Moment Matching Adaptive Importance Sampling with Skew-Student Proposals

Shijia Wang and Tim B. Swartz *

Abstract

This paper considers integral approximation via importance sampling where the importance sampler is chosen from a family of skew-Student distributions. This is an alternative class of distributions than is typically considered in importance sampling applications. We describe variate generation and propose adaptive methods for fitting a member of the skew-Student family to a particular integral. We also demonstrate the utility of the approach in several examples.

Keywords : adaptive algorithms, importance sampling, simulation

*Wang is an Assistant Professor in the School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, No 94 Weijin Road, Tianjin 300071, China. Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby British Columbia, Canada V5A1S6. Swartz has been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

1 INTRODUCTION

The evaluation of integrals is a fundamental problem that presents itself in many diverse fields such as mathematical finance, economics and physics. For statisticians, integrals are commonplace in the Bayesian framework and arise as posterior expectations. In many applications, particularly in high dimensions, the integrals in question are intractable. Therefore, one must resort to methods of integral approximation. Evans and Swartz (2000) describe the major approaches used in the approximation of integrals with a particular emphasis on integrals arising in statistics.

One of the long-standing approaches to integral approximation is importance sampling which dates back to at least Metropolis and Ulam (1949). Importance sampling proceeds by rewriting the integral of interest

$$I(f) = \int_S f(y)\ dy \tag{1}$$

as

$$I(f) = \int_S \left(\frac{f(y)}{q(y)}\right) q(y)\ dy$$

where the density function $q(y)$, with support S, is introduced and is referred to as an importance sampler, assuming $f(y) > 0$ implies $q(y) > 0$. In importance sampling, independent variates $y^{(1)}, \ldots, y^{(N)}$ are generated from the distribution corresponding to $q(y)$, and we obtain the importance sampling estimator

$$\hat{I}(f) = \frac{1}{N}\sum_{i=1}^{N} \frac{f(y^{(i)})}{q(y^{(i)})}. \tag{2}$$

We note that the estimator $\hat{I}(f)$ is unbiased and is a consistent estimator where

$$Var(\hat{I}(f)) = \frac{1}{N}\left[\int_S \frac{f^2(y)}{q(y)}\ dy - I^2(f)\right] \tag{3}$$

is finite if the integral in (3) is finite. We require a proposal distribution $q(y)$ that leads to finite variance $Var(\hat{I}(f))$. From (3), we see that the variance of $\hat{I}(f)$ is small when $f(y) \approx kq(y)$ for some constant k. Therefore our goal is to choose an importance sampler $q(y)$ which permits convenient variate generation and whose shape mimics the shape of $f(y)$.

In practice, standard importance sampling may not be directly applicable to the common problem where $f(y)$ can only be evaluated up to a normalizing constant. For example, in a Bayesian inference context, we are interested in estimating the posterior expectation of some test function $h(y)$, where $f(y)$ in Equation (1) is proportional to the product of the test function $h(y)$ multiplied by the unnormalized posterior density $g(y)$. Normalized importance sampling bypasses this problem. In normalized importance sampling, independent variates $y^{(1)}, \ldots, y^{(N)}$ are generated from the distribution corresponding to $q(y)$, and we approximate the integral of interest as

$$\hat{I}(f) = \sum_{i=1}^{N} \frac{h(y^{(i)})g(y^{(i)})}{q(y^{(i)})} \Big/ \sum_{i=1}^{N} \frac{g(y^{(i)})}{q(y^{(i)})}. \tag{4}$$

Often in Bayesian applications, there may be various test functions $h(y)$ of interest, and often these are simple functions. In these cases, the choice of importance sampler is often based on mimicing the unnormalized posterior $g(y)$ rather than $h(y)g(y)$. We note that the estimator $\hat{I}(f)$ in (4) induced by normalized importance sampling is biased and that the estimation of $Var(\hat{I}(f))$ may be challenging.

As a general purpose integration technique, it appears that importance sampling has fallen out of favour compared to some of the popular Markov chain methods such as the Metropolis-Hastings algorithm (see Gilks, Richardson and Spiegelhalter 1996). However, in principle there is nothing wrong with importance sampling. In fact, importance sampling has several advantages over Markov chain methods. For example, error assessment

of averages in a Markov chain is not straightforward due to the dependence structure in the chain. Also, Markov chain methods require the determination of "convergence to stationarity" of the chain. Furthermore, even if a practitioner wishes to use Markov chain methods, it is comforting to have an alternative technique which provides corroborating evidence of the accuracy of the approximations.

There is a growing body of work on importance sampling approaches in the past decade. Indeed, a powerful feature of the importance sampling framework is the ability to estimate the normalizing constant of the target distribution. Elvira and Martino (2021) provide a comprehensive review for importance sampling approaches. A mismatch between proposal distribution and target may lead to a large variance in the importance sampling (IS) estimator (Agapiou et al. 2017). One line of research involves a nonlinear transformation for the IS weights (Ionides 2008). Vehtari et al. (2015) propose a generalized Pareto distribution for the distribution of importance weights, to model the heavy tail caused by the transformation. Multiple importance sampling (MIS) algorithms (Elvira et al. 2019) are another class of methods which decrease the variance of the IS estimator. In MIS algorithms, the samples are simulated from multiple proposals, instead of a single one. He et al. (2014) develop control variates in an IS framework to reduce the variance of the MIS estimator. Sbert et al. (2018) explore the use of a linear combination of distributions as a proposal. Another line of work is based on adaptive importance sampling (AIS) approaches. Adaptive importance sampling algorithms iteratively adapt the parameters of the importance sampler to achieve accurate approximation of the target distribution (Bugallo et al. 2017). Bugallo et al. (2017) provide a thorough review for AIS algorithms. Adaptive schemes can be classified into three categories according to Bugallo et al. (2017): resampling, moment matching and independent adaptive processes. Cornuet et al. (2012) propose a multiple AIS scheme that adapts proposal parameters using all samples up to

the latest iteration. Martino et al. (2017) develop an AIS algorithm with a hierarchical procedure for generating samples. Elvira et al. (2017) propose novel PMC schemes that can be combined with AIS algorithms.

Often, multivariate normal or the longer tailed multivariate Student families are used as proposal distributions (Evans and Swartz 2000, Cornuet et al. 2012, Elvira et al. 2019). A drawback with these families is that they are symmetric and may not be effective when the integrand $f(y)$ in (1) is skewed. We hope that a richer family of importance samplers (as we are proposing) will reduce the mismatch between the importance sampler and the target distribution. In this paper, we propose the use of restricted skew-Student distributions for importance sampling on $\mathcal{R}^n$, based on a moment matching adaptive scheme. A comprehensive treatment of the properties and applications of skew-elliptical distributions (with particular emphasis on skew-normal distributions) appears in the volume edited by Genton (2004). The skew-Student family extends the range of integrals for which importance sampling is successful. Azzalini and Capitanio (2014) provides a comprehensive review for skew family distributions. In Section 2, we describe the restricted skew-normal family of distributions and provide the relevant details for the implementation of adaptive importance sampling. For example, we describe variate generation and propose adaptive methods for fitting a member of the restricted skew-normal family to a particular integral. Adaptive importance sampling is then extended to the family of restricted skew-Student distributions in Section 3. In Section 4, we demonstrate the utility of the approach with some examples. We conclude with a short discussion in Section 5.

2 RESTRICTED SKEW-NORMAL DISTRIBUTIONS

2.1 Standard Skew-Normal Distributions

There are a number of variations in the definition of skew-elliptical distributions in the literature (Genton 2004). We consider the standard multivariate skew-normal distribution as defined by Azzalini and Dalla Valle (1996).

Definition: A random vector $z = (z_1, \ldots, z_n)'$ is n-dimensional standard skew-normal, denoted $z \sim SSN_n(\Omega, \alpha)$ if it has probability density function

$$g(z) = 2\phi_n(z, \Omega)\Phi(\alpha' z)$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)'$ is the skew parameter, Φ is the standard univariate normal cumulative distribution function and $\phi_n(z, \Omega)$ is the probability density function of the n-dimensional normal distribution with mean vector 0 and correlation matrix $\Omega = (\omega_{ij})$.

There are many properties that can be established regarding the SSN family. For example, we immediately note that if the skew parameter α is equal to the zero vector, then z reduces to a normal vector with $g(z) = \phi_n(z, \Omega)$. In addition, it can be shown via moment generating functions that the marginal distributions $z_i \sim SSN_1(1, \lambda_i)$ where $\lambda_i = (\sum \alpha_j \omega_{ji})/[1 + \alpha'\Omega\alpha - (\sum \alpha_j \omega_{ji})^2]^{1/2}$, $i = 1, \ldots, n$.

2.2 Restricted Skew-Normal Distributions

Next, we simultaneously extend and restrict the SSN family. Following Dalla Valle (2004), the standard skew-normal family SSN is extended according to $y \sim SN_n(\epsilon, S, \Omega, \alpha)$ where $y = \epsilon + Sz$, $z \sim SSN_n(\Omega, \alpha)$ and S is a diagonal matrix described by the vector $s = (s_1, \ldots, s_n)'$. The diagonal entries s_i affect variability in the coordinates and the

vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is a location parameter. We next restrict the SN_n family by imposing $\Omega = I_n$ where I_n is the $n \times n$ identity matrix. The restriction reduces the number of parameters in the SN_n family. Figure 1 displays the relationships between the SNN family, SN family and RSN family. Later, we see that the restriction leads to an effective fitting procedure where estimated moments are equated to theoretical moments. Even though the restriction reduces the number of parameters, the resultant family provides probability distributions that are flexible and can differ markedly from the multivariate normal family.

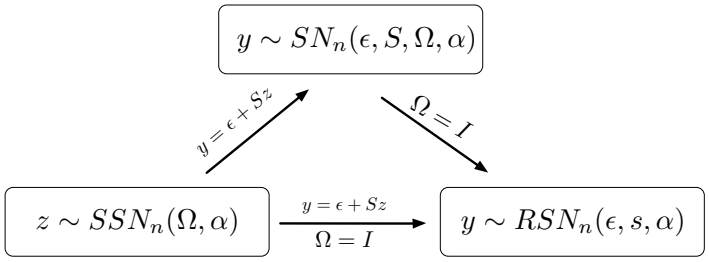

Figure 1: Relationships between the SSN_n, SN_n and RSN_n families.

In Figure 2, we provide some contour plots of bivariate densities of the restricted SN family which we denote $RSN_n(\epsilon, s, \alpha)$. We observe that this stretching (skewing) provides different shapes than the elliptical contours of multivariate normal distributions. Note that we cannot determine the skew according to a single component of α (Azzalini and Capitanio, 2014). Whereas the multivariate normal family is characterized by $n(n+3)/2$ parameters, the RSN_n family consists of $3n$ parameters (i.e. ϵ, s and α) and is more parsimonious when $n \geq 4$.

By the change of variables formula, the probability density function for $y \sim RSN(\epsilon, s, \alpha)$

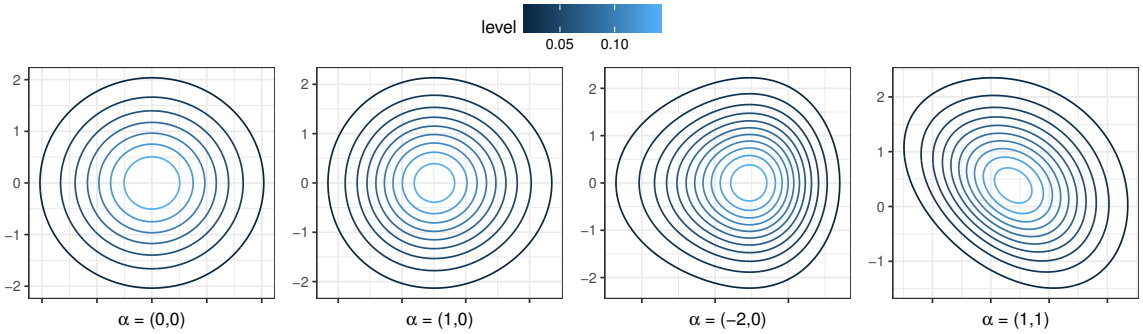

Figure 2: Contour plots of (a) $RSN_2(0, 0, (0, 0)')$ (i.e. the bivariate standard normal), (b) $RSN_2(0, 0, (1, 0)')$, (c) $RSN_2(0, 0, (-2, 0)')$ and (d) $RSN_2(0, 0, (1, 1)')$.

is given by

$$\begin{aligned} q(y) &= \frac{2}{(2\pi)^{n/2} \mid S \mid} \exp\left\{-\frac{1}{2}(y-\epsilon)'S^{-2}(y-\epsilon)\right\} \Phi\left(\alpha' S^{-1}(y-\epsilon)\right) \\ &= \frac{2}{(2\pi)^{n/2} \prod_{i=1}^{n} s_i} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\epsilon_i}{s_i}\right)^2\right\} \Phi\left(\sum_{i=1}^{n} \alpha_i (y_i-\epsilon_i)/s_i\right) . \end{aligned} \quad (5)$$

An important consideration for the implementation of importance sampling is that $q(y)$ in (5) is easily evaluated at any point y. This is not the case for all of the various skew-elliptical distributions which have been proposed in the literature.

2.3 Generating restricted skew-normal random numbers

The next relevant issue for importance sampling is the generation of random variates y having the pdf (5) given the parameters ϵ, s and α. To generate, we use the characterization in Proposition 6 of Azzalini and Dalla Valle (1996) and follow the steps:

- set $\delta = (1 + \alpha'\alpha)^{-1/2} \alpha$

- obtain the Cholesky factor A of $\Sigma = \begin{pmatrix} 1 & \delta' \\ \delta & I_n \end{pmatrix}$ (i.e. obtain the unique lower triangular matrix A with positive diagonal entries such that $AA' = \Sigma$)
- generate a sample $t_1, \ldots, t_{n+1}$ of Normal$(0, 1)$ variates
- set $\begin{pmatrix} u \\ v \end{pmatrix} = A \begin{pmatrix} t_1 \\ . \\ . \\ . \\ t_{n+1} \end{pmatrix}$ where u is a scalar
- set $y = \begin{cases} \epsilon + Sv & \text{if} \quad u > 0 \\ \epsilon - Sv & \text{if} \quad u < 0 \end{cases}$

2.4 An adaptive importance sampling with restricted skew-normal proposals

The final and most challenging issue relevant to the implementation of restricted skew-normal importance sampling is the determination of a member of the restricted skew-normal family corresponding to a particular integral. Again, we seek a restricted skew-normal density $q(y)$ which mimics the integrand $f(y)$ in (1). We consider an adaptive procedure where we fit the parameters ϵ, s and α in stages. We begin by assuming that $f(y)$ in (1) is non-negative as in Bayesian applications where $f(y)$ is proportional to the product of $h(y)$ and $g(y)$. We set $\alpha^{(0)} = 0$ which reduces the restricted skew-normal importance sampler to a multivariate normal importance sampler with mean ϵ and covariance matrix S^2. A standard approach (Evans and Swartz 2000) is to set $\epsilon = \epsilon^{(0)}$

and $S = S^{(0)}$, and using a Laplacian (normal) approximation of $f(y)$ (where f is assumed twice differentiable), we let

- $\epsilon^{(0)}$ be the solution of $\left[\frac{\partial \log f(y)}{\partial y}\right] = 0$ (i.e. $\epsilon^{(0)}$ maximizes $f(y)$)
- $S^{(0)}$ be the diagonal matrix whose diagonal entries are the same as the diagonal entries of the Cholesky factor of the matrix $\left(\frac{-\partial^2 \log f(y)}{\partial y_i \partial y_j}\right)^{-1}_{y=\epsilon^{(0)}}$

At this point, we have defined a skew-normal importance sampler $q^{(0)}(y)$ based on $\epsilon^{(0)}$, $S^{(0)}$ and $\alpha^{(0)}$. However, the shape of $q^{(0)}(y)$ may not mimic the shape of the integrand $f(y)$ very well. To improve the fit, we consider an adaptive sampling approach where we first sample $y^{(1)}, \ldots, y^{(N)}$ from $q^{(0)}(y)$ and calculate the ratio

$$\hat{R}(m) = \frac{\hat{I}(mf)}{\hat{I}(f)} = \frac{\frac{1}{N}\sum_{i=1}^{N} m(y^{(i)}) f(y^{(i)})/q^{(0)}(y^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} f(y^{(i)})/q^{(0)}(y^{(i)})} \tag{6}$$

for various functions $m(y)$. The $3n$ functions $m(y)$ are chosen so that $\hat{R}(m)$ corresponds to estimates of the mean, the second central moments and the third central moments of the distribution with density proportional to $f(y)$. In other words, we obtain the n-dimensional vector $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)'$ where $\hat{\mu}_j$ is calculated according to (6) by setting $m(y) = y_j$. We also obtain the n-dimensional vector $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_n)'$ where $\hat{\gamma}_j$ is calculated by setting $m(y) = (y_j - \hat{\mu}_j)^2$. We also obtain the n-dimensional vector $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_n)'$ where $\hat{\tau}_j$ is calculated by setting $m(y) = (y_j - \hat{\mu}_j)^3$.

In the first stage of adaptation, we update ϵ, s and α by referring to the expected values for skew-normal distributions as derived in Azzalini and Dalla Valle (1996) and in Genton, He and Liu (2001). We match sample moments with theoretical moments

$$\hat{\mu}_i = \epsilon_i + s_i \left(\frac{2}{\pi}\right)^{1/2} \delta_i \tag{7}$$

$$\hat{\gamma}_i = s_i^2 \left(1 - \frac{2}{\pi}\delta_i^2\right) \tag{8}$$

and

$$\hat{\tau}_i = \frac{\sqrt{2}(4-\pi)}{\pi^{3/2}}\ s_i^3\ \delta_i^3 \tag{9}$$

where $\delta_i = (1+\alpha'\alpha)^{-1/2}\,\alpha_i$ and $i = 1,\ldots,n$. The right-hand sides of (7)-(9) represent the mean value, the variance, and the third central moment of the restricted skew-normal proposals. Using (7), we substitute $\delta_i = (\pi/2)^{1/2}(\hat{\mu}_i - \epsilon_i)/s_i$ into (9) which yields

$$\epsilon_i^{(1)} = \hat{\mu}_i - \left(\frac{2\hat{\tau}_i}{4-\pi}\right)^{1/3} \tag{10}$$

for $i = 1,\ldots,n$. Therefore, (10) is used to provide the updated parameter $\epsilon^{(1)}$ for adaptive importance sampling.

Similarly, we use (7) and substitute $\delta_i = (\pi/2)^{1/2}(\hat{\mu}_i - \epsilon_i^{(1)})/s_i$ into (8) where the updated parameter $\epsilon_i^{(1)}$ is used. This yields

$$s_i^{(1)} \quad = \quad \left(\hat{\gamma}_i + (\hat{\mu}_i - \epsilon_i^{(1)})^2\right)^{1/2} \tag{11}$$

for $i = 1,\ldots n$ which provides the updated parameter $s^{(1)}$ for adaptive importance sampling.

Lastly, we update the skew parameter α using the estimated moments $\hat{\mu}$, $\hat{\gamma}$ and $\hat{\tau}$, and the previously updated parameters $\epsilon^{(1)}$ and $s^{(1)}$. First, using $\delta_i = (1+\alpha'\alpha)^{-1/2}\alpha_i$, it is easy to establish that $\alpha_i = (1-\delta'\delta)^{-1/2}\delta_i$. Then, we again use (7) and obtain $\delta_i^{(1)} = (\pi/2)^{1/2}(\hat{\mu}_i - \epsilon_i^{(1)})/s_i^{(1)}$ which is substituted into the expression for α_i which yields

$$\alpha_i^{(1)} = (1 - \delta^{(1)'}\delta^{(1)})^{-1/2}\delta_i^{(1)} \tag{12}$$

for $i = 1,\ldots,n$. We note that the method of moments estimation does not always respect constraints $\delta^{(1)'}\delta^{(1)} < 1$. We do not update the importance sampling parameters (and continue sampling) if the constraint $\delta^{(1)'}\delta^{(1)} < 1$ is not satisfied.

Therefore, (10), (11) and (12) provide the steps for obtaining updated parameters $\epsilon^{(1)}$, $s^{(1)}$ and $\alpha^{(1)}$. These parameters define an updated importance sampler $q^{(1)}(y)$ which hopefully better mimics the integrand $f(y)$.

In the second stage of adaptation, we sample $y^{(1)}, \ldots, y^{(N)}$ according to the importance sampler $q^{(1)}(y)$ and repeat the above fitting process. This leads to a new importance sampler $q^{(2)}(y)$.

Obviously, adaptation can continue by repeated sampling and fitting. However, in the applications which we have considered, only a few (e.g. less than six) rounds of adaptation are required since subsequent iterations typically result in marginal changes to the current importance sampler. We terminate adaptation by checking the relative difference in the standard error estimator of normalizing constant estimator $\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} g(y^{(i)})/q(y^{(i)})$. The termination criterion we use is $|SE^{(t+1)}(\hat{Z}) - SE^{(t)}(\hat{Z})|/SE^{(t)}(\hat{Z}) < \eta$, where the superscript index t denotes t-th iteration, $SE(\hat{Z})$ denotes the standard error estimator of $\hat{Z}$ and can be computed by

$$SE(\hat{Z}) = \frac{1}{N}\sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}\frac{g^2(y^{(i)})}{q^2(y^{(i)})} - \left(\frac{1}{N}\sum_{i=1}^{N}\frac{g(y^{(i)})}{q(y^{(i)})}\right)^2\right]}.$$

Also, there may be various strategies in combining the moment estimates from each round of sampling. For example, one might take weighted averages where the weights correspond to the inverse of the standard errors of the estimates. Alternatively, one might simply ignore all of the estimates obtained from the early rounds of adaptation, and instead, approximate integrals based on only the results from a long run using the final importance sampler. The computational complexity of the adaptive importance sampler is a linear function of number of importance samples N, and total number of iterations T. Hence, the total cost is $O(NT)$.

As a by-product of estimating integrals, we note that the proposed algorithm attempts

to find a good importance sampler, i.e. an importance sampler that mimics the posterior density in Bayesian calculations. In the Metropolis-Hastings algorithm, a popular strategy is to seek a candidate generating density that mimics the posterior. The resultant MCMC (Markov chain Monte Carlo) algorithm is often referred to as independence sampling. Therefore, our algorithm may also be seen as a pre-conditioner to MCMC. By determining an importance sampler, we obtain a candidate generating density for Metropolis-Hastings independence sampling. The adaptive importance sampling falls within the standard importance sampling framework. Hence, the consistency property holds for the adaptive importance sampling with a restricted skew-normal distribution. When the number of samples goes to infinity ($N \to \infty$), the approximated integral (e.g. Equation (4)) is arbitrarily close to the true value. Algorithm 1 provides the pseudo-code of the adaptive importance sampling with a restricted skew normal proposal.

3 RESTRICTED SKEW-STUDENT DISTRIBUTIONS

Although the multivariate normal distribution has been used extensively in importance sampling applications, the multivariate Student can be implemented with no real additional difficulties (see Evans and Swartz 2000). An advantage of the Student over the normal is longer "tails". In some applications, the shorter tails of the normal may lead to importance sampling estimators with infinite variance.

We therefore consider an extension of the proposed adaptive importance sampling algorithm where restricted skew-normal distributions are replaced by restricted skew-Student distributions. With some change of notation, we follow the development of skew-Student distributions as given by Azzalini and Capitanio (2003).

Algorithm 1 Adaptive importance sampling with a restricted skew-normal proposal

Inputs: (a) Observations y; (b) Target integrand f; (c) Termination threshold η; (d) Number of importance samples N.
Outputs: (a) A skew normal proposal proposal with parameters $\epsilon^{(T)}$, $s^{(T)}$, $\alpha^{(T)}$; (b) Integral of interest $\hat{I}(f)$.
Initialize skew normal parameters: $\epsilon^{(0)}$, $s^{(0)}$ and $\alpha^{(0)}$.
Initialize relative difference in the standard error estimator of $\hat{Z}$, $\tilde{\eta} = \infty$.
Initialize iteration index $t = 1$.
while $\tilde{\eta} < \eta$ **do**
    if $t = 1$ **then**
        $\alpha^{(1)} = \alpha^{(0)}$, $\epsilon^{(1)} = \epsilon^{(0)}$ and $s^{(1)} = s^{(0)}$.
    else
        Update $\alpha^{(t)}$, $\mu^{(t)}$ and $s^{(t)}$ according to Eq (10-12).
    for $i \in \{1, \ldots, N\}$ **do**
        Propose samples according to $y_i \sim RSN_n(\epsilon^{(t)}, s^{(t)}, \alpha^{(t)})$.
    Compute $\tilde{\eta} = |SE^{(t+1)}(\hat{Z}) - SE^{(t)}(\hat{Z})|/SE^{(t)}(\hat{Z})$.
    Update iteration index $t = t + 1$.
Set $T = t$, $\epsilon^{(T)} = \epsilon^{(t)}$, $s^{(T)} = s^{(t)}$, $\alpha^{(T)} = \alpha^{(t)}$ and compute integral of interest using Eq (4).

Letting $z \sim SSN_n(\Omega, \alpha)$, we define

$$\begin{aligned} y &= \epsilon + Sz\sqrt{v/W} \\ &= (\epsilon + Sz)\sqrt{v/W} + \epsilon\left(1 - \sqrt{v/W}\right) \end{aligned} \tag{13}$$

where $W \sim \chi^2_v$ is distributed independently of z, and $v > 3$ so that the third moment $\mathrm{E}(y^3)$ is finite. This is the analogue of the traditional scale mixing with respect to the chi-squared when transforming variables from multivariate normal to multivariate Student. The random variable y in (13) has a skew-Student$_n(v, \epsilon, S, \Omega, \alpha)$ where v is referred to as the degrees of freedom. We note that as $v \to \infty$, the distribution of y converges to the $SN_n(\epsilon, S, \Omega, \alpha)$ distribution, and therefore the skew-Student family can be seen as a generalization of the skew-normal family. Although the notation differs, the skew-Student$_n(v, \epsilon, S, \Omega, \alpha)$ distributions coincide with the distributions in Branco and Dey (2001).

In the restricted skew-Student setting, we require $\Omega = I$. For adaptive restricted skew-Student importance sampling, restricted skew-Student variates y are generated following the construction in (13). Specifically, we generate $(\epsilon + Sz) \sim RSN_n(\epsilon, s, \alpha)$ as previously discussed in Section 2 and then generate $W \sim \chi^2_v$. Then using the second expression in (13) with appropriate algebra yields the restricted skew-Student variate y.

Restricted skew-Student importance sampling also requires the evaluation of the density of the importance sampler and this is given by

$$q(y) = \frac{2\Gamma(\frac{n+v}{2})(v/2)^{v/2}}{(2\pi)^{n/2}\Gamma(\frac{v}{2})\prod_{i=1}^n s_i}\ (p(y)/2)^{-(n+v)/2}\mathrm{Prob}\left(T \le \sqrt{(n+v)/p(y)}\sum_{i=1}^n \alpha_i(y_i - \epsilon_i)/s_i\right)$$

where $p(y) = v + \sum_{i=1}^n (y_i - \epsilon_i)^2/s_i^2$ and $T \sim \text{Student}_{n+v}$. Therefore we also require the evaluation of the distribution function of the univariate Student distribution and this is available in many statistical software packages such as the R programming language.

The final step relevant to the implementation of adaptive restricted skew-Student importance sampling is the determination of a member of the restricted skew-Student family in each round of adaptation. Consider then the first round of adaptation where the same steps are taken in subsequent rounds. We follow the approach previously developed for adaptive restricted skew-normal importance sampling by matching moments. The initial parameters can be obtained via the approach outlined in Section 2. As in expressions (7), (8) and (9), we use coordinate-wise formulae for the first moment from the origin

$$\hat{\mu}_i = \epsilon_i + s_i \delta_i \left(\frac{\nu}{\pi}\right)^{1/2} \frac{\Gamma(\frac{1}{2}\nu - \frac{1}{2})}{\Gamma(\frac{1}{2}\nu)} \tag{14}$$

the second moment from the origin

$$\hat{\gamma}_i = s_i^2 \left[\frac{\nu}{\nu - 2} - \left\{\delta_i \left(\frac{\nu}{\pi}\right)^{1/2} \frac{\Gamma(\frac{1}{2}\nu - \frac{1}{2})}{\Gamma(\frac{1}{2}\nu)}\right\}^2\right] \tag{15}$$

and the third moment from the origin

$$\hat{\tau}_i = \frac{3\upsilon s_i^2}{(\upsilon - 3)(\upsilon - 2)}(\hat{\mu}_i - \epsilon_i) - \frac{\pi\Gamma^2(\frac{\upsilon}{2})}{(\upsilon - 3)\Gamma^2(\frac{1}{2}\upsilon - \frac{1}{2})}(\hat{\mu}_i - \epsilon_i)^3 + 2(\hat{\mu}_i - \epsilon_i)^3. \tag{16}$$

We note that a difference between (7)-(9) and (14)-(16) is that the latter equations also depend on υ. As before, we solve these equations numerically (e.g. using the Newton-Raphson method) to give updated skew-Student importance sampling parameters. Parameter ϵ_i does not have a closed form solution. Therefore, we numerically solve

$$\hat{\tau}_i = \frac{3}{(\upsilon - 3)}[\hat{\gamma}_i + (\hat{\mu}_i - \hat{\epsilon}_i)^2](\hat{\mu}_i - \epsilon_i) - \left[\frac{\pi\Gamma^2(\upsilon/2)}{(\upsilon - 3)\Gamma^2((\upsilon - 1)/2)} - 2\right](\hat{\mu}_i - \epsilon_i)^3 \tag{17}$$

to obtain the method of moments estimates. The method of moments estimates s_i and δ_i are given by

$$s_i = \left[\frac{\nu - 2}{\nu}(\hat{\gamma}_i + (\hat{\mu}_i - \hat{\epsilon}_i)^2)\right]^{1/2} \tag{18}$$

and

$$\delta_i = \frac{\hat{\mu}_i - \epsilon_i}{s_i\left(\frac{\nu}{\pi}\right)^{1/2}\frac{\Gamma(\frac{1}{2}(\nu-1))}{\Gamma(\frac{1}{2}\nu)}} \ . \tag{19}$$

Therefore equations (17), (18) and (19) provide the parameter updates for successive iterations of adaptive skew-Student importance sampling. Note that in adaptive skew-Student importance sampling, we do not currently have a good method for fitting the degrees of freedom parameter $v > 3$. Instead, we take the approach of setting v in advance where smaller degrees of freedom lead to longer tailed distributions. Our view is that longer tails (e.g. $v = 5$) are better in terms of insuring estimators with finite variances. Alternatively, one can run the algorithm several times with different choices of v, and select the value which gives an estimator with the smallest variance.

4 EXAMPLES

4.1 Example 1

We first consider a 12-dimensional integral where the integrand $f(x)$ is the density of the $RSN_{12}(\epsilon, s, \alpha)$ distribution where

$$\begin{aligned} \epsilon &= (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12)' \\ s &= (1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 3\ 3\ 3\ 3)' \\ \alpha &= (1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2)' \end{aligned}$$

This is a good test case as the dimensionality is sufficiently high to be problematic for most asymptotic approximations and quadrature methods (Evans and Swartz 2000), and ideally, we would like to see the restricted skew-normal importance sampler approach the distribution given by the integrand.

In this problem, since $f(x)$ is a density, the inverse normalizing constant $I(f) = 1.0$ and we use the estimates $\hat{I}(f)$ to assess the performance of the adaptive importance sampling algorithm. We chose $N = 30,000,000$ iterations in each adaptive step. Each round of sampling and adaptation required approximately five minutes of computation on a laptop computer. The integrand is in the family of proposals used for adaptation.

In Table 1, we give some summary results of the algorithm based on four stages of adaptation. In particular we provide $\hat{I}(f)$ and its standard error. We also provide the parameters $s^{(1)}, \ldots, s^{(4)}$ updated after each round of sampling. We observe that common approach of multivariate normal importance sampling (i.e. stage (1)) is unreliable as the estimate $\hat{I}(f)$ is nearly eight standard errors from its true value. This is a typical problem with importance sampling (Evans and Swartz 2000) when the importance sampler does not adequately mimic the integrand. However, we do see that the adaptive algorithm quickly delivers a good fitting importance sampler, where after four iterations, the s vector in the importance sampler matches the s vector of the integrand. We note that after four rounds of adaptation, the standard error of $\hat{I}(f)$ reduced to 0.000002. Therefore, ignoring the fact that the standard error in the first round underestimates the true standard error, the adaptative algorithm provides an increase in efficiency by a factor of $(0.008619/0.000002)^2 \approx 1.9 \times 10^7$. In other words, at least 1.9×10^7 times as much multivariate normal importance sampling is required to obtain the equivalent precision as restricted skew-normal importance sampling based on $(\epsilon^{(4)}, s^{(4)}, \alpha^{(4)})$. Finally, we remark that in a similar fashion to s, the importance sampling parameters ϵ also converged quickly to the parameter values corresponding to the integrand. However, the α vector does not converge quickly. After six rounds of adaptation,

$$\alpha^{(6)} = (0.97\ 0.90\ 0.96\ 0.80\ 0.86\ 0.92\ 1.93\ 1.89\ 1.90\ 1.90\ 1.88\ 1.91)'.$$

We observe a lack of speedy convergence for the skew parameter α. The slow convergence

of the estimate of α is not such an important issue, since the observed numerical differences from the true values do not translate into appreciable differences of the likelihood. Recall we are trying to find an importance sampler that mimics the integrand, and any skew-normal that fits reasonably well accomplishes our goal whether or not it is optimal.

	Restricted Skew-normal Stage of Adaptation				Restricted Skew-Student$_{10}$ Stage of Adaptation			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$\hat{I}(f)$	0.93021	0.99954	1.00000	1.00002	1.00182	1.00000	0.99999	0.99993
SE$(\hat{I}(f))$	0.008619	0.000032	0.000003	0.000002	0.003404	0.000159	0.000159	0.000159
s_1	0.96	1.03	1.00	1.00	1.04	0.88	0.89	0.89
s_2	0.99	1.04	1.00	1.00	0.46	0.89	0.89	0.89
s_3	1.30	1.04	1.00	1.00	1.05	0.89	0.89	0.76
s_4	0.95	1.03	1.00	1.00	0.66	0.88	0.89	0.89
s_5	0.71	1.93	2.00	2.00	2.00	1.77	1.77	1.77
s_6	2.30	2.03	2.00	1.99	2.71	1.77	1.77	1.77
s_7	1.55	2.07	2.00	2.00	1.70	1.71	1.71	1.71
s_8	2.47	2.13	2.00	2.00	1.94	1.71	1.71	1.71
s_9	3.68	3.14	3.00	3.00	3.28	2.57	2.57	2.57
s_{10}	2.01	3.10	3.00	3.01	3.18	2.57	2.57	2.57
s_{11}	3.15	3.05	2.99	3.01	2.96	2.57	2.57	2.57
s_{12}	2.97	2.95	3.01	3.00	2.24	2.57	2.57	2.57

Table 1: Some summary results corresponding to Example 1.

In Table 1, we also provide some summary results corresponding to four rounds of adaptive restricted skew-Student$_{10}$ importance sampling. Since the integrand is restricted skew-normal, we do not expect the algorithm to perform quite as well as adaptive restricted skew-normal importance sampling, and this is the case. However, the standard error of $\hat{I}(f)$ in restricted skew-Student$_{10}$ importance sampling quickly (i.e. two rounds) reduced to 0.000159 and this represents an increase in efficiency by a factor of $(0.003404/0.000159)^2 \approx 458$ over standard multivariate normal importance sampling. Additional simulations (not reported) indicate that the algorithm performs less well as the

degrees of freedom of the importance sampler decrease and we move away from normality. In the Appendix, we further investigate tradeoffs on the number of adaptations and samples on a lower dimensional version of this problem.

4.2 Example 2

Our second example is taken from the second test case in Evans and Swartz (1995). This is a 9-dimensional integral based on the Bayesian analysis of a contingency table with parameter vector $\theta = (\theta_1, \ldots, \theta_9)'$. The data involves the cross-classification of 132 long-term schizophrenic patients into three row categories describing the frequency of hospital visits and three column categories describing the length of stay. Evans and Swartz (1995) applied various integration techniques to the calculation of posterior means in this test case. In the example, they concluded that subregion adaptive integration (Genz 1991) proved excessively time consuming, and a MCMC approach based on a Metropolis independence chain suffered from high correlations.

We applied the proposed adaptive importance sampling approach based on the restricted skew-Student distribution to this non-trivial problem. We chose $N = 10,000,000$ iterations in each adaptive step for direct comparison with Student$_5$ importance sampling reported in Evans and Swartz (1995). We consider the estimation of the posterior mean of θ_1 as this is the most problematic integral considered in Evans and Swartz (1995). In Table 2, we provide the exact value (based on 41 hours of Student importance sampling) and the adaptive skew-Student estimates after the fourth round of adaptation. We also provide the Normal (non-adaptive) importance sampling results. We observe that adaptive restricted skew-normal importance sampling provides an improvement over Normal importance sampling as the standard errors are reduced. We suspect that the improvements are not as great as in Example 1 as the posterior is not as heavily skewed. We

observe a slight benefit in adaptive restricted skew-Student importance sampling over adaptive restricted skew-normal importance sampling. This supports the assertion that little is lost by choosing the longer tailed skew-Student over the skew-normal.

We also compare the aforementioned importance sampling approaches in terms of computing time using a 2.3 GHz Intel Core 19 processor. With $N = 10,000$, normal (non-adaptive) importance sampling required 0.31 seconds of computation, each round of adaption for adaptive importance sampling with the restricted skew-normal proposal required 0.86 seconds, and each round of adaption for adaptive importance sampling with the restricted skew-Student$_{10}$ proposal required 0.62 seconds.

			After Four Rounds of Adaptation	
	Exact	Normal	Skew-normal	Skew-Student$_{10}$
Estimates	0.4215	0.4323	0.4298	0.4269
Std Error		(0.054)	(0.0486)	(0.04301)

Table 2: Estimates of the posterior mean of θ_1 and standard errors corresponding to Example 2.

4.3 Example 3

In the groundbreaking book "Basketball on Paper", by Oliver (2004), novel statistics for player evaluation in the National Basketball Association (NBA) were developed. These statistics have given rise to various aggregate offensive statistics (provided by *www.basketball-reference.com*) such as

- $X_1 \equiv$ offensive rating,
- $X_2 \equiv$ offensive win shares,

and defensive aggregate statistics such as

- $X_3 \equiv$ defensive rating,
- $X_4 \equiv$ defensive win shares.

It would be informative to investigate these new measures as they relate to player salaries (s). In particular, we model

$$ln(s) = (\beta_1 x_1 + \beta_2 x_2)(\beta_3 x_3 + \beta_4 x_4) + \epsilon, \tag{20}$$

where $\epsilon \sim N(0, \sigma^2)$. Therefore, salaries are modelled as log-normal distributions, where there is a long history of using lognormal distributions to successfully model incomes (see chapter 11 of Aitchison and Brown, 1966). In (20), we have expressed mean player salary as a product of offensive and defensive attributes. In this formulation, a 1% increase in either offensive or defensive attributes provides a 1% increase in the player's mean log salary.

Using data from the 2018-2019 NBA season involving $i = 1, \ldots, n = 358$ players who have played at least 500 minutes during the season, this leads to the likelihood

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}\big(ln(s_i) - (\beta_1 x_{i1} + \beta_2 x_{i2})(\beta_3 x_{i3} + \beta_4 x_{i4})\big)^2 \right\},$$

involving unknown parameters $\theta = (\beta_1, \beta_2, \beta_3, \beta_4, \sigma)$. Salary data is available from the website *www.kaggle.com*.

One may be interested in providing probability assessments on the parameters and also wish to introduce prior knowledge regarding the parameters via $\pi(\theta)$. This is suggestive of a Bayesian analysis and leads to integrals of the form

$$I(m) = \int m(\theta) L(\theta) \pi(\theta) d\theta \tag{21}$$

where $m(\theta)$ is some function of interest regarding the parameters.

The 5-dimensional integral $I(m)$ in (21) is not analytically available and requires some sort of approximation technique. Note also that the dimensionality of the problem increases as we add more covariates X. For simplicity, we set $m(\theta) = \pi(\theta) = 1$ and compare importance sampling approaches. The computation of (21) may be a challenging task for a Gibbs sampling algorithm as the determination of full conditional distributions is not straightforward. In Table 3, we provide the exact value of $I(m)$ (based on 3×10^9 iterations of normal importance sampling), the adaptive skew normal estimates and the adaptive skew-Student estimates after four rounds of adaptation. Note that alternative and perhaps "smarter" importance samplers might have been chosen especially as the parameter σ differs in structure from the β parameters. However, in this demonstration, we are interested in choosing a rough and ready importance sampling technique. We chose $N = 3,000,000$ iterations in each adaptive step. We also provide the Normal (non-adaptive) importance sampling results ($N = 3,000,000$). We observe that adaptive restricted skew-normal importance sampling provides an improvement over Normal importance sampling as the standard errors are reduced.

From Table 3, we observe a slight benefit in adaptive restricted skew-Student importance sampling over adaptive restricted skew-normal importance sampling. This supports the assertion that little is lost by choosing the longer tailed skew-Student over the skew-

			After Four Rounds of Adaptation	
	Exact	Normal	Skew-normal	Skew-Student$_{10}$
Estimates	-608.66	-617.55	-610.75	-609.86
Std Error		(-617.89)	(-612.84)	(-611.22)

Table 3: Estimates of the normalizing constant and standard errors corresponding to Example 3.

normal. It also suggests that if you are going to use an importance sampling algorithm, the adaptive algorithms presented here may help some, but don't appear as they will be detrimental.

We also compare the importance sampling approaches in terms of computing time using a 2.3 GHz Intel Core 19 processor. With $N = 10,000$, the normal (non-adaptive) importance sampling required 0.26 seconds of computation, each round of adaption for adaptive importance sampling with skew-normal proposal required 0.45 seconds and each round of adaption for adaptive importance sampling with the skew-Student$_{10}$ proposal required 0.43 seconds.

5 DISCUSSION

This paper introduces new importance sampling algorithms for integrals defined on $\mathcal{R}^n$. The approach is adaptive and relies on the restricted skew-normal and restricted skew-Student families of distributions. Although there is no "universal" approach to multivariate integration, the proposed algorithms generalize the widely used approach based on multivariate normal importance sampling.

Like all integration techniques, our approach suffers from the curse of dimensionality where sampling and fitting become more complex as the dimension of the integral rises. Nevertheless, we have demonstrated the utility of the approach in several challenging

examples.

No doubt, there are many variations of our algorithm that may be considered in future research. One potentially fruitful direction is the extension of our approach to more general skew-elliptical distributions. A second research direction involves the combination of skew family proposals with other adaptive schemes proposed in the recent literature (e.g. Cornuet et al. 2012, Elvira et al. 2019). Lastly, we plan to explore the theoretical efficiency of our approach.

6 REFERENCES

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). "Importance sampling: Intrinsic dimension and computational cost", *Statistical Science*, 405-431.

Aitchison, J. and Brown, J.A.C. (1966). *The Lognormal Distribution*, Cambridge University Press.

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

Azzalini, A. and Capitanio, A. (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution", *Journal of the Royal Statistical Society Series B*, 65, 367-389.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families.* IMS monographs. Cambridge, UK: Cambridge University Press.

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew-normal distribution", *Biometrika*, 83, 715-726.

Branco, M.D. and Dey, D.K. (2001). "A general class of multivariate skew-elliptical distributions", *Journal of Multivariate Analysis*, 79, 99-113.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J. and Djuric, P. M. (2017). "Adaptive importance sampling: the past, the present, and the future", *IEEE Signal Processing Magazine*, 34(4), 60-79.

Cornuet, J. M., Marin, J. M., Mira, A. and Robert, C. P. (2012). "Adaptive multiple importance sampling", *Scandinavian Journal of Statistics*, 39(4):798–812.

Dalla Valle, A. (2004). "The skew-normal distribution", In *Skew-Elliptical Distributions and their Applications* (M.G. Genton, editor), Chapman and Hall, 3-24.

Elvira, V. and Martino, L. (2021). "Advances in importance sampling", *arXiv preprint arXiv:2102.05407.*

Elvira, V. and Martino, L., Luengo, D. and Bugallo, M. F. (2017). "Improving Population Monte Carlo: Alternative weighting and resampling schemes", *Signal Processing*, 131(12):77–91.

Elvira, V. and Martino, L., Luengo, D. and Bugallo, M. F. (2019). "Generalized multiple importance sampling", *Statistical Science*, 34(1):129–155.

Evans, M. and Swartz, T.B. (1995). "Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems", *Statistical Science*, 10, 254-272.

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

Genton, M.G., He, L. and Liu, X. (2001). "Moments of skew-normal random vectors and their quadratic forms", *Statistics & Probability Letters*, 51, 319-325.

Genton, M.G. (editor) (2004). *Skew-Elliptical Distributions and their Applications*, Chapman and Hall.

Genz, A. (1991). "Subregion adaptive algorithms for multiple integrals", In *Statistical Multiple Integration* (N. Flournoy and R.K. Tsutakawa, editors), American Mathematical Society, Providence, Rhode Island, 23-31.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (editors) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

He, H. Y. and Owen, A. B. (2014). "Optimal mixture weights in multiple importance sampling", *arXiv preprint arXiv:1411.3954*.

Ionides, E.L. (2008). "Truncated importance sampling", *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Martino, L., Elvira, V., Luengo, D. and Corander, J. (2017). "Layered adaptive importance sampling", *Statistics and Computing*, 27(3):599–623.

Metropolis, N. and Ulam, S. (1949). "The Monte Carlo method", *Journal of the American Statistical Association*, 44, 335-341.

Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis. Potomac Books*, Inc..

Sbert, M., Havran, V. and Szirmay-Kalos, L. (2019). "Optimal Deterministic Mixture Sampling", *In Eurographics (Short Papers)*, 73-76.

Sbert, M., Havran, V., Szirmay-Kalos, L. and Elvira, V. (2018). "Multiple importance sampling characterization by weighted mean invariance", *The Visual Computer*, 34(6), 843-852.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). "Pareto smoothed importance sampling", *arXiv preprint arXiv:1507.02646*.

7 Appendix

We are interested in the tradeoffs between the number of adaptations and the number of samples in adaptive importance sampling (AIS). We consider a simpler version of Example 1 where the integral is 8-dimensional. The integrand $f(x)$ is the density of the $RSN_8(\epsilon, s, \alpha)$ distribution where

$$\begin{aligned} \epsilon &= (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)' \\ s &= (1\ 1\ 1\ 2\ 2\ 2\ 3\ 3)' \\ \alpha &= (1\ 1\ 1\ 1\ 2\ 2\ 2\ 2)'. \end{aligned}$$

We select two levels of iterations $N = 10^6$ and $N = 3 \cdot 10^6$ in each adaptive step, and investigate the performance of AIS as a function of N. We set the maximum number of adaptive stages to 10. For each level of N, we repeat AIS 10 times. Figure 3 displays the normalizing constant estimates $\hat{I}(f)$ and the standard deviation estimates as a function of N. The normalizing constant with a higher value of iterations admits lower bias and standard deviation. The AIS improves as stage of adaption increases; they tend to be stable after stage of adaption reaches 4. A larger value of iterations can improve the performance of AIS more significantly than an increase in iterations.

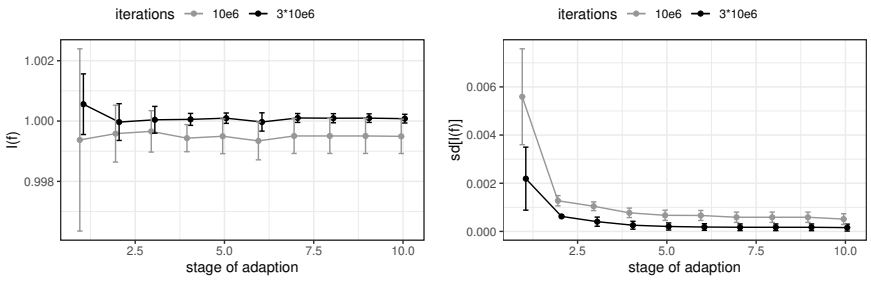

Figure 3: Normalizing constant estimates $\hat{I}(f)$ and the standard deviation estimates as a function of adaptive steps for two levels of iterations $N = 10^6$ and $N = 3 \cdot 10^6$.