/n)$. Therefore, when $\widetilde{\boldsymbol{\alpha}}$ is replaced by

some other construction that is consistent but with different rate (e.g., Remark 3), the orders in Theorem 2 may change.

Remark 7. When some coordinates of $\boldsymbol{\alpha}^*$ are on the boundary of $[\delta, 1-\delta]$, the exact second-order property is complicated. However, it is still reasonable to use the same n_0 as indicated in Theorem 2. For example, when α_1^* is on the boundary, $\widehat{\boldsymbol{\alpha}}$, as an M-estimator, will converge to $\boldsymbol{\alpha}^*$ with a rate faster than or equal to $O(1/\sqrt{n_0})$ (Geyer 1994). In the proof of Theorem 2, when the convergence rate of $\widehat{\boldsymbol{\alpha}}_1$ changes from $O(1/\sqrt{n_0})$ to $O(1/n_0^{\varepsilon})$ with $\varepsilon \geq \frac{1}{2}$, the second order $O(n_0/n^2) + O(1/(n_0 n))$ changes to $O(n_0/n^2) + O(1/(n_0^{2\varepsilon} n))$. Then by choosing $n_0 = O(\sqrt{n})$, $\mathrm{MSE}[\widehat{Z}^*] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$ is still $O(1/(n\sqrt{n}))$ and the accuracy of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ remains the same.

4. EXTENSION TO RATIO ESTIMATORS

4.1 Extension of IS Techniques to Ratio Estimators

As mentioned in Section 1, the integral (2) can be estimated by the ratio estimator

$$\widehat{\mu}_{\mathrm{IS}} = \frac{\frac{1}{n}\sum_{i=1}^{n} h(x_i)\pi(x_i)/q(x_i)}{\frac{1}{n}\sum_{i=1}^{n} \pi(x_i)/q(x_i)}, \tag{11}$$

(Liu 2008; Rubinstein and Kroese 2008). By the delta method, it is easy to show that the asymptotic variance of $\widehat{\mu}_{\mathrm{IS}}$ is

$$\mathrm{var}_q\left(\frac{h(x)\pi(x) - \mu\pi(x)}{q(x)}\right). \tag{12}$$

In the sense of minimizing (12), the optimal choice of $q(x)$ is the probability density proportional to $|h(x)\pi(x) - \mu\pi(x)|$. Therefore, it is preferred to choose $q(x)$ that mimics the shape of $|h(x)\pi(x) - \mu\pi(x)|$. Similar to estimating the normalizing constant, multiple proposals may be needed and the techniques in Section 2 may be beneficial.

Given p proposal distributions $q_1(x), \ldots, q_p(x)$ and mixture proportions $\{\alpha_k\}_{k=1}^p$ satisfying $\sum_{k=1}^p \alpha_k = 1$. Observations $\{x_{k1}, \ldots, x_{kn_k}\}$ are generated from proposal q_k with size $n_k = \alpha_k n$ for each k. In Hesterberg (1995), the mixture IS and stratified sampling were applied to $\widehat{\mu}_{\mathrm{IS}}$ by using the mixture proposal $q_{\boldsymbol{\alpha}}$ in numerator and denominator separately as follows:

$$\widehat{\mu}_{\mathrm{SIS}} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} h(x_i)\pi(x_{ki})/q_{\boldsymbol{\alpha}}(x_{ki})}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \pi(x_{ki})/q_{\boldsymbol{\alpha}}(x_{ki})}.$$

Control variates and likelihood approach can also be applied to $\widehat{\mu}_{\mathrm{IS}}$. With the same control variates $\boldsymbol{g}(x)$ as in (8), μ can be estimated by the following:

$$\widehat{\mu}_{\mathrm{Reg}} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_1^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_2^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}},$$

$$\widehat{\mu}_{\mathrm{MLE}} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})}}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})}},$$

where

$$\widehat{\boldsymbol{\beta}}_1 = \widetilde{\mathrm{var}}\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1} \widetilde{\mathrm{cov}}^T\left(\frac{h(X)\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right) \quad \text{and}$$

$$\widehat{\boldsymbol{\beta}}_2 = \widetilde{\mathrm{var}}\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1} \widetilde{\mathrm{cov}}^T\left(\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)$$

$$\widetilde{\boldsymbol{\zeta}} = \underset{\zeta}{\mathrm{argmax}} \sum_{k=1}^{p}\sum_{i=1}^{n_k} \log\left[q_{\boldsymbol{\alpha}}(x_{ki}) + \zeta^T \boldsymbol{g}(x_{ki})\right].$$

Remark 8. The optimality of the above estimators can be seen by extending the optimality results of $\widehat{Z}_{\mathrm{Reg}}$ in Owen and Zhou (2000) and $\widehat{Z}_{\mathrm{MLE}}$ in Tan (2004) from scalar case to vector case. Specifically, under conditions (C1)–(C3) for $\pi(x)$ and $h(x)\pi(x)$, the two estimators

$$\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_1^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_2^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \end{pmatrix} \quad \text{and}$$

$$\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})} \end{pmatrix}$$

can be shown to be consistent and asymptotic normal with the minimum covariance matrix among all estimators in the form of

$$\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}_1^T \\ \boldsymbol{\beta}_2^T \end{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}$$

for arbitrary real vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Here $A \geq B$ means $A - B$ is nonnegative definite for two square matrices A and B. Then by the delta method, it is straightforward to show the optimality of $\widehat{\mu}_{\mathrm{Reg}}$ and $\widehat{\mu}_{\mathrm{MLE}}$. Their asymptotic variances are identical and equal to

$$\sigma_\mu^2(\boldsymbol{\alpha}) = \frac{1}{Z^2}\mathrm{var}_{\boldsymbol{\alpha}}\left(\frac{h(X)\pi(X) - \mu\pi(X) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right), \tag{13}$$

where

$$\boldsymbol{\beta}_{\boldsymbol{\alpha}} = \mathrm{var}\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1} \mathrm{cov}^T\left(\frac{h(X)\pi(X) - \mu\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right).$$

4.2 Two-Stage Procedure For Ratio Estimators

Take $\widehat{\mu}_{\mathrm{Reg}}$ and $\widehat{\mu}_{\mathrm{MLE}}$ as functions of $\boldsymbol{\alpha}$ and denote by $\widehat{\mu}_{\mathrm{Reg}}(\boldsymbol{\alpha})$ and $\widehat{\mu}_{\mathrm{MLE}}(\boldsymbol{\alpha})$. The two-stage procedure in Section 3 can be applied here:

1. First stage: Given initial proportion $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ satisfying $\sum_{k=1}^p \gamma_k = 1$, generate n_0 independent stratified

sample $\{x_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\gamma}}(x)$. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing

$$\widehat{\tau}^2(\boldsymbol{\alpha}) = \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\left[h(x_i)\pi(x_i) - \widehat{\mu}\pi(x_i) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}\mathbf{g}(x_i)\right]^2}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}, \quad (14)$$

where

$$\widehat{\mu} = \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{h(x_i)\pi(x_i)}{q_{\boldsymbol{\gamma}}(x_i)} \Big/ \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\pi(x_i)}{q_{\boldsymbol{\gamma}}(x_i)},$$

and

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\boldsymbol{g}(x_i)\boldsymbol{g}(x_i)^T}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right)^{-1} \times \left[\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{(h(x_i)\pi(x_i) - \widehat{\mu}\pi(x_i))\,\boldsymbol{g}(x_i)}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right],$$

with respect to $\boldsymbol{\alpha}$ over Θ.

2. Second stage: Generate $n - n_0$ independent stratified observations $\{x_i\}_{i=n_0+1}^{n}$ from $q_{\widehat{\boldsymbol{\alpha}}}(x)$. Estimate integral μ by $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ or $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ with all n observations, where $\widetilde{\boldsymbol{\alpha}} = n_0/n \cdot \boldsymbol{\gamma} + (n - n_0)/n \cdot \widehat{\boldsymbol{\alpha}}$.

In the first stage, $\widehat{\tau}^2(\boldsymbol{\alpha})$ is the Monte Carlo estimate of $Z^2\sigma_\mu^2(\boldsymbol{\alpha})$. Similar to the results in Section 3.3.1, $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ for μ have proper asymptotic results and the case for two proposal distributions is stated below.

Theorem 3. Under conditions (C1)–(C5) with $\pi(x)$ replaced by $h(x)\pi(x) - \mu\pi(x)$, $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ are consistent and

$$\sqrt{n}\left(\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}}) - \mu\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_\mu^2(\boldsymbol{\alpha}^*)\right)$$

and

$$\sqrt{n}\left(\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}}) - \mu\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_\mu^2(\boldsymbol{\alpha}^*)\right),$$

where $\boldsymbol{\alpha}^*$ is the minimizer of $\sigma_\mu^2(\boldsymbol{\alpha})$.

4.3 Consideration of Selecting Component Proposal Distributions

In this article, we focus on finding the optimal mixture weights to construct a mixture proposal distribution for IS, assuming that the set of component proposals to be included in the mixture has been preselected. Since the proposed mixture proportion determination automatically discriminates the high-quality proposals from the poor ones, our procedure in a way alleviates the difficulty of selecting the set of proposal distributions. It also allows a larger set of proposals to be considered as the procedure serves as a selection tool. Nevertheless, preselection of the proposals is extremely important as it provides the basis for efficient inference of optimal mixture weights. This is an area of active research. Here we provide some remarks and practical guidance.

Consider the asymptotic variances of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$, $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$, $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$, and $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ in (10) and (13). Owen and Zhou (2000) and Tan (2004) showed that when $\pi(x)$ is a linear combination of the component proposals, $\sigma_Z^2(\boldsymbol{\alpha}) = 0$ for any $\boldsymbol{\alpha}$. Therefore for estimating Z, it is preferred that the component proposals have a linear combination close to the shape of $\pi(x)$. This can be achieved by using proposals that separately approximate the modes and tails of $\pi(x)$. Alternatively, one can decompose $\pi(x)$ into a linear combination

$$\pi(x) = \sum_{k=1}^{r} c_k\pi_k(x). \quad (15)$$

Then the component proposals can be obtained by approximating each $\pi_k(x)$. Owen and Zhou (2000) gave some illustrations of this strategy.

For the ratio estimator, it can be shown similarly that when $h(x)\pi(x) - \mu\pi(x)$ is a linear combination of the component proposals, $\sigma_\mu^2(\boldsymbol{\alpha}) = 0$ for any $\boldsymbol{\alpha}$. Therefore, the strategy used for estimating Z can be used here as well. In particular, we can find a decomposition

$$h(x)\pi(x) - \mu\pi(x) = \sum_{k=1}^{r} c_k h(x)\pi_k(x) - \sum_{k=1}^{r} \mu c_k^*\pi_k^*(x)$$

and find component proposals to approximate the individual terms. If $h(x)$ takes negative values, additional terms corresponding to $h(x) = h^+(x) - h^-(x)$ will be needed. Example 3 in Section 5 provides an illustration of this approach.

Another consideration is the tail requirement. For estimating Z, $q_{\boldsymbol{\alpha}^*}(x)$ needs to have heavier tail than $\pi(x)$; and for estimating μ, $q_{\boldsymbol{\alpha}^*}(x)$ needs to have heavier tail than $h(x)\pi(x) - \mu\pi(x)$. In cases where $\pi(x)$'s tail decreases exponentially, the requirements can be satisfied by including some Student's t distributions or other heavy-tailed distributions in the set of component proposals (Geweke 1989).

Oh and Berger (1993) and West (1993) proposed adaptive procedures to find better proposal distributions. Liang, Liu, and Carroll (2007) proposed a stochastic approximation procedure to partition the sample domain and used truncations of the target distribution in the subregions as component proposal distributions. The normalizing constant of each component is estimated in a pilot stage. These procedures can be used here for finding the component proposals in our setting. In fact, the pilot stage of our proposed procedure can also be used as well. The estimated optimal mixture weights from the pilot stage may provide hints on potentially useful proposals to be considered. For example, a large weight for a component proposal that mainly covers the tail in one direction may suggest to use additional proposals to cover the more extreme part of the tail in that direction. However, caution should be exercised when considering the removal of a proposal distribution because of its small weight, since it may be used to serve as a defensive proposal that guarantees finite variance of the IS estimator.

5. EMPIRICAL STUDIES

Here, we present several examples to illustrate the performance of the proposed procedure. In all examples, the standard restricted optimization algorithm BFGS (Broyden-Hetcher-Goldfarb-Shanno) (Battiti and Masulli 1990) is used in the pilot stage to find $\widehat{\boldsymbol{\alpha}}$.

Example 1. Let $\phi(\boldsymbol{x};\sigma)$ be the normal density with mean 0 and standard error σ, and $\psi_k(\boldsymbol{x})$ be the density of t distribution with degree of freedom k. In this example, we consider two target distributions and two sets of proposal distributions. The combination is listed in Table 1. The case (A1) represents the

Table 1. Parameter settings of four cases in Example 1

	Proposal distributions			
	$q_1 = \prod_{i=1}^{10} \psi_k(x_i)$ and $q_2 = \prod_{i=1}^{10} \phi(x_i;\sigma)$			
Target distribution	$k=1$ $\sigma=1.1$	$k=1$ $\sigma=0.4$	$k=1$ $\sigma=1$	$k=2$ $\sigma=1$
$\prod_{i=1}^{10} \phi(x_i;1)$	(A1)	(A2)		
$0.2\prod_{i=1}^{10} \psi_4(x_i) + 0.8\prod_{i=1}^{10} \phi(x_i;1)$			(B1)	(B2)

situation that one of the proposal distribution, $q_2(\boldsymbol{x})$, is a good approximation to $\pi^*(\boldsymbol{x})$ by itself, and $q_1(\boldsymbol{x})$, being a product of Cauchy distributions, is a relatively poor proposal. We expect that the two-stage procedure will be helpful to decrease the contamination of $q_2(\boldsymbol{x})$. The case (A2) represents the situation that both proposals are not good approximation to the target and an appropriate proportion is not immediately clear. Both (B1) and (B2) represent the situation that one of the proposals, $q_2(\boldsymbol{x})$, is a good approximation to the center of the target, but with a lighter tail, and the other proposal, $q_1(\boldsymbol{x})$, has a heavier tail, for protection. The case (B1) uses a more conservative protection (Cauchy) and (B2) is more aggressive (t_2).

We compare five methods. The first three methods generate independent and stratified observations $\{\boldsymbol{x}_i\}_{i=1}^{n}$ from $q_{\boldsymbol{\alpha}_0}(\boldsymbol{x}) = \alpha_0 q_1(\boldsymbol{x}) + (1-\alpha_0)q_1(\boldsymbol{x})$, where $\boldsymbol{\alpha}_0 = (0.5, 1-0.5)$. The last two methods generate independent and stratified observations $\{\boldsymbol{x}_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\alpha}_0}(\boldsymbol{x})$ and $\{\boldsymbol{x}_i\}_{i=n_0+1}^{n}$ from $q_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{x})$, where $\widetilde{\alpha}_1 = \alpha_0 n_0/n + \widehat{\alpha}_1(n-n_0)/n$ and $\widehat{\alpha}_1$ is obtained by the corresponding method. Since the simulation results of regression method are nearly identical to the likelihood approach, we only list MLE and 2MLE here. Specifically, the methods are as follows. For simplicity, only formulas for estimating Z are listed.

UIS (Unprotected Importance Sampling): This is estimator (1) with $q(\boldsymbol{x}) = q_2(\boldsymbol{x})$.

SIS (Stratified Importance Sampling): This is estimator (3) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.

MLE (MLE method): This is estimator (6) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.

2SIS (Two-Stage Stratified Importance Sampling):

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i)}{q_{\widehat{\boldsymbol{\alpha}}}(x_i)}, \text{ where}$$

$$\widehat{\boldsymbol{\alpha}}_1 = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left(\alpha_1 \widetilde{\text{var}}_1 \left[\frac{\pi(x)}{q_{\boldsymbol{\alpha}_0}(x)} \right] + (1-\alpha_1) \widetilde{\text{var}}_2 \left[\frac{\pi(x)}{q_{\boldsymbol{\alpha}_0}(x)} \right] \right)$$

and $\widetilde{\text{var}}_k$ denotes the sample variance with the subset of $\{x_i\}_{i=1}^{n_0}$, which comes from $q_k(x)$. This is the method used in Raghavan and Cox (1998).

2MLE (Two-Stage MLE): This is our proposed method.

The results are shown in Table 2 for estimating Z and μ. Simulation is replicated for 1000 times independently with $n = 4000$ and $n_0 = 400$ in each simulation. We report the means of $\widehat{Z}$ or $\widehat{\mu}$, the means of $\widehat{\alpha}$ and the MSE

$$n\widehat{V} = \frac{n}{1000}\sum^{1000}(\widehat{Z}_i - Z)^2 \text{ or } \frac{n}{1000}\sum^{1000}(\widehat{\mu}_i - \mu)^2,$$

where Z and μ are theoretical values.

It is seen that, in (A1), where q_2 is a good proposal by itself, 2SIS and 2MLE choose α_1 close or equal to the smallest allowed value (0.001) for q_1, which minimizes its contamination, and achieves the same efficiency as UIS (using the good proposal only). They are more efficient than SIS and MLE, which use equal proportions for both proposal distribution.

In (A2), both 2SIS and 2MLE choose $\hat{\alpha}_1 = 0.98$, giving much higher proportion to the heavy-tail t proposal. It is seen that the normal proposal has a much lighter tail ($\sigma = 0.4$) than that the target ($\sigma = 1$). In this case, $q_1(x)$ is the better proposal. UIS, which uses $q_2(x)$ exclusively, does not have finite variance. Comparing to one stage MLE, the two-stage procedure reduces MSE by about 43% and 34% for estimating Z and μ, respectively.

In (B1) and (B2), UIS has the largest variance as expected. By using control variates, 2MLE and MLE perform much better than SIS and 2SIS. With the estimated mixture proportions, 2MLE reduces MSE by 10% and 30% for estimating Z in (B1) and (B2), respectively, comparing the one-stage MLE. Note that 2MLE obtains a larger estimated optimal proportion for $q_1(x)$ in (B2) than in (B1). Intuitively this is because $q_1(x)$ in (B2) is "closer" to the target integrand. In estimating μ, 2MLE and

Table 2. Comparison of methods for Example 1, with each column for one setting

		Z				μ			
	Method	(A1)	(A2)	(B1)	(B2)	(A1)	(A2)	(B1)	(B2)
$\overline{\widehat{\alpha}}_1$	UIS	0	0	0	0	0	0	0	0
	SIS	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	2SIS	0.001	0.98	0.21	0.13	0.001	0.93	0.40	0.37
	MLE	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	2MLE	0.004	0.98	0.72	0.999	0.001	0.91	0.42	0.30
nMSE	UIS	0.16	9.4×10^3	3.0	0.47	0.19	1.2×10^3	66	68
	SIS	0.45	28	0.15	0.16	0.34	3.2	0.38	0.27
	2SIS	0.15	16	0.087	0.028	0.20	2.1	0.37	0.26
	MLE	0.27	28	0.041	0.0094	0.34	3.2	0.37	0.16
	2MLE	0.15	16	0.037	0.0066	0.20	2.1	0.35	0.15

NOTE: $\overline{\widehat{\alpha}}_1$ is the mean of 1000 estimated mixture proportions and MSE is mean square error of integral estimators.

Table 3. Comparison between finite sample and asymptotic results

	Z				μ			
	(A1)	(A2)	(B1)	(B2)	(A1)	(A2)	(B1)	(B2)
$\widehat{\widehat{\alpha}}_1$	0.004	0.98	0.72	0.999	0.001	0.91	0.42	0.30
α_1^*	0.001	0.98	0.77	0.999	0.001	0.999	0.42	0.30
$n\widehat{V}$	0.150	15.5	0.037	0.0066	0.20	2.06	0.35	0.15
$\sigma^2(\boldsymbol{\alpha}^*)$	0.155	15.9	0.035	0.0061	0.19	1.97	0.36	0.15

NOTE: α_1^* is the mixture proportion giving the minimum asymptotic variance, $\widehat{V}$ is the sample variance of integral estimators, and $\sigma^2(\boldsymbol{\alpha}^*)$ is the minimum asymptotic variance.

MLE perform better than SIS and 2SIS, but the two-stage 2MLE and one-stage MLE are similar, since the estimated optimal proportions are close to 0.5.

To check the convergence properties of 2MLE, in all four cases we report, in Table 3, a comparisons between the theoretical minimum asymptotic variances and the sample variance of 2MLE, as well as a comparison between the optimal proportions and the average estimated proportions. It is seen that both of them are quite close to the optimal values.

Example 2. Consider a rare event problem in Hesterberg (1995). Let $\boldsymbol{X}$ be a three-dimensional random variable with independent components (X_1, X_2, X_3) and

$$\boldsymbol{X} = (X_1, X_2, X_3) = \max(0, \boldsymbol{Y}_1 + 10\boldsymbol{d} - \boldsymbol{Z}_1 - \boldsymbol{Z}_2 - \max(500, 3000 - \boldsymbol{Y}_2 - 40\boldsymbol{d})),$$

where $\boldsymbol{Y}_1 \sim N((1600, 1650, 1600), 100^2 I_3)$, $\boldsymbol{Y}_2 \sim N((1600, 1700, 1600), 100^2 I_3)$, $\boldsymbol{Z}_1 \sim \Gamma(100\mathbf{1}_3, (5, 6, 7))$ with Γ(scale,shape) denoting the gamma distribution, $\boldsymbol{Z}_2$ has density proportional to $e^{x/100} I_{x\in(0,300)}$, and $\boldsymbol{d} = \max(0,\ 60 - \boldsymbol{t})$, where $\boldsymbol{t} \sim N((54, 52, 55), 5^2 I_3)$. Denote the density of $\boldsymbol{X}$ to be $f(\boldsymbol{x}) = \prod_{j=1}^3 f_j(x_j)$. The targets of interest are

$$P = P\left[\sum_{i=1}^3 X_i > 1200\right]$$

and

$$\mu = E\left[80 \cdot \max\left(\sum_{i=1}^3 X_i - 1200, 0\right)\right].$$

The true value of P is about 0.003 and therefore the probability measures the area in the tail of $f(\boldsymbol{x})$. Hesterberg (1995) used $\widehat{Z}_{\text{SIS}}$ to estimate P and μ and constructed the proposal distributions by exponential tilting, using $q(\boldsymbol{x}) = c(\boldsymbol{\beta})\exp(\sum_{j=1}^3 \beta_j x_j) f(\boldsymbol{x})$ with parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. Seven proposals are constructed by setting $\boldsymbol{\beta} = c \cdot (I_1, I_2, I_3)$, where I_j is binary and c is set so that $E[\sum_{i=1}^3 X_i']$ is equal to some predetermined value, where (X_1', X_2', X_3') follows $q(\boldsymbol{x})$. Including $f(\boldsymbol{x})$ as another proposal component, there are eight proposal components. Hesterberg (1995) provided preset mixture proportions for these proposals, listed in Table 4.

Here, we compare the proposed two-stage procedure with the estimator used in Hesterberg (1995) for estimating both P and μ. The results are listed in Table 5. Again, simulation is replicated 1000 times independently with $n = 4000$ and $n_0 = 400$ in each simulation. We report the sample means and variances of $\widehat{P}$ and $\widehat{\mu}$, and the means of the mixture proportion $\widehat{\alpha}_i$ for $i = 1, \ldots, 8$.

Table 4. Parameters setting of the mixture proposal

Proposal	(I_1, I_2, I_3)	$E[\sum_{j=1}^3 X_i']$	α_i
$q_1(\boldsymbol{x}) = f(\boldsymbol{x})$			0.5
$q_2(\boldsymbol{x})$	(1, 0, 0)	1416	0.0035
$q_3(\boldsymbol{x})$	(0, 1, 0)	1266	0.028
$q_4(\boldsymbol{x})$	(0, 0, 1)	1616	0.0005
$q_5(\boldsymbol{x})$	(1, 1, 0)	1482	0.236
$q_6(\boldsymbol{x})$	(1, 0, 1)	1832	0.018
$q_7(\boldsymbol{x})$	(0, 1, 1)	1682	0.0635
$q_8(\boldsymbol{x})$	(1, 1, 1)	1898	0.151

NOTE: Each $q_i(\boldsymbol{x})$ is proportional to $\exp(\sum_{j=1}^3 \beta_j x_j) f(\boldsymbol{x})$, where $\boldsymbol{\beta} = c \cdot (I_1, I_2, I_3)$ and c is selected such that the expectation is equal to the corresponding expectation, for example, 1416. α_i is the mixture proportion for $q_i(\boldsymbol{x})$. Here (X_1', X_2', X_3') has density $q_i(\boldsymbol{x})$.

Comparing to SIS, it is seen that while the means are the same, 2MLE reduces the variance by 47% for estimating P and 44% for estimating μ. When comparing the proportion set selected by 2MLE and the predetermined proportion set used by SIS, it is seen that some of the proposal considered to be important for SIS is also determined important by 2MLE, such as $q_5(\boldsymbol{x})$ and $q_8(\boldsymbol{x})$. The major difference is that SIS puts too much proportion on $q_1(\boldsymbol{x})$, while 2MLE only selects a very small proportion for it, indicating that only a small proportion is needed for $q_1(\boldsymbol{x})$ to guarantee the bounded estimating variance.

Example 3. In this example, we examine the performance of 2MLE on estimating VaR using a Bayesian GARCH(1,1) model for S&P500 index series. Given a probability p and a time horizon d, VaR is the value that a portfolio would encounter a loss greater than or equal to, with probability p over the horizon.

Suppose at time T, we have historical log returns $\boldsymbol{y} = \{y_1, \ldots, y_T\}$. Let $R(\boldsymbol{y}_d) = \sum_{k=1}^d y_{T+k}$ be the cumulative return in the next d periods, where $\boldsymbol{y}_d = (y_{T+1}, \ldots, y_{T+d})$ and denote $F_{\boldsymbol{y}_d}$ as the cumulative distribution function (CDF) of R. Then the d days ahead VaR is defined as

$$\text{VaR}_p = \inf\{x \in \mathbb{R} | F_{\boldsymbol{y}_d}(x) \le p\}.$$

VaR is a widely used measure of market risk (Duffie and Pan 1997; Jorion 1997). To obtain the CDF $F_{\boldsymbol{y}_d}$, we model the return series using GARCH model (Engle 1982; Bollerslev 1986), a commonly used model for return series and modeling volatility dynamics. Specifically, we use a Bayesian GARCH(1,1) model with normal innovations (Geweke 1994; Bauwens and Lubrano 2008),

$$y_t = \varepsilon_t h_t^{1/2}, \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, 1), \quad h_t = \phi_0 + \phi_1 y_{t-1}^2 + \beta h_{t-1},$$

where $\phi_0 \ge 0$, $\phi_1 \ge 0$, and $\beta \ge 0$ and $\phi_1 + \beta < 1$ to ensure stationarity. Following Geweke (1994), the prior distributions of $\log\phi_0$ and (ϕ_1, β) are selected to be $N(a_0, \sigma_a^2)$ and $U(\phi_1 \ge 0, \beta \ge 0, \phi_1 + \beta < 1)$. Here ϕ_0 is transformed to have the real line as domain, and (ϕ_1, β) follows a uniform distribution in the stationary domain. The hyperparameters a_0 and σ_a^2 are set to be 1 and 2, respectively. We also use the sample variance for h_0^2 for simplicity.

The Bayesian approach has the advantage of taking into account of parameter estimation variability in the estimation of VaR. Due to the complexity, Monte Carlo method is used. Since

Table 5. Comparison between two methods of Example 2

	Method	Mean	Var	Mixture proportions							
				$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\alpha}_5$	$\widehat{\alpha}_6$	$\widehat{\alpha}_7$	$\widehat{\alpha}_8$
P	SIS	3.4×10^{-3}	3.0×10^{-8}	0.500	0.0035	0.028	0.0005	0.236	0.018	0.064	0.151
	2MLE	3.4×10^{-3}	1.6×10^{-8}	0.001	0.0040	0.038	0.0009	0.420	0.051	0.170	0.310
μ	SIS	41	1.8	0.500	0.0035	0.028	0.0005	0.236	0.018	0.064	0.151
	2MLE	41	1.0	0.001	0.002	0.021	0.0003	0.380	0.040	0.150	0.410

NOTES: SIS is the method of Hesterberg (1995) and 2MLE is our method. $\overline{\widehat{\mu}}$ and $\overline{\widehat{P}}$ are the means of 1000 point estimators, $\overline{\widehat{\alpha}}_i$ are the average mixture proportions, and $\widehat{V}$ is the sample variance of 1000 estimators.

VaR is largely a tail property, an appropriate implementation of IS may significantly improve the efficiency. Although VaR is not in the form of integral, it can be estimated easily by empirical quantiles from the Monte Carlo samples. Note that, CDF and probability are in the form of integral. Related literatures about estimating VaR using IS can be found in Hoogerheide and Van Dijk (2010), Glasserman, Heidelberger, and Shahabuddin (2000), and Dunkel and Weber (2007).

The two-stage algorithm is tested to estimate VaR with $p = 0.05$ and 0.01 and horizons 1, 2, and 5 days, corresponding to 4-, 5-, and 8-dimensional problem, as there are three parameters in the GARCH(1,1) model. Denote $\boldsymbol{\theta} = (\log\phi_0, \phi_1, \beta)$. For each VaR, following the strategy discussed in Section 4.3, we construct the proposal distributions based on the asymptotic variance of the empirical posterior CDF at VaR

$$\sigma_p^2(\boldsymbol{\alpha}) = \int \frac{\left[\left(1_{\{R(\boldsymbol{y}_d)\le \mathrm{VaR}\}}(\boldsymbol{y}_d) - p\right)\pi(\boldsymbol{y}_d,\boldsymbol{\theta}) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(\boldsymbol{y}_d,\boldsymbol{\theta})\right]^2}{q_{\boldsymbol{\alpha}}(\boldsymbol{y}_d,\boldsymbol{\theta})} \times d\boldsymbol{y}_d\, d\boldsymbol{\theta}, \tag{16}$$

where $\pi(\boldsymbol{y}_d,\boldsymbol{\theta}) = \prod_{k=1}^{T+d} p(y_k|y_{k-1},\boldsymbol{\theta})p(\boldsymbol{\theta})$, $p(y_k|y_{k-1},\boldsymbol{\theta})$ is the innovation density and $p(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$.

Expression (16) is not the variance of the VaR estimator. However, Hoogerheide and Van Dijk (2010) showed that the asymptotic variance of $\widehat{\mathrm{VaR}}_p$ can be approximated by $\sigma_p^2(\boldsymbol{\alpha})$ times a constant, which does not depend on the proposal density. Since it is difficult to sample from $\pi(\boldsymbol{y}_d,\boldsymbol{\theta})$ directly, we approximate it by the mixture of

$$q_1(\boldsymbol{y}_d,\boldsymbol{\theta}) = \prod_{k=1}^{T+d} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_N(\boldsymbol{\theta})$$

and

$$q_2(\boldsymbol{y}_d,\boldsymbol{\theta}) = q_*(y_{T+d}|y_{T+d-1},\boldsymbol{\theta}) \prod_{k=1}^{T+d-1} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_N(\boldsymbol{\theta}),$$

where $q_N(\boldsymbol{\theta})$ is the normal distribution with the mean vector being the MLE $\widehat{\boldsymbol{\theta}}$ and the covariance matrix Σ_N being the negative inverse Hessian matrix of $\pi(\boldsymbol{y},\boldsymbol{\theta})$ at $\widehat{\boldsymbol{\theta}}$, inflated by a constant to allow a wider coverage. We use $q_*(y_{T+d}|y_{1:T+d-1},\boldsymbol{\theta}) \sim N(-h_{T+d}^{1/2}, h_{T+d})$ for the proposal $q_2(\boldsymbol{y}_d,\boldsymbol{\theta})$. It tries to cover the tail (large loss on the last day of the horizon). Similar proposals can be constructed by considering other potential situations of large loss, but only this one is included in the current example.

With the approximation of $\pi(\boldsymbol{y}_d,\boldsymbol{\theta})$, the heavier tail components can be constructed by modifying the tails of q_1 and q_2. Then the following two proposals are included as the heavier tail components:

$$q_3(\boldsymbol{y}_d,\boldsymbol{\theta}) = \prod_{k=1}^{T+d} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_t(\boldsymbol{\theta})$$

and

$$q_4(\boldsymbol{y}_d,\boldsymbol{\theta}) = q_*(y_{T+d}|y_{1:T+d-1},\boldsymbol{\theta}) \prod_{k=1}^{T+d-1} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_t(\boldsymbol{\theta}),$$

where $q_t(\boldsymbol{\theta})$ is the product of three location-scale generalization of t_1 densities with the means being $\widehat{\boldsymbol{\theta}}$ and the squared scale parameters being the diagonal elements of Σ_N, and q_* is the same as in the construction of q_2. Since $\pi(\boldsymbol{y}_d,\boldsymbol{\theta})/q_3(\boldsymbol{y}_d,\boldsymbol{\theta}) = p(\boldsymbol{\theta})/q_t(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ is the prior distribution with exponentially decreasing tail, $q_3(\boldsymbol{y}_d,\boldsymbol{\theta})$ has heavier tail than $\pi(\boldsymbol{y}_d,\boldsymbol{\theta})$. Similarly, q_3 and q_4 have heavier tail than q_1, q_2, and proposals below, and therefore only mixture proportions for q_3 and q_4 need to be restricted.

To incorporate the integrand as discussed in Section 4.3, we further extend $q_1(\boldsymbol{y}_d,\boldsymbol{\theta})$ and $q_2(\boldsymbol{y}_d,\boldsymbol{\theta})$ to include

$$q_5(\boldsymbol{y}_d,\boldsymbol{\theta}) \propto 1_{\{y_{T+d}\le \mathrm{VaR}_{0.05} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d})q_1(\boldsymbol{y}_d,\boldsymbol{\theta})$$

and

$$q_6(\boldsymbol{y}_d,\boldsymbol{\theta}) \propto 1_{\{y_{T+d}\le \mathrm{VaR}_{0.05} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d})q_2(\boldsymbol{y}_d,\boldsymbol{\theta}),$$

Here the truncation is done only on y_{T+d}, instead of the more accurate but computationally expensive truncation of $\sum_{k=1}^{d} y_{T+k} \le \mathrm{VaR}_{0.05}$ under joint normal distribution.

The estimation of $\mathrm{VaR}_{0.01}$ can be done simultaneously by including the following component proposals:

$$q_7(\boldsymbol{y}_d,\boldsymbol{\theta}) \propto 1_{\{y_{T+d}\le \mathrm{VaR}_{0.01} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d})q_1(\boldsymbol{y}_d,\boldsymbol{\theta})$$

and

$$q_8(\boldsymbol{y}_d,\boldsymbol{\theta}) \propto 1_{\{y_{T+d}\le \mathrm{VaR}_{0.01} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d})q_2(\boldsymbol{y}_d,\boldsymbol{\theta}).$$

Overall, $q_1(\boldsymbol{y}_d,\boldsymbol{\theta})$ to $q_8(\boldsymbol{y}_d,\boldsymbol{\theta})$ are used as component proposal distributions.

Since our objective is to estimate $\mathrm{VaR}_{0.05}$ and $\mathrm{VaR}_{0.01}$ simultaneously, in the pilot stage we estimate the optimal mixture proportions by minimizing the sum of variances of the two estimators. Since $q_5, \ldots, q_8$ involve the unknown $\mathrm{VaR}_{0.05}$ and $\mathrm{VaR}_{0.01}$, the first stage sampling is modified as follows.

1. Generate pilot samples from q_1 to q_4 with sample size $n_0/8$ each.
2. Estimate $\mathrm{VaR}_{0.05}$ using the pilot samples from step 1. Replace $\mathrm{VaR}_{0.05}$ in q_5 and q_6 with the estimate and generate pilot samples from them, with sample size $n_0/8$ each.

Table 6. Comparison between MLE and 2MLE in Example 3

		$p = 0.05$		$p = 0.01$	
Horizon	Method	$\widehat{\text{VaR}}$	$\widehat{V}$	$\widehat{\text{VaR}}$	$\widehat{V}$
1 day	MLE	−1.332	14e-5	−1.894	20e-5
	2MLE	−1.333	3.8e-5	−1.895	4.6e-5
2 days	MLE	−1.886	5.1e-4	−2.773	12e-4
	2MLE	−1.886	1.5e-4	−2.771	3.5e-4
5 days	MLE	−2.997	17e-4	−4.432	5.9e-3
	2MLE	−2.996	5.4e-4	−4.424	1.8e-3

NOTE: $\widehat{\text{VaR}}$ is the average of 300 point estimators and $\widehat{V}$ is the sample variance of 300 estimators.

3. Estimate $\text{VaR}_{0.01}$ using the pilot samples from steps 1 and 2. Replace $\text{VaR}_{0.01}$ in q_7 and q_8 with the estimate and generate pilot samples from them, with sample size $n_0/8$ each.
4. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing $\hat{\tau}^2_{0.05}(\boldsymbol{\alpha}) + \hat{\tau}^2_{0.01}(\boldsymbol{\alpha})$, where $\hat{\tau}^2_p(\boldsymbol{\alpha})$ is the estimator for $\sigma^2_p(\boldsymbol{\alpha})$ using all samples in the first three steps.

Here, we compare the two-stage procedure 2MLE with the one-stage MLE with equal mixture proportions. The log returns of S&P500 index from September 28, 2010 to July 13, 2011 are used, with total 200 observations. The simulation is replicated for 300 times independently with $n = 4 \times 10^6$ and $n_0 = 8 \times 10^4$ in each simulation. δ is selected to be 0.001.

The summary of estimation results and the estimated mixture proportions $\widehat{\boldsymbol{\alpha}}$ are listed in Tables 6 and 7. From Table 6, it is seen that 2MLE's Monte Carlo variance is about 23%–32% of the variance of MLE, while there is almost no difference in the mean. Table 7 shows that the two-stage algorithm assigns most of the mixture proportions to q_3 and q_4, which indicates that these two heavy-tail component proposals are more important than the others. This is probably because q_1 and q_2 do not cover the high-density area of target distribution sufficiently, resulted in the preference to q_3 and q_4. Compared with MLE, the optimization in the pilot stage of 2MLE requires additional computing time, which is about 20% more in practice.

Table 7. Summary of mixture proportions estimated from stage 1

VaR	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\alpha}_5$	$\widehat{\alpha}_6$	$\widehat{\alpha}_7$	$\widehat{\alpha}_8$
1 day	1e-3	8e-4	3.4e-1	6.2e-1	9e-3	1.7e-2	8e-3	2e-3
2 days	1e-3	1e-3	3.5e-1	6.0e-1	2.4e-2	3e-3	9e-3	2e-3
5 days	1e-3	6e-4	4.3e-1	5.4e-1	2.0e-2	8e-4	4e-3	5e-4

NOTE: The average over 300 simulations are reported. $\widehat{\alpha}_1$ to $\widehat{\alpha}_8$ correspond to the mixture proportions assigned to q_1 to q_8.

Finally, we report some interesting insights on the comparison between MLE and 2MLE. By multiplying a scaling constant c^2 to the covariance matrix Σ_N used in the proposal q_1, all the related component proposals are made either more dispersed for $c > 1$ or more concentrated for $c < 1$. Since $\sigma^2_p(\boldsymbol{\alpha})$ is proportional to the estimation variance of VaR and the proportion does not depend on the proposal density, the trajectories of estimated $\sigma^2_p(\boldsymbol{\alpha}^*)$ and $\sigma^2_p(\boldsymbol{\alpha_0})$ as function of c are given in Figure 1 to illustrate how the quality of proposal distribution affect the performance of 2MLE and MLE.

It is seen that 2MLE is always better than MLE. Most interestingly, it shows that the performance of both methods depends on the quality of the proposal distributions, but 2MLE is much less sensitive to the proposal distributions and has more robust performance than MLE. This is due to 2MLE's ability to automatically adjust mixture proportion for the most efficient estimation. The simulation results (not shown here) show that, when c is small, 2MLE tends to assign most of the mixture proportions to the heavy tail q_3 and q_4. This insight reinforces the notion that the two-stage approach not only improves upon the one-stage approach, but also alleviates to some extent the difficulty of selecting proposal distributions.

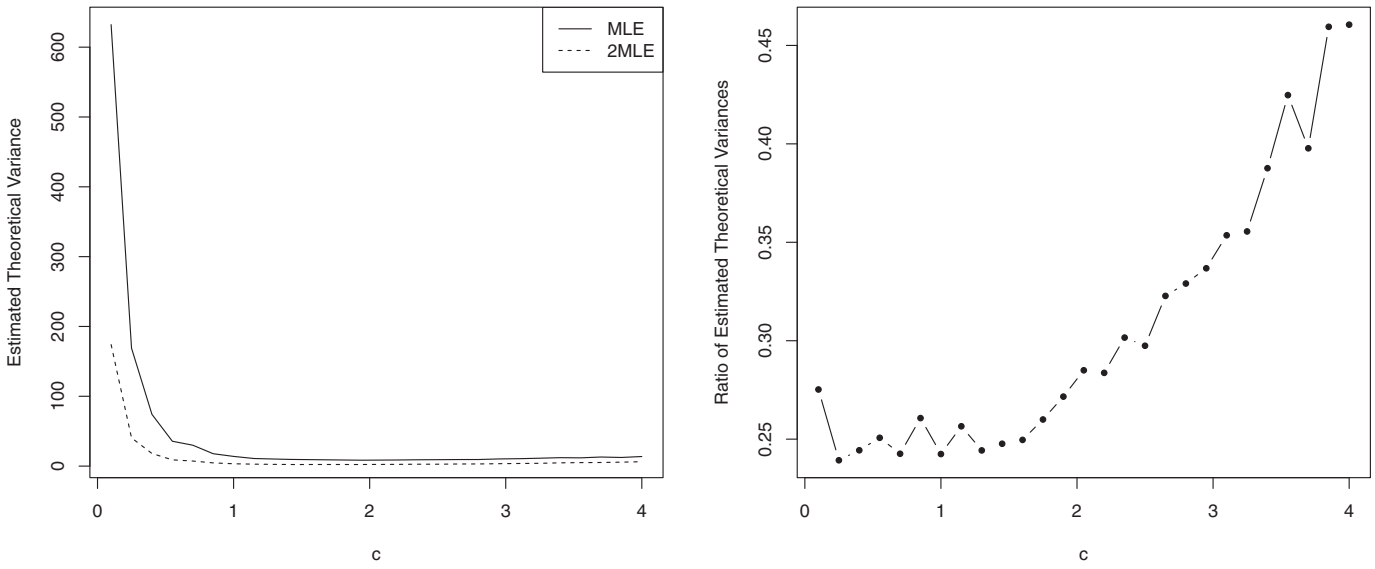

Figure 1. The left figure gives trajectories of estimated $\sigma^2_p(\boldsymbol{\alpha}^*)$ and $\sigma^2_p(\boldsymbol{\alpha_0})$, corresponding to MLE and 2MLE methods, respectively, with respect to the scaling constant c. c ranges from 0.1 to 4. For each c, the theoretical variances are estimated using one Monte Carlo sample, and the average over 10 replicates is reported. The right figure gives the trajectory of ratio of estimated $\sigma^2_p(\boldsymbol{\alpha}^*)$ over estimated $\sigma^2_p(\boldsymbol{\alpha_0})$.

6. SUMMARY

In this article, we proposed a two-stage procedure to select the optimal mixture proportions for the regression estimator in Owen and Zhou (2000) and MLE estimator in Tan (2004), and established the corresponding theoretical framework. The two-stage procedure significantly improved the existing methods in four aspects. First, the proposed estimator is asymptotically the best among all the estimators proposed in Owen and Zhou (2000), Tan (2004), and Raghavan and Cox (1998). Second, the criterion function of our pilot stage optimization is convex in its arguments, and therefore it is guaranteed that the optimization converges to the global minimum. Third, since there is no simple intuition in selecting the proportions for Owen and Zhou's (2000) regression estimator and Tan's (2004) MLE estimator, the proposed automatic procedure makes it much easier and safer to use mixture distributions for IS. Finally, the automatic determination of the mixture proportion alleviates the difficulty of choosing the set of proposal distributions to be considered in the mixture, as it serves as a selection and discrimination tool and hence allows users to include more potential proposal distributions for consideration.

APPENDIX A. PROOF OF RESULTS

For simplicity, we only consider two proposal distributions. Then $\boldsymbol{\alpha} = (\alpha_1, 1-\alpha_1)$ and $\boldsymbol{\gamma} = (\gamma_1, 1-\gamma_1)$. The proofs can be extended to the case of more than two proposals. To begin with, we establish the consistency of $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, 1-\widehat{\alpha}_1)$. Note that $\widehat{\alpha}_1$ is equivalently a component of the bivariate M-estimate $(\widehat{\alpha}_1, \widehat{\beta}) = \operatorname{argmin}_{\alpha,\beta} n_0^{-1} \sum_{i=1}^{n_0} m(x_i; \alpha, \beta)$, where $m(x; \alpha_1, \beta) = [\pi(x) - \beta g(x)]^2 / [q_{\boldsymbol{\alpha}}(x) q_{\boldsymbol{\gamma}}(x)]$. Let $M(\alpha_1, \beta) = \int m(x; \alpha_1, \beta) q_{\boldsymbol{\gamma}}(x) dx$ and $(\alpha_1^*, \beta^*) = \operatorname{argmin}_{\alpha_1,\beta} M(\alpha_1, \beta)$. Meanwhile, $M(\alpha_1, \beta)$ and $\sigma_Z^2(\boldsymbol{\alpha})$ are strictly convex functions.

Lemma 1. It holds that

$$(\widehat{\alpha}_1, \widehat{\beta}) \xrightarrow{P} (\alpha_1^*, \beta^*),$$
$$(\widehat{\alpha}_1, \widehat{\beta}) = (\alpha_1^*, \beta^*) - \frac{1}{2\sqrt{n_0}} V^{-1} \widehat{U} + o_p\left(\frac{1}{\sqrt{n_0}}\right),$$

where V and $\widehat{U}$ are given in Lemma 2. Then $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, 1-\widetilde{\alpha}_1) \xrightarrow{P} (\alpha_1^*, 1-\alpha_1^*)$. Meanwhile, $M(\alpha_1, \beta)$, $\sigma_Z^2(\boldsymbol{\alpha})$, and $\widehat{\sigma}^2(\boldsymbol{\alpha})$ are strictly convex functions.

Proof. Note that $m(x; \alpha, \beta)$ is convex since its Hessian matrix

$$D^2 m(x; \alpha_1, \beta) = \frac{2g(x)^2}{q_{\boldsymbol{\alpha}}(x) q_{\boldsymbol{\gamma}}(x)} \begin{pmatrix} \dfrac{(\pi(x) - \beta g(x))^2}{q_{\boldsymbol{\alpha}}(x)^2} & \dfrac{\pi(x) - \beta g(x)}{q_{\boldsymbol{\alpha}}(x)} \\ \dfrac{\pi(x) - \beta g(x)}{q_{\boldsymbol{\alpha}}(x)} & 1 \end{pmatrix}$$

is a positive semidefinite matrix. Then the consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ can be proved by verifying conditions 1–3 in Haberman (1989) for M-estimators by convex minimization. First, the parameter set $\Theta = [\delta, 1-\delta] \times \mathbb{R}$ of (α_1, β) is convex and closed. Second, (α_1^*, β^*) is unique. By Durrett (1996, Appendix 9), the differentiation and integration in $M(\alpha, \beta)$ can be exchanged so that

$$D^2 M(\alpha_1, \beta) = \begin{pmatrix} 2\displaystyle\int \frac{[\pi(x) - \beta g(x)]^2 g(x)^2}{q_{\boldsymbol{\alpha}}(x)^3} dx & 2\displaystyle\int \frac{[\pi(x) - \beta g(x)] g(x)^2}{q_{\boldsymbol{\alpha}}(x)^2} dx \\ 2\displaystyle\int \frac{[\pi(x) - \beta g(x)] g(x)^2}{q_{\boldsymbol{\alpha}}(x)^2} dx & 2\displaystyle\int \frac{g(x)^2}{q_{\boldsymbol{\alpha}}(x)} dx \end{pmatrix}.$$

For any bivariate vector v, $v^T \{D^2 M(\alpha_1, \beta)\} v \geq 0$ and the equality holds only when $\pi(x) \equiv c_1 q_1(x) + c_2 q_2(x)$ for some c_1 and c_2. By condition $(C5)$, $D^2 M(\alpha_1, \beta)$ is positive definite. Therefore, $M(\alpha_1, \beta)$ is strictly convex and (α_1^*, β^*) is unique. Third, let $W = (\delta, 1-\delta) \times \mathbb{R}$. By condition $(C3)$, $M(\alpha_1, \beta) < \infty$ for any $(\alpha_1, \beta) \in W$.

The expansion of $(\widehat{\alpha}_1, \widehat{\beta})$ can be found in the proof of Haberman (1989, Theorem 6.1) by verifying his conditions 7 and 10. First, $D^2 M(\alpha_1^*, \beta^*)$ is positive definite as mentioned above. Second, the gradient of $m(x; \alpha_1, \beta)$ satisfies $E|Dm(x; \alpha_1, \beta)|^2 < \infty$. Therefore, the convergence of $(\widetilde{\alpha}_1, 1-\widetilde{\alpha}_1)$ holds because $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\gamma} n_0/n + \widehat{\boldsymbol{\alpha}}(n-n_0)/n$ and $n_0 = o(n)$.

Finally, with the strict convexity of $M(\alpha_1, \beta)$, which is stated above, the strict convexity of $\sigma_Z^2(\boldsymbol{\alpha})$ can be seen by the facts that $\sigma_Z^2(\boldsymbol{\alpha}) = \min_{\beta} M(\alpha_1, \beta)$ and

$$\min_{\beta} M(\lambda\alpha_1 + (1-\lambda)\alpha_2, \beta) = \min_{\beta_1} \min_{\beta_2} M(\lambda\alpha_1 + (1-\lambda)\alpha_2, \lambda\beta_1 + (1-\lambda)\beta_2)$$

for any α_1, α_2 and $\lambda \in [0, 1]$. The strict convexity of $\widehat{\sigma}^2(\boldsymbol{\alpha})$ can be proved similarly. □

The following expansion of $(\widehat{\alpha}_1, \widehat{\beta})$ will be used in the higher-order calculation of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$.

Lemma 2. It holds that

$$(\widehat{\alpha}_1, \widehat{\beta}) = (\alpha_1^*, \beta^*) + \frac{1}{2\sqrt{n_0}} V^{-1} \widehat{U} - \frac{1}{2n_0} V^{-1} \times \left\{ (\widehat{V} - \sqrt{n_0} V) V^{-1} \widehat{U} + \widehat{W} \right\} + o_p\left(\frac{1}{n_0}\right),$$

where $\widehat{W}$ is a random variable of order $O_p(1)$,

$$\widehat{U} = \begin{pmatrix} \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{(\pi(x_i) - \beta^* g(x_i))^2 g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} \\ \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{2(\pi(x_i) - \beta^* g(x_i)) g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i) q_{\boldsymbol{\gamma}}(x_i)} \end{pmatrix},$$

$$V = \begin{pmatrix} \displaystyle\int \frac{[\pi(x) - \beta^* g(x)]^2 g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3} dx & \displaystyle\int \frac{[\pi(x) - \beta^* g(x)] g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^2} dx \\ \displaystyle\int \frac{[\pi(x) - \beta^* g(x)] g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^2} dx & \displaystyle\int \frac{g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)} dx \end{pmatrix},$$

and

$$\widehat{V} = \begin{pmatrix} \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)]^2 g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^3 q_{\boldsymbol{\gamma}}(x_i)} & \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)] g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} \\ \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)] g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} & \dfrac{1}{\sqrt{n_0}} \displaystyle\sum_{i=1}^{n_0} \frac{g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i) q_{\boldsymbol{\gamma}}(x_i)} \end{pmatrix}.$$

Proof. Note that $\widehat{U} = n_0^{-1/2} \sum_{i=1}^{n_0} Dm(x_i; \alpha_1^*, \beta^*)$, $\widehat{V} = n_0^{-1/2} \sum_{i=1}^{n_0} D^2 m(x_i; \alpha_1^*, \beta^*)$, and $V = \int D^2 m(x; \alpha_1^*, \beta^*) dx$. Then by Taylor expansion around (α_1^*, β^*) on $n_0^{-1} \sum_{i=1}^{n_0} Dm(x_i; \widehat{\alpha}_1, \widehat{\beta}) = 0$ and the convergence of $(\widehat{\alpha}_1, \widehat{\beta})$, we have

$$0 = -\frac{1}{\sqrt{n_0}} \widehat{U} + \frac{2}{\sqrt{n_0}} \widehat{V} \begin{pmatrix} \widehat{\alpha}_1 - \alpha_1^* \\ \widehat{\beta} - \beta^* \end{pmatrix} + \frac{1}{n_0} \widehat{W} + o_p\left(\frac{1}{n_0}\right),$$

then

$$\begin{pmatrix}\widehat{\alpha}_1-\alpha_1^*\\ \widehat{\beta}-\beta^*\end{pmatrix}=\frac{1}{2\sqrt{n_0}}V^{-1}\widehat{U}-\frac{1}{n_0}V^{-1}\left\{(\widehat{V}-\sqrt{n_0}V)\cdot\sqrt{n_0}\begin{pmatrix}\widehat{\alpha}_1-\alpha_1^*\\ \widehat{\beta}-\beta^*\end{pmatrix}+\widehat{W}\right\}+o_p\left(\frac{1}{n_0}\right).$$

The expansion of $(\widehat{\alpha}_1,\widehat{\beta})$ follows by substituting $(\widehat{\alpha}_1-\alpha_1^*,\widehat{\beta}-\beta^*)$ to the right-hand side (RHS) of the above equation. □

The combined sample $\{x_1,\ldots,x_n\}$ can be split into four parts by distributions q_1 or q_2 and first or second stages. Denote I_{jk} to be the index set of observations from the jth stage and q_k, that is, $I_{11}=\{1,\ldots,n_0\gamma_1\}$, $I_{12}=\{n_0\gamma_1+1,\ldots,n_0\}$, $I_{21}=\{n_0+1,\ldots,n_0+\lceil(n-n_0)\widehat{\alpha}_1\rceil\}$, $I_{22}=\{n_0+\lceil(n-n_0)\widehat{\alpha}_1\rceil+1,\ldots,n\}$, where $\lceil x\rceil$ means the largest integer smaller than x, and n_{jk} to be the size of I_{jk}. Here, we can use for the index, where $\lfloor x\rfloor$ is the largest integer smaller than x, to define I_{jk}. But for investigating the asymptotic behavior, the difference can be ignored. We will use the decomposition

$$\begin{aligned}&G_n\tau(x)\\&=\sqrt{\frac{n_0}{n}}\left\{\sum_{k=1}^{2}\sqrt{\gamma_k}\cdot\sqrt{n_{1k}}\left(\frac{1}{n_{1k}}\sum_{i\in I_{1k}}\tau(x_i)-\int\tau(x)q_k(x)dx\right)\right\}\\&\quad+\sqrt{\frac{n-n_0}{n}}\left\{\sum_{k=1}^{2}\sqrt{\widehat{\alpha}_k}\cdot\sqrt{n_{2k}}\left(\frac{1}{n_{2k}}\sum_{i\in I_{2k}}\tau(x_i)-\int\tau(x)q_k(x)dx\right)\right\}\\&\equiv\sqrt{\frac{n_0}{n}}\left\{\sqrt{\gamma_1}G_{11}\tau(x)+\sqrt{\gamma_2}G_{12}\tau(x)\right\}\\&\quad+\sqrt{\frac{n-n_0}{n}}\left\{\sqrt{\widehat{\alpha}_1}G_{21}\tau(x)+\sqrt{\widehat{\alpha}_2}G_{22}\tau(x)\right\}.\end{aligned}\tag{A.1}$$

The following lemma shows the convergence of $\widehat{Z}_{\mathrm{SIS}}$ with $\widetilde{\boldsymbol{\alpha}}$ as mixture proportion.

Lemma 3. For any integrable function $h(x)$ satisfying $\mathrm{var}_{\boldsymbol{\alpha}}[h(X)/q_{\boldsymbol{\alpha}}(X)]<\infty$ for every $\alpha_1\in[\delta,1-\delta]$, it holds that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}\xrightarrow{P}\int h(x)dx,$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}-\int h(x)dx\right)\xrightarrow{\mathcal{L}}N\left(0,\sum_{k=1}^{2}\alpha_k^*\mathrm{var}_k\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right).$$

where var_k denote the variance under distribution density $q_k(x)$.

Proof. We only need to prove asymptotic normality since it implies the consistency. Using decomposition (A.1), we have

$$\begin{aligned}&\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}-\int h(x)dx\right)\\&=\sqrt{\frac{n_0}{n}}\left\{\sqrt{\gamma_1}G_{11}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)+\sqrt{\gamma_2}G_{12}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)\right\}\\&\quad+\sqrt{\frac{n-n_0}{n}}\left\{\sqrt{\widehat{\alpha}_1}G_{21}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)+\sqrt{\widehat{\alpha}_2}G_{22}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)\right\}\end{aligned}\tag{A.2}$$

The asymptotic normality will be implied by showing that the first two terms in (A.2) are of order $o_p(1)$ and the remaining is asymptotic normal.

For the first two terms, we prove that the collection of functions $\{h(x)/q_{\boldsymbol{\alpha}}(x)\}_{\alpha_1\in[\delta,1-\delta]}$ is a Donsker class under either probability measures q_1 or q_2 by verifying the three conditions in van der Vaart (2000, Example 19.7). In fact, the parameter α is in a bounded set; $|h(x)/q_{\alpha_1}(x)-h(x)/q_{\alpha_2}(x)|\le|m(x,\alpha_1,\alpha_2)|\cdot|\alpha_1-\alpha_2|$ for every α_1,α_2, where $m(x,\alpha_1,\alpha_2)=h(x)g(x)/(q_{\alpha_1}(x)q_{\alpha_2}(x))$, and $\int|m(x,\alpha_1,\alpha_2)|^2q_k(x)dx<\infty$. By van der Vaart (2000, Lemma 19.24) and Lemma 1, we have $G_{1k}(h/q_{\widetilde{\boldsymbol{\alpha}}})=G_{1k}(h/q_{\boldsymbol{\alpha}^*})+o_p(1)$, $k=1,2$. Then by Central Limit Theorem and $n_0=o(n)$, the first two terms in (A.2) are of order $o_p(1)$.

For the last two terms, similarly, we argue that $G_{2k}(h/q_{\widetilde{\boldsymbol{\alpha}}})=G_{2k}(h/q_{\boldsymbol{\alpha}^*})+o_p(1)$ by a modification of van der Vaart (2000, Lemma 19.24) to handle random sample size. In fact, the key condition for his results, namely, weak convergence of $G_{2k}(h/q_{\boldsymbol{\alpha}^*})$, is guaranteed by van Der Vaart and Wellner (1996, Theorem 3.5.1). Then by the independence between $\hat{\alpha}_1$ and observations in $\{x_i\}_{i=n_0+1}^n$ and an extension of Chow and Teicher (2003, sec. 9.4), we have

$$\begin{pmatrix}G_{21}(h/q_{\widetilde{\boldsymbol{\alpha}}})\\ G_{22}(h/q_{\widetilde{\boldsymbol{\alpha}}})\end{pmatrix}\xrightarrow{\mathcal{L}}N\left(0,\begin{pmatrix}\mathrm{var}_1\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right] & 0\\ 0 & \mathrm{var}_2\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\end{pmatrix}\right)$$

and by Slutsky's theorem, $\sqrt{\widehat{\alpha}_1}G_{21}(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)})+\sqrt{\widehat{\alpha}_2}G_{22}(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)})\xrightarrow{\mathcal{L}}N(0,\sum_{k=1}^{2}\alpha_k^*\mathrm{var}_k[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}])$. Therefore the lemma holds. □

In the above proof, only the consistency of $\widetilde{\boldsymbol{\alpha}}$ is used. If $\widetilde{\boldsymbol{\alpha}}$ is replaced by other consistent mixture proportion, the convergence properties still hold.

Corollary 1. For any $\boldsymbol{\alpha}$ satisfying $\boldsymbol{\alpha}\xrightarrow{P}\boldsymbol{\alpha}^*$, it holds that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\boldsymbol{\alpha}}(x_i)}\xrightarrow{P}\int h(x)dx,$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\boldsymbol{\alpha}}(x_i)}-\int h(x)dx\right)\xrightarrow{\mathcal{L}}N\left(0,\sum_{k=1}^{2}\alpha_k^*\mathrm{var}_k\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right).$$

We also need the convergence of $\widetilde{\zeta}$ in $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$, where $\widetilde{\zeta}=\mathrm{argmin}_{\zeta}\sum_{i=1}^{n}\log[q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\boldsymbol{\zeta}^T\boldsymbol{g}(x_i)]$.

Lemma 4. The following convergence properties for $\widetilde{\zeta}$ hold:

$$\widetilde{\zeta}\xrightarrow{P}0\text{ and }\sqrt{n\widetilde{\zeta}}\xrightarrow{\mathcal{L}}N\left(0,\frac{\sum_{k=1}^{2}\alpha_k^*\mathrm{Var}_k\,[g(X)/q_{\boldsymbol{\alpha}^*}(X)]}{\left(\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx\right)^2}\right).$$

Proof. The random variable $\sqrt{n\widetilde{\zeta}}$ is the minimizer of convex function $\psi(s)=\sum_{i=1}^{n}\log(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+g(x_i)s/\sqrt{n})$. By verifying the condition of Hjort and Pollard (1994, basic corollary), we have the expansion

$$\sqrt{n\widetilde{\zeta}}=\frac{n^{-1/2}\sum_{i=1}^{n}g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}+o_p(1)$$

and then the convergence of $\widetilde{\zeta}$ follows by Lemma 3. By Taylor expansion around 0, we have

$$\begin{aligned}\psi(s)&=\sum_{i=1}^{n}\log\left(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)\right)+\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}\right\}s\\&\quad-\left\{\frac{1}{2n}\sum_{i=1}^{n}\frac{g(x_i)^2}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}\right\}s^2+R_n(s),\end{aligned}$$

where

$$R_n(s)=\left\{\frac{1}{3n\sqrt{n}}\sum_{i=1}^{n}\frac{g(x_i)^3}{q_{\xi}(x_i)^3}\right\}s^3$$

and ξ_1 is between $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_1+s/\sqrt{n}$.

For fixed s, $\boldsymbol{\xi}\xrightarrow{P}\boldsymbol{\alpha}^*$ by $\widetilde{\boldsymbol{\alpha}}\xrightarrow{P}\boldsymbol{\alpha}^*$. Then $R_n(s)\xrightarrow{P}0$ by Corollary 1, and condition of Hjort and Pollard (1994, basic corollary) holds. □

Proof of Theorem 1. By Taylor expansion of $n^{-1}\sum_{i=1}^n g(x_i)/(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\widehat{\zeta}g(x_i))=0$, $\widehat{\zeta}$ can be expanded as

$$\begin{aligned}\widehat{\zeta} &= \frac{\frac{1}{n}\sum_{i=1}^n g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}{\frac{1}{n}\sum_{i=1}^n g(x_i)^2/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2+\widehat{\zeta}\cdot\frac{1}{n}\sum_{i=1}^n g(x_i)^3/(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\dot{\zeta}g(x_i))^3}\\ &\equiv S_g/S_{gg},\ \text{where } \dot{\zeta} \text{ is between 0 and } \widetilde{\zeta}.\end{aligned}$$

By Taylor expansion, we have

$$\begin{aligned}\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}}) &= \frac{1}{n}\sum_{i=1}^n \frac{\pi(x_i)}{(\widetilde{\alpha}_1+\widetilde{\zeta})q_1(x_i)+(\widetilde{\alpha}_2-\widetilde{\zeta})q_2(x_i)}\\ &= S_\pi-\left(S_{gg}^{-1}S_{\pi g}-\beta^*\right)S_g, \qquad (A.3)\end{aligned}$$

where

$$S_\pi=\frac{1}{n}\sum_{i=1}^n\frac{\pi(x_i)-\beta^*g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)},\ S_{\pi g}=\frac{1}{n}\sum_{i=1}^n\frac{\pi(x_i)g(x_i)}{(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\ddot{\zeta}g(x_i))^2},$$

and $\ddot{\zeta}$ is between 0 and $\widetilde{\zeta}$.

Since $\widetilde{\boldsymbol{\alpha}}\xrightarrow{P}\boldsymbol{\alpha}^*, \widetilde{\zeta}\xrightarrow{P}0$, and $q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\dot{\zeta}g(x_i)=q_{\widetilde{\boldsymbol{\alpha}}+(\dot{\zeta},-\dot{\zeta})}(x_i)$, we have

$$S_\pi\xrightarrow{P}Z,\ \sqrt{n}(S_\pi-Z)\xrightarrow{\mathcal{L}}N\left(0,\mathrm{var}_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)-\beta^*g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right),$$

$$\sqrt{n}S_g\xrightarrow{\mathcal{L}}N\left(0,\mathrm{var}_{\boldsymbol{\alpha}^*}\left[\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right),$$

$$S_{\pi g}\xrightarrow{P}\mathrm{cov}_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}^*}(X)},\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]$$

and

$$S_{gg}\xrightarrow{P}\mathrm{var}_{\boldsymbol{\alpha}^*}\left[\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right].$$

by Lemma 3 and Corollary 1. Then plugging the above results in (A.3), Slutsky's theorem gives that

$$\begin{aligned}\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})&\xrightarrow{P}Z \text{ and } \sqrt{n}\left(\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})-Z\right)\\ &\xrightarrow{\mathcal{L}}N\left(0,\mathrm{var}_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)-\beta^*g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right).\end{aligned}$$

Similarly, the consistency and asymptotic normality of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ hold by the decomposition

$$\begin{aligned}\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}}) &= \frac{1}{n}\sum_{i=1}^n\frac{\pi(x_i)-\widehat{\beta}_{\widetilde{\boldsymbol{\alpha}}}g(x_i)}{q_{\boldsymbol{\alpha}}(x_i)}\\ &= S_\pi-\left[\widetilde{\mathrm{var}}\left(\frac{g(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)}\right)^{-1}\widetilde{\mathrm{cov}}\left(\frac{\pi(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)},\frac{g(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)}\right)-\beta^*\right]\cdot S_g.\end{aligned}$$

□

Proof of Proposition 1. Denote $G_n\tau(x)=\sqrt{n}[n^{-1}\sum_{i=1}^n\tau(x_i)-\int\tau(x)q_{\widetilde{\boldsymbol{\alpha}}}(x)dx]$. By (A.3) and Taylor expansion around α_1^*, we have

$$\begin{aligned}&\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})-Z\\ &=\frac{1}{\sqrt{n}}\left\{G_n\frac{\pi(x)-\beta^*g(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right\}+\frac{1}{\sqrt{n}}\left\{G_n\frac{g(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right\}(S_{gg}^{-1}S_{\pi g}-\beta^*)\\ &=\frac{1}{\sqrt{n}}\left\{G_n\frac{\pi(x)-\beta^*g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}+\frac{1}{\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}\\ &\quad\times(\widetilde{\alpha}_1-\alpha_1^*)\\ &\quad+\frac{1}{\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\right\}(\widetilde{\alpha}_1-\alpha_1^*)^2\\ &\quad+o\left(\frac{1}{\sqrt{n}}\left(\frac{n_0}{n}+\frac{1}{\sqrt{n_0}}\right)^2\right)\\ &\quad+\frac{1}{\sqrt{n}}\left\{G_n\frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}(\widetilde{\beta}-\beta^*)+o\left(\frac{1}{n}\right),\end{aligned}$$

where

$$\widetilde{\beta}=\frac{\frac{1}{n}\sum_{i=1}^n\pi(x_i)g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\left\{2-\frac{\frac{1}{n}\sum_{i=1}^n g(x_i)^2/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\right\}.$$

Note that $S_{gg}^{-1}S_{\pi g}-\beta^*=\widetilde{\beta}-\beta^*+o(1/\sqrt{n})$ by Taylor expansion. The expansion of $\widehat{\alpha}_1$ in Lemma 2 can be plugged into the above equation. After some algebra and note that $1/\sqrt{n}\le n_0/n+1/n_0$ by the inequality $2ab\le a^2+b^2$, we obtain the expansion of $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ as follows:

$$\begin{aligned}&Z+\frac{1}{\sqrt{n}}G_n\frac{\pi(x)-\beta^*g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\\ &\quad+\frac{n_0}{n\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1-\alpha_1^*)\\ &\quad+\frac{1}{\sqrt{n_0}\sqrt{n}}\cdot\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}A_{n_0}\\ &\quad+\frac{1}{n_0\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\cdot A_{n_0}^2\right.\\ &\quad\left.-G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\cdot B_{n_0}\right\}\\ &\quad+\frac{1}{n}\left\{G_n\frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\{\sqrt{n}(\widetilde{\beta}-\beta^*)\}+o\left(\frac{n_0}{n\sqrt{n}}\right)+o\left(\frac{1}{n_0\sqrt{n}}\right)\\ &\equiv Z+g_1(\widetilde{\boldsymbol{\alpha}})+g_2(\widetilde{\boldsymbol{\alpha}})+o\left(\frac{n_0}{n\sqrt{n}}\right)+o\left(\frac{1}{n_0\sqrt{n}}\right), \qquad (A.4)\end{aligned}$$

where

$$\begin{aligned}A_{n_0}&=(1,0)\cdot V^{-1}\widehat{U}/2,\\ B_{n_0}&=(1,0)\cdot V^{-1}\left((\widehat{V}-\sqrt{n_0}V)V^{-1}\widehat{U}/2+\widehat{W}\right),\\ g_1(\widetilde{\boldsymbol{\alpha}})&=\frac{1}{\sqrt{n}}G_n\frac{\pi(x)-\beta^*g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\end{aligned}$$

and

$$\begin{aligned}g_2(\widetilde{\boldsymbol{\alpha}})=&\frac{n_0}{n\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1-\alpha_1^*)\\ &+\frac{1}{\sqrt{n_0}\sqrt{n}}\cdot\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}A_{n_0}\\ &+\frac{1}{n_0\sqrt{n}}\left\{G_n\frac{(\pi(x)-\beta^*g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\cdot A_{n_0}^2\right.\\ &\left.-G_n\frac{(\pi(x)-\beta^*g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\cdot B_{n_0}\right\}\\ &+\frac{1}{n}\left\{G_n\frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\{\sqrt{n}(\widetilde{\beta}-\beta^*)\}.\end{aligned}$$

The expansion of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ follows similarly, except the definition of $\widetilde{\beta}$ is changed to

$$\widetilde{\beta}=\frac{\widetilde{\mathrm{cov}}\left[\pi(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X),g(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X)\right]}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\left\{2-\frac{\widetilde{\mathrm{var}}\left[g(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X)\right]}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\right\}.$$

□

Proof of Theorem 2. The calculation of moments of $\widehat{Z}^*$ involves calculating the moments of (A.4) including

$$E\left[\left(\frac{1}{n}\sum_{i=1}^n\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)}-\int\frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\widetilde{\boldsymbol{\alpha}}}(x)dx\right)^{k_1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{h_2(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)}-\int\frac{h_2(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\boldsymbol{\gamma}}(x)dx\right)^{k_2}(\widetilde{\beta}-\beta^*)^{k_3}\right]$$

for functions $h_1(x)$ and $h_2(x)$, $k_1=1,2$, $k_2=0,1,2$, and $k_3=0,1,2$,

From (A.4), note that the calculation of $E[\widehat{Z}^*-Z]$ involves calculating the cases of $k_1=1$. For $(k_1,k_2,k_3)=(1,1,0)$ and $(1,2,0)$, by

plugging in the decomposition

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int \frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)} q_{\widetilde{\boldsymbol{\alpha}}}(x)dx \\
&= \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int \frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)} q_{\boldsymbol{\gamma}}(x)dx\right) \\
&\quad + \frac{n-n_0}{n}\left(\frac{1}{n-n_0}\sum_{i=1}^{n-n_0}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int \frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)} q_{\widetilde{\boldsymbol{\alpha}}}(x)dx\right), \quad \text{(A.5)}
\end{aligned}$$

the expectations are of order $O(1/n)$ and $O(1/(n_0 n))$, respectively, by the law of iterated expectations conditioning on $\{x_i\}_{i=1}^{n_0}$. For $(k_1, k_2, k_3) = (1, 0, 1)$, the expectation is of order $O(1/n)$ by the inequality $2ab \le a^2 + b^2$ and the fact $E[\sqrt{n}(\widetilde{\beta} - \beta^*)]^2 < \infty$. For $(k_1, k_2, k_3) = (1, 0, 0)$, the expectation is 0. Therefore, $E[\widehat{Z}^* - Z] = O(1/n)$.

From (A.4), note that the calculation of $\mathrm{var}[\widehat{Z}^* - Z]$ involves calculating the cases of $k_1 = 2$. For $(k_1, k_2, k_3) = (2, 0, 0)$, $(2, 1, 0)$, and $(2, 2, 0)$, the expectations are of order $O(1/n)$, $O(n_0/n^2)$, and $O(1/(n_0 n))$, respectively, by (A.5) and the law of total variance conditioning on $\{x_i\}_{i=1}^{n_0}$. For$(k_1, k_2, k_3) = (2, 0, 1)$ and $(2, 0, 2)$, the expectations are of order $O(1/(n\sqrt{n}))$ and $O(1/n^2)$, respectively, by the inequality $2ab \le a^2 + b^2$ and the fact $E[\sqrt{n}(\widetilde{\beta} - \beta^*)]^4 < \infty$. The other terms are dominated by $O(n_0/n^2) + O(1/(n_0 n))$. Therefore by noting that $1/\sqrt{n} \le n_0/n + 1/n_0$,

$$\mathrm{var}\left[\widehat{Z}^* - Z\right] = \frac{1}{n}\mathrm{var}\left[G_n \frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right] + O\left(\frac{1}{nn_0}\right) + O\left(\frac{n_0}{n^2}\right).$$

Again by (A.5) and the law of total variance conditioning on $\{x_i\}_{i=1}^{n_0}$, some algebra gives that

$$\begin{aligned}
&\frac{1}{n}\mathrm{var}\left[G_n \frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right] \\
&= \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + \frac{n_0}{n^2}\left\{\mathrm{var}_{\boldsymbol{\gamma}}\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - \sigma_Z^2(\boldsymbol{\alpha}^*)\right\} \\
&\quad + \frac{1}{n}\left(1 - \frac{n_0}{n}\right)\left\{\mathrm{var}_1\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - \mathrm{var}_2\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right)\right\} \\
&\quad \times E(\widehat{\alpha}_1 - \alpha_1^*) \quad \text{(A.6)} \\
&\approx \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O\left(\frac{n_0}{n^2}\right) + O\left(\frac{1}{nn_0}\right).
\end{aligned}$$

Therefore, $\mathrm{var}[\widehat{Z}^* - Z] = \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O(n_0/n^2) + O(1/(nn_0))$. □

The proofs of Proposition 1 and Theorem 2 reveal the sources of the higher orders $O(n_0/n^2)$ and $O(1/(nn_0))$ in $\mathrm{MSE}[\widehat{Z}^*] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. These two orders come from three sources, which can be seen by investigating each term in (A.4) and (A.6). One source is due to using pilot samples, which leads to terms

$$\frac{n_0}{n\sqrt{n}}\left\{G_n \frac{(\pi(x) - \beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1 - \alpha_1^*) \text{ in (A.4)}$$

and

$$\frac{n_0}{n^2}\left\{\mathrm{var}_{\boldsymbol{\gamma}}\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - \sigma_Z^2(\boldsymbol{\alpha}^*)\right\} \text{ in (A.6)},$$

and results in the order $O(n_0/n^2)$. When $\boldsymbol{\gamma} = \boldsymbol{\alpha}^*$, these two terms are equal to 0 and thus they are derived from the difference between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}^*$. Another one is the variability of random coefficient of control variates, which leads to the term

$$\frac{1}{n}\left\{G_n \frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\left\{\sqrt{n}(\widetilde{\beta} - \beta^*)\right\} \text{ in (A.4)}.$$

This variability results in the order $O(1/(n\sqrt{n}))$, which is also $O(n_0/n^2) + O(1/(nn_0))$ when $n_0 = \sqrt{n}$, because $2/\sqrt{n} \le n_0/n + 1/n_0$. The other source is the variability of estimated mixture proportion $\widetilde{\boldsymbol{\alpha}}$, which leads to all other terms in (A.4) and (A.6) except the previous three terms and $n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. This variability results in the order $O(1/(nn_0))$.

For the asymptotic properties of $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$, the proof differs in two aspects with that of $\widehat{Z}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$. One is the M-estimator $\widehat{\boldsymbol{\alpha}}$ contains an estimated parameter $\widehat{\mu}$ in the criterion function. The other one is the ratio form of $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$.

Lemma 5. $\widehat{\alpha}_1 \xrightarrow{P} \alpha_1^*$ as $n \to \infty$ for $\widehat{\alpha}_1$ defined in (14).

Proof. $\widehat{\alpha}_1$ can be equivalently obtained as a component of the bivariate estimator $(\widehat{\alpha}_1, \widehat{\beta}) = \mathrm{argmin}_{\alpha_1, \beta \in \Theta}\ n_0^{-1}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \widehat{\mu})$, where $\rho(x; \alpha_1, \beta, \mu) = \frac{[h(x)\pi(x) - \mu\pi(x) - \beta g(x)]^2}{q_{\boldsymbol{\alpha}}(x)q_{\boldsymbol{\gamma}}(x)}$ and $\Theta = [\delta, 1 - \delta] \times \mathbb{R}$. The proof of consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ contains two steps.

First, although the domain of β is unbounded, $\widehat{\beta}$ stays in a compact set almost surely when $n \to \infty$, because

$$|\widehat{\beta}| = \left|\frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{(h(x_i)\pi(x_i) - \widehat{\mu}\pi(x_i))g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}}{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{g(x_i)^2}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}}\right| \le \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{|h(x_i)\pi(x_i)|}{q_{\boldsymbol{\gamma}}(x_i)} + \frac{|\widehat{\mu}\pi(x_i)|}{q_{\boldsymbol{\gamma}}(x_i)}\right)\frac{2}{\delta}}{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{g(x_i)^2}{(q_1(x_i) + q_2(x_i))q_{\boldsymbol{\gamma}}(x_i)}}$$

and the RHS converges to a constant almost surely since $\widehat{\mu} \to \mu$ almost surely. Then the consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ and the minimizer $(\widehat{\alpha}_1', \widehat{\beta}')$ restricted in some compact set $C \subset \Theta$, that is, $\mathrm{argmin}_{\alpha_1, \beta \in C} n_0^{-1}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \widehat{\mu})$, are equivalent since $P((\widehat{\alpha}_1, \widehat{\beta}) \in C) \to 1$.

Second, the consistency of $(\widehat{\alpha}_1', \widehat{\beta}')$ and the minimizer with $\widehat{\mu}$ replaced by μ, that is, $\mathrm{argmin}_{\alpha_1, \beta \in C}\ n_0^{-1}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \mu)$, are equivalent because

$$\begin{aligned}
&\sup_{\alpha_1, \beta \in C}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \widehat{\mu}) - \frac{1}{n_0}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \mu)\right| \\
&\quad \le (\widehat{\mu}^2 - \mu^2)\max_{\alpha_1, \beta \in C}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\pi(x_i)^2}{q_{\alpha_1}(x_i)q_{\boldsymbol{\gamma}}(x_i)} \\
&\qquad + (\widehat{\mu} - \mu)\max_{\alpha_1, \beta \in C}\frac{2}{n_0}\sum_{i=1}^{n_0}\frac{(h(x_i)\pi(x_i) - \beta g(x_i))\pi(x_i)}{q_{\alpha_1}(x_i)q_{\boldsymbol{\gamma}}(x_i)} \\
&\quad \to 0 \text{ almost surely},
\end{aligned}$$

and the argument similar to van der Vaart (2000, Theorem 5.7). Then since the consistency of $\mathrm{argmin}_{\alpha_1, \beta \in C}\ n_0^{-1}\sum_{i=1}^{n_0}\rho(x; \alpha_1, \beta, \mu)$ holds by replacing $\pi(x)$ in Lemma 1 by $h(x)\pi(x) - \mu\pi(x)$, the consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ follows. □

Proof of Theorem 3. The consistency and asymptotic normality of $\widehat{\mu}_{\mathrm{MLE}}(\widetilde{\boldsymbol{\alpha}})$ follow the extension of proof of Theorem 1 to random vector

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} h(x_i)\pi(x_i)/q_{\widetilde{\boldsymbol{\alpha}}+\widetilde{\zeta}} \\ \pi(x_i)/q_{\widetilde{\boldsymbol{\alpha}}+\widetilde{\zeta}}\end{pmatrix} - \begin{pmatrix}\int h(x)\pi(x)/q_{\boldsymbol{\alpha}^*}(x)dx \\ \int \pi(x)/q_{\boldsymbol{\alpha}^*}(x)dx\end{pmatrix}\right\}$$

and the delta method. The proof for $\widehat{\mu}_{\mathrm{Reg}}(\widetilde{\boldsymbol{\alpha}})$ is similar. □

APPENDIX B. VARIANCE MATRICES FOR $\widehat{\boldsymbol{\alpha}}$

Denote

$$\begin{aligned}
I_{jkl} &= \int \frac{(\pi(x) - \boldsymbol{\beta}^{*T}\boldsymbol{g}(x))^j \boldsymbol{g}(x)\boldsymbol{g}(x)^T}{q_{\boldsymbol{\alpha}^*}(x)^k q_{\boldsymbol{\gamma}}(x)^l}dx, \\
A &= I_{230} - I_{120}I_{010}^{-1}I_{120}, \quad B = I_{010} - I_{120}I_{230}^{-1}I_{120}, \\
C &= I_{441} - 2I_{331}I_{010}^{-1}I_{120}, \text{ and } D = I_{331} - 2I_{221}I_{010}^{-1}I_{120}.
\end{aligned}$$

When estimating Z, $(\widehat{\alpha}_2, \ldots \widehat{\alpha}_p)$ has the asymptotic variance matrix

$$\frac{1}{\sqrt{n_0}}\left(A^{-1}CA^{-1} - 2I_{230}^{-1}I_{120}B^{-1}DA^{-1}\right).$$

When estimating μ, similar expression can be obtained by replacing $\pi(x)$ in I_{jkl} by $(h(x) - \mu)\pi(x)$.

[Received March 2012. Revised June 2013.]

REFERENCES

Battiti, R., and Masulli, F. (1990), "BFGS Optimization for Faster and Automated Supervised Learning," in *Proceedings of the International Neural Network Conference (INNC 90)-Paris-France*, pp. 757–760. [1356]

Bauwens, L., and Lubrano, M. (2008), "Bayesian Inference on GARCH Models Using the Gibbs Sampler," *The Econometrics Journal*, 1, 23–46. [1358]

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327. [1358]

Chow, Y., and Teicher, H. (2003), *Probability Theory: Independence, Interchangeability, Martingales*, New York: Springer Verlag. [1362]

Cochran, W. (1977), *Sampling Techniques*, New York: Wiley. [1352]

Denny, M. (2001), "Introduction to Importance Sampling in Rare-Event Simulations," *European Journal of Physics*, 22, 403. [1350]

Duffie, D., and Pan, J. (1997), "An Overview of Value at Risk," *The Journal of Derivatives*, 4, 7–49. [1358]

Dunkel, J., and Weber, S. (2007), "Efficient Monte Carlo Methods for Convex Risk Measures in Portfolio Credit Risk Models," in *Proceedings of the 2007 Winter Simulation Conference*, pp. 958–966. [1359]

Durrett, R. (1996), *Probability: Theory and Examples*, Pacific Grove, CA: Duxbury Press. [1361]

Emond, M., Raftery, A., and Steele, R. (2001), "Easy Computation of Bayes Factors and Normalizing Constants for Mixture Models via Mixture Importance Sampling," Technical Report No. 398, Department of Statistics, Washington University Seattle. [1351]

Engle, R. (1982), "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica: Journal of the Econometric Society*, 50, 987–1007. [1358]

Fan, S., Chenney, S., Hu, B., Tsui, K., and Lai, Y. (2006), "Optimizing Control Variate Estimators for Rendering," in *Computer Graphics Forum* (Vol. 25), eds. E. Gröller and L. Szirmay-Kalos, Goslar, Germany: Eurographics Association, pp. 351–357. [1351]

Ford, E., and Gregory, P. (2007), "Bayesian Model Selection and Extrasolar Planet Detection," in *Statistical Challenges in Modern Astronomy IV* (Vol. 371), eds. G. J. Babu and E. D. Feigelson, San Francisco, CA: Astron. Soc. Pacific, p. 189. [1350]

Gelman, A., and Meng, X. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185. [1350]

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica: Journal of the Econometric Society*, 57, 1317–1339. [1356]

——— (1994), "Bayesian Comparison of Econometric Models," Working Paper No. 532, Federal Reserve Bank of Minneapolis. [1358]

Geyer, C. (1994), "On the Asymptotics of Constrained M-Estimation," *The Annals of Statistics*, 22, 1993–2010. [1355]

Giordani, P., and Kohn, R. (2010), "Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixtures of Normals," *Journal of Computational and Graphical Statistics*, 19, 243–259. [1350]

Givens, G., and Raftery, A. (1996), "Local Adaptive Importance Sampling for Multivariate Densities With Strong Nonlinear Relationships," *Journal of the American Statistical Association*, 91, 132–141. [1350]

Glasserman, P., Heidelberger, P., and Shahabuddin, P. (2000), "Variance Reduction Techniques for Estimating Value-at-Risk," *Management Science*, 46, 1349–1364. [1359]

Haberman, S. (1989), "Concavity and Estimation," *The Annals of Statistics*, 17, 1631–1661. [1354,1361]

Hesterberg, T. (1988), "Advances in Importance Sampling," Ph.D. dissertation, Department of Statistics, Stanford University. [1352]

——— (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194. [1350,1351,1355,1358]

Hjort, N., and Pollard, D. (1994), "Asymptotics for Minimisers of Convex Processes," Statistical Research Report, Department of Mathematics, University of Oslo. [1362]

Hoogerheide, L., and Van Dijk, H. (2010), "Bayesian Forecasting of Value at Risk and Expected Shortfall Using Adaptive Importance Sampling," *International Journal of Forecasting*, 26, 231–247. [1359]

Jorion, P. (1997), *Value At Risk: The New Benchmark for Controlling Market Risk*, Chicago: McGraw-Hill. [1358]

Kong, A., McCullagh, P., Meng, X., Nicolae, D., and Tan, Z. (2003), "A Theory of Statistical Models for Monte Carlo Integration," *Journal of the Royal Statistical Society,* Series B, 65, 585–604. [1352]

Liang, F., Liu, C., and Carroll, R. J. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320. [1351,1356]

Liu, J. (2008), *Monte Carlo Strategies in Scientific Computing*, New York: Springer Verlag. [1355]

Oh, M., and Berger, J. (1993), "Integration of Multimodal Functions by Monte Carlo Importance Sampling," *Journal of the American Statistical Association*, 88, 450–456. [1350,1351,1356]

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical Association*, 95, 135–143. [1351,1352,1353,1355,1356,1361]

——— (1999), "Adaptive Importance Sampling by Mixtures of Products of Beta Distributions," Technical Report No. 1999-1, Department of Statistics, Stanford University. [1350]

Raghavan, N., and Cox, D. (1998), "Adaptive Mixture Importance Sampling," *Journal of Statistical Computation and Simulation*, 60, 237–260. [1351,1352,1357,1361]

Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer Verlag. [1351]

Rothenberg, T. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics* (Vol. 2), eds. Z. Griliches and M. D. Intriligator, Amsterdam: North Holland, pp. 881–935. [1354]

Rubinstein, R., and Kroese, D. (2008), *Simulation and the Monte Carlo Method*, Hoboken, NJ: Wiley. [1351,1355]

Smith, P., Shafi, M., and Gao, H. (1997), "Quick Simulation: A Review of Importance Sampling Techniques in Communications Systems," *IEEE Journal on Selected Areas in Communications*, 15, 597–613. [1350]

Tan, Z. (2004), "On a Likelihood Approach for Monte Carlo Integration," *Journal of the American Statistical Association*, 99, 1027–1036. [1351,1352,1353,1355,1356,1361]

van der Vaart, A. (2000), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [1362,1364]

van der Vaart, A., and Wellner, J. (1996), *Weak Convergence and Empirical Processes*, New York: Springer Verlag. [1362]

Veach, E., and Guibas, L. (1995), "Optimally Combining Sampling Techniques for Monte Carlo Rendering," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 419–428. [1350,1351,1352]

West, M. (1993), "Approximating Posterior Distributions by Mixture," *Journal of the Royal Statistical Society,* Series B, 55, 409–422. [1350,1356]