Robust Estimation and Inference for Importance Sampling Estimators with Infinite Variance*

Joshua C. C. Chan
Purdue University

Chenghan Hou
Australian National University

Thomas Tao Yang
Australian National University

August 2018

Abstract

Importance sampling is a popular Monte Carlo method used in a variety of areas in econometrics. When the variance of the importance sampling estimator is infinite, the central limit theorem does not apply and estimates tend to be erratic even when the simulation size is large. We consider asymptotic trimming in such a setting. Specifically, we propose a bias-corrected tail-trimmed estimator such that it is consistent and has finite variance. We show that the proposed estimator is asymptotically normal, and has good finite-sample properties in a Monte Carlo study.

Keywords: simulated maximum likelihood, bias correction, stochastic volatility, importance sampling

JEL classification: C11, C32, C52

*A previous version of this paper was circulated under the title "Asymptotic Trimming for Importance Sampling Estimators with Infinite Variance".

1 Introduction

Importance sampling is widely used in econometrics to estimate integrals that do not have closed-form solutions. This Monte Carlo technique is especially important in maximum likelihood estimation of latent variable models—such as stochastic volatility and discrete choice models—where it is used to approximate intractable likelihood functions. Examples of maximum likelihood estimators computed using importance sampling for stochastic volatility models include Sandmann and Koopman (1998) and Koopman and Hol Uspensky (2002); for discrete choice models, see, e.g., Geweke, Keane and Runkle (1994) and Hajivassiliou and McFadden (1998). In the Bayesian literature, importance sampling is used in a variety of areas, including model comparison (e.g., Frühwirth-Schnatter, 1995; Frühwirth-Schnatter and Wagner, 2008; Chan and Eisenstat, 2015) and posterior simulation (e.g., Hoogerheide, Opschoor and Van Dijk, 2012; Pitt et al., 2012; Tran et al., 2014).

In his seminal paper, Geweke (1989) cautions that importance sampling should only be used when one can ensure the variance of estimator is finite. This is because when this finite variance condition fails, the central limit theorem does not apply, and the importance sampling estimator converges slower than the usual parametric rate. Even though in principle the estimator remains consistent, in practice it can be strongly biased and erratic even when the simulation size is huge. Despite these warnings, practitioners often ignore to check the finite variance condition, as it is challenging to verify in high-dimensional settings.

Koopman, Shephard and Creal (2009) make an important contribution by proposing a test to assess the validity of this finite variance assumption. However, it is not clear how one should proceed when the finite variance condition fails. This paper proposes a way forward in such settings. We ask if the original infinite-variance estimator can be modified such that the new estimator is consistent, and more importantly, asymptotically normal.

To that end, we consider asymptotic trimming (see, e.g., Hill, 2010, 2013). Specifically, we trim the right tail of the importance sampling weights in a way that the tail-trimmed estimator converges to the estimand faster than the untrimmed estimator. Trimming large importance sampling weights obviously introduces bias, and we show that this bias dominates the variance asymptotically. To overcome this problem, we propose a biased-corrected tail-trimmed estimator that has finite variance. We further prove that this estimator, after proper studentization, is asymptotically normal.

We demonstrate the good properties of the proposed estimator in a Monte Carlo study.

We show that in cases when the variance of the original importance sampling estimator does not exist, its sampling distribution can be highly skewed. In contrast, the biased-corrected tail-trimmed estimator performs substantially better with an approximate normal distribution. We illustrate the proposed methodology with an application on fitting daily financial returns using a standard stochastic volatility model.

2 Importance Sampling

Importance sampling is a variance reduction technique that can be traced back to Kahn and Marshall (1953) and Marshall (1956). Kloek and Van Dijk (1978) appear to be the first application of importance sampling in econometrics. To define importance sampling, suppose we wish to evaluate the following integral that does not have an analytical expression:

$$\Psi = \int_{\mathcal{X}} H(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} < \infty,$$

where $\mathbf{x}$ is a $k \times 1$ vector, H is a function mapping from $\mathbb{R}^k$ to $\mathbb{R}$, and f is a density function with support $\mathcal{X} \subset \mathbb{R}^k$. Let g be another density function that *dominates* Hf, i.e., $g(\mathbf{x}) = 0$ implies that $H(\mathbf{x})f(\mathbf{x}) = 0$. Then, Ψ can be rewritten as

$$\Psi = \int_{\mathcal{X}} H(\mathbf{x})\frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Therefore, we can estimate Ψ by the *importance sampling estimator*:

$$\widehat{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{H(\mathbf{X}_i)f(\mathbf{X}_i)}{g(\mathbf{X}_i)}, \tag{1}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent draws from the density g, which is often called the *importance sampling density*.

It is easy to see that $\mathbb{E}_g\widehat{\Psi}_n = \Psi$, where we make it explicit that the expectation is taken with respect to g. In addition, $\widehat{\Psi}_n$ is a consistent estimator of Ψ. We refer the readers to Kroese, Taimre and Botev (2013) for a more detailed discussion of importance sampling. For later reference, we define $W(\mathbf{X}) = H(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ and write $W(\mathbf{X}_i)$ as simply W_i hereafter.

2.1 Testing for Existence of Variance

One can establish the asymptotic distribution of the importance sampling estimator $\widehat{\Psi}_n$ if W_i has a finite second moment. However, this is difficult to verify in practice. This problem has been recognized by Monahan (1993, 2001) and Koopman, Shephard and Creal (2009). In particular, Monahan (1993, 2001) provides a test for the existence of the second moment of W_i when the right tail of $|W_i|$ satisfies the condition

$$1 - F(w) = c_w w^{-\alpha} \left[1 + O\left(w^{-\beta}\right)\right] \tag{2}$$

as $w \to \infty$, where c_w, α and β are positive constants, and F is the cumulative distribution function of $|W_i|$. It follows from (2) that $\mathbb{E}|W_i|^j < \infty$ if $j < \alpha$ and $\mathbb{E}|W_i|^j = \infty$ otherwise. Consequently, testing the existence of the second moment is equivalent to testing $\alpha > 2$.

Alternatively, Koopman, Shephard and Creal (2009) construct a test statistics from the maximum likelihood estimates of the generalised Pareto distribution with density

$$f(z;\xi,\eta) = \eta^{-1}\left(1 + \frac{z}{\xi\eta}\right)^{-(\xi+1)},$$

for $z \in D(\xi,\eta) > 0, \eta > 0$, where

$$D(\xi,\eta) = \begin{cases} [0,\infty), & \xi \geq 0, \\ [0,-\eta\xi], & \xi < 0. \end{cases}$$

The parameter ξ here plays the same role as α above. In fact, the generalised Pareto satisfies equation (2), with $\alpha = \xi$, $\beta \to \infty$ when $\xi > 0$.[1]

In our context H is the likelihood function and therefore $W_i = H(\mathbf{X}_i)f(\mathbf{X}_i)/g(\mathbf{X}_i) \geq 0$. From here onwards we assume W_i is bounded from the left, i.e., there exists a positive number c_l that $W_i > -c_l$, and we only need to discuss the right tail of W_i.

Below we formally present the assumptions and results in Monahan (1993, 2001). These assumptions are also needed when we develop our bais-corrected tail-trimmed estimator in the following sections.

To motivate the test in Monahan (1993, 2001), we first consider the following technical conditions.

[1]The result can be seen by a change of variable $w = z + \xi\eta$. Then the density function of w becomes $f(w;\xi,\eta) = \eta^{-1}(\xi^{-1}w/\eta)^{-\xi-1}$, for $z \in \tilde{D}(\xi,\eta) > 0, \eta > 0$, where $\tilde{D}(\xi,\eta) = \begin{cases} [\xi\eta,\infty), & \xi \geq 0, \\ [\xi\eta,0], & \xi < 0. \end{cases}$

Assumption 1 *There exists a positive number* c_l *such that* $W_i > -c_l$. *In addition, we assume* $\mathbb{E}(W_i) = \Psi < \infty$ *and*

$$\mathbb{P}(W_i > w) = L(w)\, w^{-\alpha}\left(1 + o\left(w^{-\beta}\right)\right),$$

as $w \to \infty$, $\alpha, \beta > 0$ *and* $L(w)$ *is a slowly varying function.*[2]

Assumption 2 *Let* $\{m_n\}$ *be a sequence of positive integers such that* $m_n \to \infty$ *and* $m_n = o\left(n^{2\beta/(2\beta+\alpha)}\right)$ *as* $n \to \infty$.

Assumption 3 *The random vectors* $\{\boldsymbol{X}_i\}$ *are independent and identically distributed.*

Next, we discuss the estimation of α. To that end, we define the following quantities

$$\begin{aligned} \bar{F}(y) &= 1 - F(y), \\ b\left(y^{-1}\right) &= F^{\leftarrow}(1-y), \end{aligned}$$

where $F^{\leftarrow}(y) = \inf\{w : F(w) \geq y\}$, $0 < y < 1$. We define the sample order statistics $W_{(i)}$ of from $W_1, \ldots, W_n$ such that $W_{(1)} \geq W_{(2)} \geq \cdots \geq W_{(n)}$.

One popular estimator of α proposed by Hill (1975) is:

$$\widehat{\alpha}^{-1} = \frac{1}{m_n} \sum_{i=1}^{m_n} \left(\log W_{(i)} - \log W_{(m_n)}\right), \tag{3}$$

where $m_n \to \infty$, $m_n/n \to 0$ as $n \to \infty$.

Assumptions 1–3 are one set of sufficient conditions for the asymptotic normality of $\widehat{\alpha}$. The role of β in Assumption 1 and Assumption 2 is to ensure that $\widehat{\alpha}$ is asymptotically normal. For a more detailed discussion on the sufficient conditions for the asymptotic normality of $\widehat{\alpha}$, see Haeusler and Teugels (1985).

Under Assumptions 1–3, the asymptotic normality of $\widehat{\alpha}$ can be established. We summarize this result in the following theorem.

Theorem 2.1 *Under Assumptions 1–3, we have*

$$m_n^{1/2}\left(\widehat{\alpha}^{-1} - \alpha^{-1}\right)/\alpha^{-1} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

[2] Slow variation is defined by $\lim_{w\to\infty} L(aw)/L(w) = 1$, for any $a > 0$. For more details and examples, see Resnick (1987).

The above theorem is a direct result of Corollary 4 in Haeusler and Teugels (1985).

Here the convergence rate of $\widehat{\alpha}^{-1}$ is $m_n^{1/2}$. Assumption 2 restricts m_n such that $m_n = o\left(n^{2\beta/(2\beta+\alpha)}\right)$. It implies that the larger the value α has—corresponding to a thinner tail of W_i—the slower the convergence rate of $\widehat{\alpha}^{-1}$ can possibly achieve. The parameter β also plays an important role here. However, the estimation of β is unfortunately very hard; to the best of our knowledge, its estimation remains an open question in the literature.

Finally, Monahan (1993, 2001) proposes testing the existence of the second moment of W_i using the following test.

$$\mathbb{H}_0 : \alpha \geq 2 \text{ versus } \mathbb{H}_1 : \alpha < 2. \tag{4}$$

If $\mathbb{H}_0$ is rejected, we then conclude that the variance of the importance sampling estimator is infinite.

We can proceed the hypothesis test in (4) using Theorem 2.1. Specifically, we construct a test statistic $m_n^{1/2}\left(\widehat{\alpha}^{-1} - 2^{-1}\right)/\widehat{\alpha}^{-1}$. Since this test statistic is asymptotically normal, we calculate the P-value as $1 - \Phi\left(m_n^{1/2}\left(\widehat{\alpha}^{-1} - 2^{-1}\right)/\widehat{\alpha}^{-1}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

2.2 A Tail-Trimmed Estimator

If α is larger than 2, then the variance of the importance sampling estimator exists. It follows that the estimator converges to Ψ at the usual parametric rate and is asymptotically normal. Conversely, if the test rejects the null hypothesis that $\alpha \geq 2$, then the central limit theorem does not apply and one might need a prohibitively large simulation size to obtain a reliable estimate.

A natural question is: how can we proceed in the latter scenario? The literature has focused on providing a test of $\mathbb{H}_0$ and has not offered any solution if $\mathbb{H}_0$ is rejected. We aim to take a first step in filling this gap. Specifically, we consider tail trimming, i.e., dropping some large values of W_i, in a such way that the resulting tail-trimmed estimator converges to Ψ faster than otherwise. More importantly, the modified estimator is asymptotically normal after proper studentization.

To that end, let $\{k_n\}$ be an intermediate order sequence. That is, $k_n \to \infty$ and $k_n/n \to 0$

as $n \to \infty$. We define a sequence $\{l_n\}$ as the $(1 - k_n/n)$-th quantile of W_i, i.e.,

$$\bar{F}(l_n) = \frac{k_n}{n}.$$

Next, we consider the following tail-trimmed estimator:

$$\widehat{\Psi}_n^* = \frac{1}{n} \sum_{i=1}^{n} W_i I(W_i < W_{(k_n)}), \tag{5}$$

where $I(\cdot)$ is the indicator function. In other words, the estimator $\widehat{\Psi}_n^*$ drops the largest k_n values of W_i.

Obviously, $\widehat{\Psi}_n^*$ is biased downward. To discuss the bias and variance, let $W_{n,i}^* = W_i I(W_i < l_n)$. Then the bias B_n and the variance S_n^2 of the tail-trimmed estimator are, respectively,

$$B_n = \mathbb{E}[W_i I(W_i \geq l_n)], \qquad S_n^2 = \mathbb{E}\left([W_{n,i}^* - \mathbb{E}W_{n,i}^*]^2\right).$$

Below we derive the asymptotic distribution of $\widehat{\Psi}_n^*$. For that purpose, we consider the following technical assumption.

Assumption $1'$ *There exists a positive number c_l such that $W_i > -c_l$. In addition, we assume $\mathbb{E}(W_i) = \Psi < \infty$ and*

$$\mathbb{P}(W_i > w) = L(w)\, w^{-\alpha}(1 + o(1)),$$

as $w \to \infty$, $c_w, \alpha, \beta > 0$ and $L(w)$ is a slowly varying function.

Assumption $1'$ imposes a tail restriction on W_i that is similar to, but weaker than, that in Assumption 1. We are able to do it because the stronger restriction in Assumption 1 is only used to ensure the asymptotic normality of $\widehat{\alpha}^{-1}$, which is not needed here.

Remark 1 (Assumptions on tails) The literature on heavy-tailed problems is huge. Assumptions 1 and $1'$ are some of the most popular sets of assumptions used to study these problems. Similar conditions have been imposed to analyze problems in production frontiers (see Daouia, Florens and Simar, 2010, and references therein), auctions (Hill and Shneyerov, 2013), Value at Risk (Linton and Xiao, 2013), data networks (Leland et al., 1994), and many others. However, it is very hard to verify these two assumptions in most cases. For the simple example in the simulation section (see Section 3), we verify Assumption 1′ in Appendix A and we show that Assumption 1′ is satisfied with $\alpha = (1+\epsilon_1)/\epsilon_1$. For a slightly more general multivariate case

$$
\begin{aligned}
f(\mathbf{x}) &= (2\pi)^{-d/2}\exp\left(-\sum_{j=1}^{d}\frac{x_j^2}{2}\right), \\
g(\mathbf{x}) &= (2\pi)^{-d/2}\Pi_{j=1}^{d}(1+\epsilon_j)^{1/2}\exp\left(-\sum_{j=1}^{d}\frac{(1+\epsilon_j)\,x_j^2}{2}\right), \text{ and } H(x)=1,
\end{aligned}
$$

where d is a positive integer, $\epsilon_j > 0$ for all $1 \le j \le d$, and $\mathbf{x} = (x_1, x_2, \ldots, x_d)'$, we show in Appendix A that Assumption 1′ is satisfied with $\alpha = (1+\max\{\epsilon_j\}_{j=1}^{d})/\max\{\epsilon_j\}_{j=1}^{d}$.

The following theorem establishes the asymptotic normality of $\widehat{\Psi}_n^*$, even when W_i does not have a finite variance.

Theorem 2.2 *Suppose $1 < \alpha \le 2$, $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$.. Under Assumptions 1′ and 3,*

$$
n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* - \Psi + B_n\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).
$$

The proof is given in Appendix C.

Remark 2 (Multivariate case) For multivariate $\mathbf{X}$, W remains a scalar. Consequently, everything would go through exactly as in the univariate X case. As noted in the literature, the heavy-tailed problem may happen more frequently and may be more severe for multivariate $\mathbf{X}$, because an appropriate choice of g is much harder. This can be seen from the simple k-variates example in Remark 1 where α is determined by the worse dimension of $\mathbf{X}$. The convergence rate is likely to be slower, with a smaller α, for multivariate $\mathbf{X}$. However, although the initial choice of g is hard in multivariate case, once g or a class of g is fixed, our analysis can be carried out exactly as in the univariate case.

Even though the tail-trimmed estimator $\widehat{\Psi}_n^*$ is asymptotically normal, the bias B_n dominates its distribution. We derive an approximate expression for B_n below. As an illustration, we strengthen the distribution condition in Assumption 1′ to $1 - F(w) =$

$c_w w^{-\alpha}\left[1+O\left(w^{-\beta}\right)\right]$ as $w \to \infty$ and $c_w, \alpha, \beta > 0$, then we have

$$
\begin{aligned}
B_n = \mathbb{E}[W_i I(W_i \geq l_n)] &\approx \int_0^{k_n/n} b\left(u^{-1}\right) \mathrm{d}u \\
&\approx c_w^{1/\alpha} \frac{\alpha}{\alpha-1}\left(\frac{k_n}{n}\right)^{-1/\alpha+1} \\
&= \frac{\alpha}{\alpha-1}\frac{k_n}{n} l_n.
\end{aligned} \tag{6}
$$

In the case when $1 < \alpha < 2$, we have $S_n^2 \approx K(n/k_n)^{1/\alpha-1}$ for some constant K. Hence, the ratio of the bias and standard deviation is given by

$$
n^{1/2}\frac{B_n}{S_n} \approx n^{1/2}\left[K\left(\frac{n}{k_n}\right)^{2/\alpha-1}\right]^{-1/2}\left(\frac{n}{k_n}\right)^{1/\alpha-1} \approx K^{-1/2}k_n^{1/2},
$$

which tends to infinity as $n \to \infty$. This implies that the bias B_n dominates the distribution of $\widehat{\Psi}_n^*$ asymptotically. Consequently, it is vital to correct for this bias. In the next section, we consider a bias-corrected version of the tail-trimmed estimator.

2.3 Bias-Corrected Tail-Trimmed Estimator

In this section we introduce the bias-corrected tail-trimmed estimator and show that it is asymptotically normal. To that end, we need to restrict the tails of the importance sampling estimator to satisfy Assumption 1 instead of Assumption 1′—the former is needed to establish the convergence rate of $\widehat{\alpha}$.

It follows from (6) that a natural estimator for the bias is:

$$
\widehat{B}_n = \frac{\widehat{\alpha}}{\widehat{\alpha}-1}\frac{k_n}{n} W_{(k_n)}, \tag{7}
$$

where we use $W_{(k_n)}$ as an approximation of l_n. Then, the bias-corrected estimator is given by

$$
\widehat{\Psi}_n^{(b)} = \widehat{\Psi}_n^* + \widehat{B}_n. \tag{8}
$$

To derive a central limit theorem for $\widehat{\Psi}_n^{(b)}$, we impose the following technical condition on k_n:

Assumption 4 *Let $\{k_n\}$ be a sequence of positive integers such that $k_n \to \infty$ and*

$k_n/m_n \to 0$ *as* $n \to \infty$.

Recall that the sequence $\{m_n\}$ is used to estimate α in (3). Under Assumption 4, m_n diverges faster than k_n so that the first step estimation of α does not affect the asymptotics of the bias-corrected estimator $\widehat{\Psi}_n^{(b)}$. This can be seen from Theorem 2.1, where the convergence rate of $\widehat{\alpha}$ is $m_n^{1/2}$. In the case where we choose k_n to be the same order of m_n, the asymptotic distribution of $\widehat{\Psi}_n^{(b)}$ remains normal, but the asymptotic variance has a complicated expression. For more details, see Peng (2001).

Under Assumption 4, the estimation of α does not affect the asymptotics of $\widehat{\Psi}_n^{(b)}$. However, the use of $W_{(k_n)}$ in $\widehat{B}_n$ introduces a term in the influence function of $\widehat{\Psi}_n^{(b)}$, which affects the asymptotic variance. We define

$$
\begin{aligned}
\mathbf{\Upsilon}_{n,t} &= \left(W_{n,i}^* - \mathbb{E}W_{n,i}^*, \frac{n^{1/2}}{k_n^{1/2}}\left[I(W_i > l_n) - \mathbb{P}(W_i > l_n)\right]\right)', \\
\mathbf{T}_n &= \left(1, -\frac{1}{\alpha - 1}\frac{k_n^{1/2}}{n^{1/2}} l_n\right)'.
\end{aligned}
$$

Then, the asymptotic variance of the bias-corrected estimator is given by

$$
V_n = \mathbf{T}_n' \mathbf{\Omega}_n \mathbf{T}_n,
$$

where $\mathbf{\Omega}_n = \mathbb{E}\left[\mathbf{\Upsilon}_{n,t}\mathbf{\Upsilon}_{n,t}'\right]$.

Finally, we state the central limit theorem for $\widehat{\Psi}_n^{(b)}$ below. The proof is given in Appendix C.

Theorem 2.3 *Suppose Assumptions 1–3 and 4 hold, then we have*

$$
n^{1/2} V_n^{-1} \left(\widehat{\Psi}_n^{(b)} - \Psi\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).
$$

Remark 3 (A limitation) Our bias correction method would not work when only the first moment of $f(X)/g(X)$ exists. For example, if f is a Student-t distribution and f a normal density, then the highest order of finite moments of $f(X)/g(X)$ is 1. It means that $\alpha = 1$ if we use our tail assumption to approximate the tail of $f(X)/g(X)$. Since the denominator in our bias correction term involves $\widehat{\alpha} - 1$, the bias correction term does not exist in this case. More generally, our bias correction method would not work if one adopts a very "bad" choice of g such that only the first moment of $f(X)/g(X)$ exists.

Remark 4 Estimating α is not an easy problem. The convergence rate of the estimates of α can be slow because one only uses observations on the tails for the estimation. This can affect the bias corrected estimator because the estimates of α appear in this estimator. In the special case when X is a scalar and $W(X)$ is monotonic in X, we propose a simple alternative in Appendix B where we do not use the estimates of α to approximate the bias from the trimming.

Here we discuss the choice of the tuning parameters m_n and k_n. Since in our context n is the simulation size which the user controls, the exact choice of m_n and k_n is less important. Our baseline recommendation is to set $m_n = \sqrt{n}\log n$ and $k_n = \sqrt{n}$. This choice then satisfies Assumption 2 if $2\beta/(2\beta+\alpha) > 1/2$ and Assumption 4, and we can establish Theorem 2.3. This seems to work well in the Monte Carlo study and it gives reasonable results in our empirical illustration.

Below we discuss other alternatives. In general, the optimal—in the sense of minimizing root mean square errors of $\widehat{\alpha}$—data-driven procedure of choosing m_n is difficult. This is because the optimal choice depends critically on the second order tail behavior of W_i (the unknown parameter β in our case), which is hard to verify in practice.

In one strand of the literature, Hall (1990), Gomes and Oliviera (2001) and Danielsson et al. (2001) suggest bootstrap methods by further imposing certain strong conditions on the tails. In another strand of literature, Resnick (1997) and Resnick and Starica (1997) propose graphical tools—the Hill plot. Recently, Hill (2013) suggests similar methods for the weakly dependent case. The bootstrap methods are computational intensive and require stronger conditions on the tails, while the Hill plot is easy to construct but is more *ad hoc*. However, both methods are often used in the literature (e.g., McNeil, 1997; Drees, de Haan and Resnick, 2000; Danielsson et al., 2001).

Here we suggest an empirical rule to choose m_n based on Daouia, Florens and Simar (2010). More specifically, we locate a stable region in the alternative Hill plot of $(u, \widehat{\alpha}_u^{-1})$, $0 < u < 1$, and

$$\widehat{\alpha}_u^{-1} = (\lceil n^u \rceil)^{-1} \sum_{t=1}^{\lceil n^u \rceil} \left(\log W_{(t)} - \log W_{(n^u)}\right),$$

where $\lceil x \rceil$ returns the smallest integer no less than x.

This alternative Hill plot is defined in Resnick and Starica (1997) and has nice properties and advantages over the original Hill plot; see Resnick and Starica (1997) for details. To find the stable region, we calculate a sequence of estimates of α, namely, $\{\widehat{\alpha}_{1/L}, \widehat{\alpha}_{2/L}, \ldots, \widehat{\alpha}_1\}$.

Then, to measure whether the sequence of estimates has become stable, we compute the standard deviation of a rolling-window of estimates as follows:

$$\mathrm{sd}_i = \sqrt{\mathrm{Var}\left(\widehat{\alpha}_{i/L}, \ldots, \widehat{\alpha}_{\left(i+(0.05n)^{0.5}\right)/L}\right)}, \quad i = 1, \ldots, L - \sqrt{0.05n}.$$

We then locate the "least volatile region" as the index i such that the standard deviation attains the minimum, i.e., $\widehat{\imath}_{\min} = \mathrm{argmin}_i\{\mathrm{sd}_i\}$. Finally, we set $m_n = \left\lceil n^{\widehat{\imath}_{\min}/L} \right\rceil$.

For the choice of k_n, we suggest taking $k_n = m_n / \log n$. The reason is twofold. First, this choice is in line with Assumption 4. Second, a larger k_n implies smaller S_n and V_n, which in turn suggests a faster convergence rate of $\widehat{\Psi}_n^*$ and $\widehat{\Psi}_n^{(b)}$.

In the Monte Carlo experiments next section, we present results based on the baseline recommendation of setting $m_n = \sqrt{n} \log n$ and $k_n = \sqrt{n}$. We also perform a robustness check of the Monte Carlo results using the empirical rule discussed above, which are reported in Appendix D. It turns out that the results are not sensitive to how m_n and k_n are chosen.

3 Monte Carlo Experiments

In this section we investigate the properties of the proposed bias-corrected tail-trimmed estimator via a series of Monte Carlo experiments. Recall that the goal is to estimate the integral $\Psi = \int_{\mathcal{X}} H(x) f(x) \mathrm{d}x$ using importance sampling. Following Koopman, Shephard and Creal (2009), we assume that $f(x)$ is a Gaussian density and approximate it using various Gaussian densities with thinner tails.

More specifically, we set $f(x)$ to be $\mathcal{N}(0, 1)$ and take $H(x) = 1$. It is easy to see that $\Psi = \int_{-\infty}^{\infty} H(x) f(x) \mathrm{d}x = 1$. The importance sampling density $g(x)$ is chosen to be $\mathcal{N}(0, (1 + \epsilon)^{-1})$ with $\epsilon > 0$. Hence, the tails of $g(x)$ are thinner than those of the original density $f(x)$. When ϵ is sufficiently large, the variance of the importance sampling estimator becomes infinite. To show that, first note that the importance weight $W(x)$ is given by

$$W(x) = \frac{f(x)H(x)}{g(x)} = \frac{1}{\sqrt{1+\epsilon}} \mathrm{e}^{\frac{\epsilon}{2}x^2}.$$

It is easy to verify that $\mathbb{E}_g[W(X)] = 1$. Next, we compute the second moment of $W(x)$:

$$\begin{aligned}\mathbb{E}_g[W(X)^2] &= \int_{-\infty}^{\infty} \frac{1}{1+\epsilon} \mathrm{e}^{\epsilon x^2} \times \sqrt{\frac{1+\epsilon}{2\pi}} \mathrm{e}^{-\frac{1+\epsilon}{2}x^2} \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi(1+\epsilon)}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1-\epsilon}{2}x^2} \mathrm{d}x.\end{aligned}$$

It implies that $\mathbb{E}_g[W(X)^2] < \infty$ only if $\epsilon < 1$. When $\epsilon \geq 1$, the variance of the importance sampling estimator is infinite.

We consider three values of ϵ: 0.5, 1 and 3. Hence, only in the first case does the unmodified importance sampling estimator $\widehat{\Psi}_n$ have a finite variance. In the baseline case we set the tuning parameters for the bias-corrected tail-trimmed estimator as $m_n = \sqrt{n}\log n$ and $k_n = \sqrt{n}$. These tuning parameters satisfy Assumptions 2 and 4. We also perform a robustness check of these Monte Carlo results by using an empirical rule based on the alternative Hill plot to choose m_n and k_n as discussed in Section 2.3. The results are reported in Appendix D.

Figures 1-3 plot the sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ in (1) and the bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ in (8) under different settings. Each data point in the histograms consists of an estimate using a simulation size of $n = 10000$, and we use a total of 10000 independent estimates.

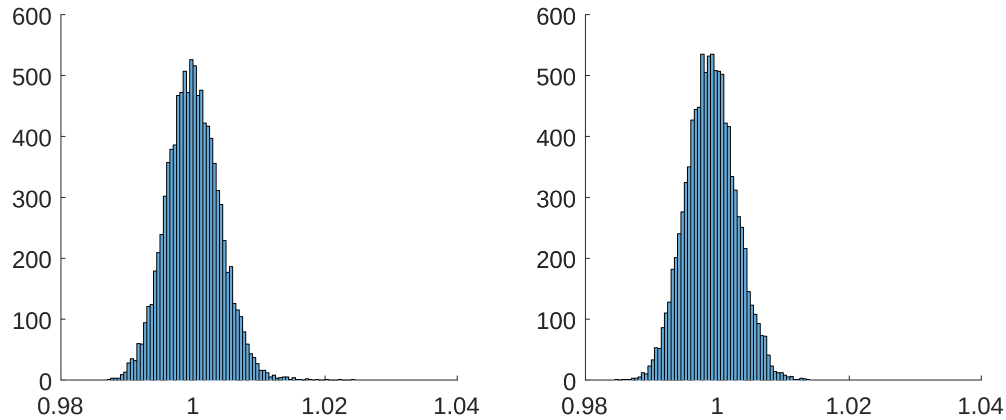

Figure 1: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 0.5$.

It is clear from Figure 1 that when the variance of $\widehat{\Psi}_n$ is finite and the central limit theorem applies, the sampling distributions of both estimators are approximately Gaussian and centered around the true value of 1. In fact, Table 1 below shows that the quantiles of the two distributions are essentially identical.

Figure 2 depicts the results for $\epsilon = 1$, when the variance of $\widehat{\Psi}_n$ is infinite and the central limit theorem fails to apply. The sampling distribution of $\widehat{\Psi}_n$ is noticeably more skewed with many large estimates, despite a substantial simulation size of $n = 10000$. In contrast, the sampling distribution of the tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ remains approximately Gaussian and is centered around the true value of 1. This is of course not surprising—the right panel of Figure 2 is simply an empirical verification of the central limit theorem proved in Theorem 2.3.

When $\epsilon = 3$, the tails of the importance sampling density $g(x)$ are substantially thinner than those of $f(x)$. Hence, it is of no surprise that the unmodified importance sampling estimator $\widehat{\Psi}_n$ performs very poorly, with the largest estimate being over 26 times larger than the true value. In fact, Table 1 shows that the 99th percentile of the estimates is 1.42. In other words, 1% of the estimates are over 42% larger than the true value.

In contrast, the proposed tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ performs reasonably well even in this severe setting. Although there seems to be a small finite-sample bias—the median estimate is 0.968 compared to the true value of 1—its sampling distribution remains approximately Gaussian and the mass of the density is concentrated between 0.9 and 1.05.

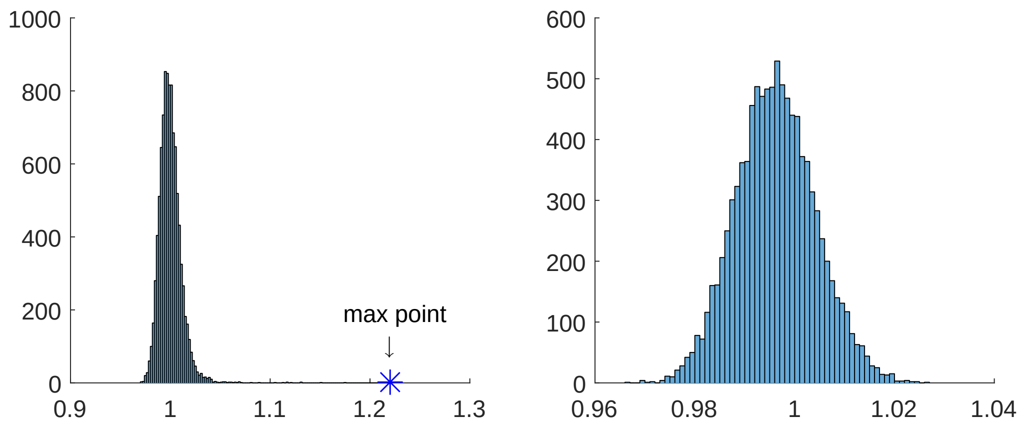

Figure 2: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 1$.

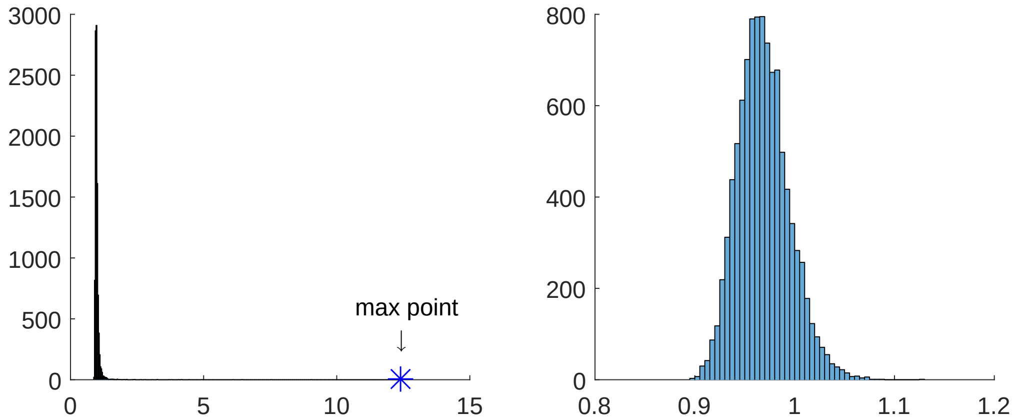

Figure 3: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 3$.

Table 1 also presents the quantiles of the two sampling distributions when the simulation size is increased to $n = 100000$. The basic conclusions remain the same: when the variance of $\widehat{\Psi}_n$ is finite, the properties of both estimators are essentially identical. When the variance of $\widehat{\Psi}_n$ is infinite, its sampling distribution can be severely skewed with many extremely large estimates, even when the simulation size is very large. On the other hand, the proposed tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ performs reasonably well with an approximate Gaussian sampling distribution.

Table 1: Quantiles of the sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$.

	$n = 10000$					$n = 100000$				
$\epsilon = 0.5$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.987	0.991	1.000	1.009	1.022	0.995	0.997	1.000	1.003	1.005
$\widehat{\Psi}_n^{(b)}$	0.985	0.991	1.000	1.007	1.017	0.995	0.997	1.000	1.003	1.004
$\epsilon = 1$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.971	0.979	0.999	1.032	1.340	0.989	0.993	1.000	1.011	1.090
$\widehat{\Psi}_n^{(b)}$	0.971	0.979	0.996	1.016	1.031	0.987	0.992	0.998	1.005	1.010
$\epsilon = 3$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.878	0.910	0.973	1.420	26.42	0.935	0.952	0.986	1.212	56.47
$\widehat{\Psi}_n^{(b)}$	0.881	0.915	0.968	1.039	1.098	0.947	0.959	0.982	1.009	1.031

Next, we study the empirical convergence rates of the two estimators under the settings $\epsilon = 1$ and $\epsilon = 3$. Since the mean squared error for $\widehat{\Psi}_n$ does not exist for these settings, we instead consider the interquartile range, defined as the difference between the 75th percentile and the 25th percentile. For a normal random variable with variance σ^2/n, its interquartile range is $2z_{0.75}\sqrt{\sigma^2/n}$, where $z_{0.75}$ is the 75th percentile of the standard normal distribution. Theorem 2.3 indicates that the convergence rate of the bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ depends of the tail parameter α and is typically slower than the usual root-n rate.

Figure 4 plots the interquartile range (in log) against the simulation size n (in log) for both estimators. In the case $\epsilon = 1$, the slopes corresponding to $\widehat{\Psi}_n$ and $\widehat{\Psi}_n^{(b)}$ are respectively -0.439 and -0.455—the convergence rate of the proposed estimator is slightly faster. When $\epsilon = 3$, however, the difference becomes much larger—the empirical convergence rates of $\widehat{\Psi}_n$ and $\widehat{\Psi}_n^{(b)}$ become $n^{0.274}$ and $n^{0.359}$, respectively. These values imply that if we want to halve the width of the interquartile range of $\widehat{\Psi}_n^{(b)}$, we need to increase the simulation size by 5.5 times (compared to 4 times in usual settings). For $\widehat{\Psi}_n$, however, we need to increase the simulation size by 7.3 times.

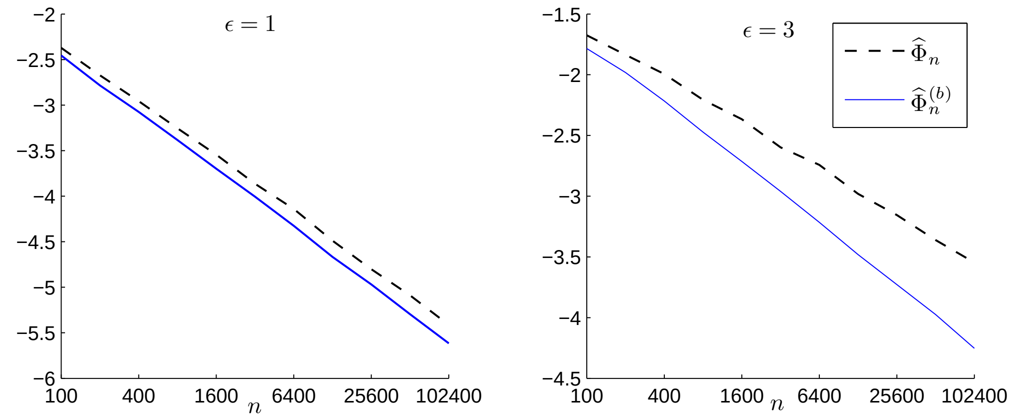

Figure 4: The interquartile range (in log) against the simulation size n (in log) for the importance sampling estimator $\widehat{\Psi}_n$ and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$.

4 An Illustration

In this section we illustrate the performance of the bias-corrected tail-trimmed estimator using an empirical application in which we fit a stochastic volatility model using daily returns on the Standard & Poor 500 (S&P 500) index. Stochastic volatility models are widely used to model financial returns. Notable early examples include Danielsson (1994), Durbin and Koopman (1997) and Shephard and Pitt (1997).

The daily returns on S&P 500 index are from September 2012 to August 2016. More specifically, let P_t denote the S&P 500 index on day t. We compute the return on day t as $y_t = 100\log(P_t/P_{t-1})$. The standard stochastic volatility model is given as

$$y_t = \mu + \epsilon_t^y, \qquad \epsilon_t^y \sim \mathcal{N}(0, \mathrm{e}^{h_t}), \tag{9}$$

$$h_t = \mu_h + \phi_h(h_{t-1} - \mu_h) + \epsilon_t^h, \qquad \epsilon_t^h \sim \mathcal{N}(0, \omega_h^2), \tag{10}$$

where the initial log volatility is initialized by $h_1 \sim \mathcal{N}(\mu_h, \omega_h^2/(1-\phi_h^2))$.

We estimate this model using Markov chain Monte Carlo methods (e.g., Kim, Shepherd and Chib, 1998). In particular, we obtain the posterior mean of the parameter vector $(\mu, \mu_h, \phi_h, \omega_h^2)$. We then evaluate the likelihood at this posterior mean by integrating out the log volatility $\mathbf{h} = (h_1, \ldots, h_T)'$ using importance sampling.

The importance sampling density is obtained by approximating the posterior distribution of $\mathbf{h}$ given the parameters using a Gaussian distribution as in Durbin and Koopman (1997). But instead of using the Kalman filter to obtain the importance sampling density, band matrix routines as discussed in Chan and Grant (2016) and Chan (2017) are used. Furthermore, independent draws from this Gaussian importance sampling density are obtained using the precision sampler in Chan and Jeliazkov (2009), which is more efficient than Kalman filter-based algorithms.

Figure 5 plots the sampling distributions of the importance sampling estimator $\log\widehat{\Psi}_n$ and the bias-corrected tail-trimmed estimator $\log\widehat{\Psi}_n^{(b)}$. Each data point in the histograms consists of an estimate using a simulation size of $n = 100000$.

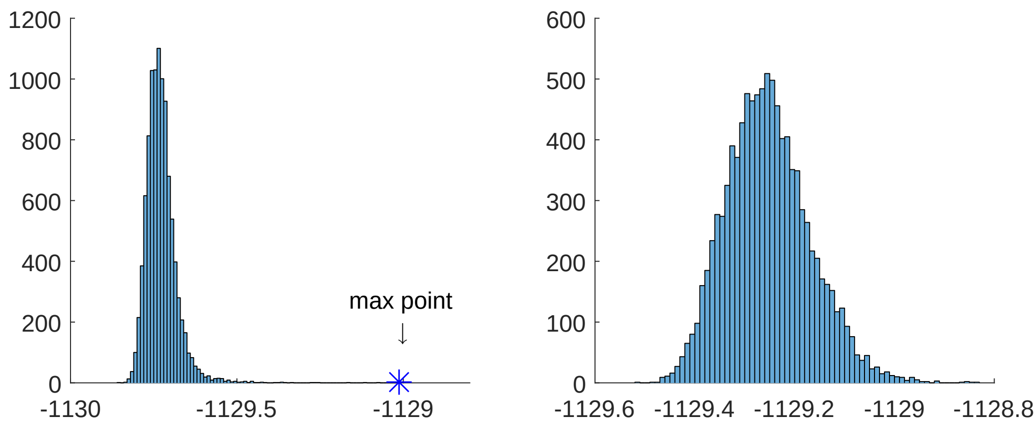

Figure 5: Sampling distribution of the importance sampling estimator $\log\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\log\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 100000$.

As is evident from the figure, the sampling distribution of $\log \widehat{\Psi}_n$ is substantially more skewed even though the simulation size is $n = 100000$. In contrast, the sampling distribution of the proposed estimator is less dispersed and remains approximately Gaussian.

Next, we plot in Figure 6 the sampling distribution of the ratio of the two estimators $\widehat{\Psi}_n^{(b)}/\widehat{\Psi}_n$. Here the same importance sampling weights are used to compute the two estimates. Again each data point in the histogram consists of an estimate using a simulation size of $n = 100000$.

The distribution is centered at around 1.6 with virtually no mass near unity, indicating that the $\widehat{\Psi}_n$ estimates tend to be substantially smaller than those of $\widehat{\Psi}_n^{(b)}$. This suggests that the original importance sampling estimator might underestimate the likelihood value Ψ.

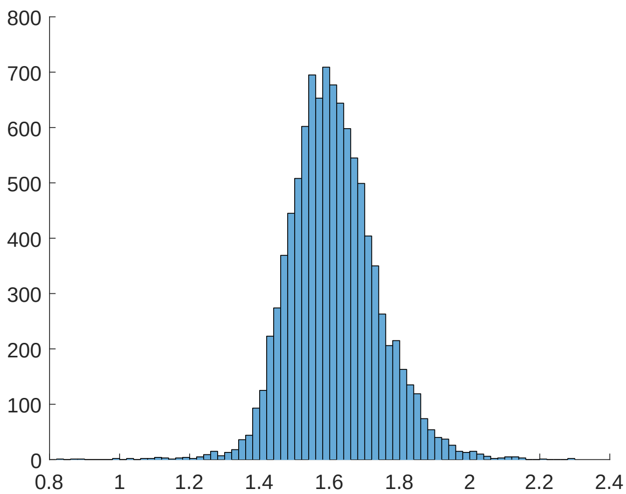

Figure 6: Sampling distribution of the ratio $\widehat{\Psi}_n^{(b)}/\widehat{\Psi}_n$ with a simulation size of $n = 100000$.

5 Concluding Remarks and Future Research

We have proposed a way forward when an importance sampling estimator has an infinite variance. Specifically, we have shown how one could modify the original infinite-variance estimator to obtain a consistent estimator that is asymptotically normal. The proposed estimator performed well in the Monte Carlo study as well as the empirical illustration.

In this paper we have only considered the case where independent samples are used to compute the importance sampling estimate. In many Bayesian applications, MCMC draws are used instead—and by construction these draws are correlated. Hence, for future work it would be worthwhile to first develop a similar test as Koopman, Shephard and Creal (2009) for the weakly dependent case. In addition, it would also be fruitful to consider asymptotic trimming in such a setting.

Appendix A Verification of Assumption 1′ for the Examples in Simulation Studies

A1 Univariate Case

For the univariate case, Assumption 1′ is equivalent to the following condition:[3]

$$\lim_{w\to\infty}\frac{\mathbb{P}\left(W_i > aw\right)}{\mathbb{P}\left(W_i > w\right)} = a^{-\alpha},$$

for $\alpha > 0$. For notational convenience, let $c_1 = \frac{1}{\sqrt{1+\epsilon}}$ and $c_2 = \frac{\epsilon}{2}$. Then,

$$\begin{aligned}
\mathbb{P}\left(W_i > aw\right) &= \mathbb{P}\left(c_1 e^{c_2 X_i^2} > aw\right) = 2\mathbb{P}\left(X_i > \sqrt{1/c_2\left(\log aw - \log c_1\right)}\right) \\
&= 2\int_{\sqrt{1/c_2(\log aw - \log c_1)}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(1+\epsilon\right)x^2}{2}\right)\mathrm{d}x.
\end{aligned}$$

By a similar calculation, we have

$$\mathbb{P}\left(W_i > w\right) = 2\int_{\sqrt{1/c_2(\log w - \log c_1)}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(1+\epsilon\right)x^2}{2}\right)\mathrm{d}x.$$

Hence, it follows that

$$\begin{aligned}
\lim_{w\to\infty}\frac{\mathbb{P}\left(W_i > aw\right)}{\mathbb{P}\left(W_i > w\right)} &= \lim_{w\to\infty}\frac{\int_{\sqrt{1/c_2(\log aw - \log c_1)}}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1+\epsilon)x^2}{2}\right)\mathrm{d}x}{\int_{\sqrt{1/c_2(\log w - \log c_1)}}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1+\epsilon)x^2}{2}\right)\mathrm{d}x} \\
&= \lim_{w\to\infty}\frac{\sqrt{1/c_2\left(\log w - \log c_1\right)}}{\sqrt{1/c_2\left(\log aw - \log c_1\right)}}\left(\frac{aw}{c_1}\right)^{-\frac{1+\epsilon}{2c_2}}\Bigg/\left(\frac{w}{c_1}\right)^{-\frac{1+\epsilon}{2c_2}} \\
&= a^{-\frac{1+\epsilon}{2c_2}} = a^{-\frac{1+\epsilon}{\epsilon}},
\end{aligned}$$

where we apply L'Hopital's Rule to obtain the second equality and we substitute in $c_2 = \frac{\epsilon}{2}$ to get the last equality. We have therefore shown that Assumption 1′ holds for the univariate example with $\alpha = \frac{1+\epsilon}{\epsilon}$.

[3]For this result, see Resnick (1987).

A2 Multivariate Case

We study the following simple multivariates case. Let

$$\begin{aligned} f(\mathbf{x}) &= (2\pi)^{-d/2}\exp\left(-\sum_{j=1}^{d}\frac{x_j^2}{2}\right), \\ g(\mathbf{x}) &= (2\pi)^{-d/2}\,\Pi_{j=1}^{d}\left(1+\epsilon_j\right)^{1/2}\exp\left(-\sum_{j=1}^{d}\frac{\left(1+\epsilon_j\right)x_j^2}{2}\right), \text{ and } H(x)=1, \end{aligned}$$

where d is a positive integer, $\epsilon_j > 0$ for all $1 \le j \le d$, and $\mathbf{x} = (x_1, x_2, \ldots, x_d)'$. To simplify analysis, we consider $d = 2$ for now. We first assume $\epsilon_1 > \epsilon_2 > 0$, and we deal with the case $\epsilon_1 = \epsilon_2$ later. First

$$W(x) = (1+\epsilon_1)^{-1/2}(1+\epsilon_2)^{-1/2}\exp\left(\frac{\epsilon_1 x_1^2}{2}+\frac{\epsilon_2 x_2^2}{2}\right).$$

Let $c_1 = (1+\epsilon_1)^{-1/2}(1+\epsilon_2)^{-1/2}$, $c_2 = (2\pi)^{-1}(1+\epsilon_1)^{1/2}(1+\epsilon_2)^{1/2}$. Then,

$$\mathbb{P}(W>w) = \int\int_{\epsilon_1 x_1^2+\epsilon_2 x_2^2 > 2\log\left(\frac{w}{c_1}\right)} c_2\exp\left(-\frac{1}{2}\left[(1+\epsilon_1)x_1^2+(1+\epsilon_2)x_2^2\right]\right)\mathrm{d}x_1\mathrm{d}x_2.$$

To calculate the above quantity, let

$$u = \epsilon_1 x_1^2+\epsilon_2 x_2^2, \qquad v = (1+\epsilon_1)x_1^2+(1+\epsilon_2)x_2^2. \tag{A-11}$$

Then, we have

$$\begin{aligned} \mathbb{P}(W>w) &= \int\int_{u>2\log\left(\frac{w}{c_1}\right),\,\frac{1+\epsilon_1}{\epsilon_1}u\le v\le\frac{1+\epsilon_2}{\epsilon_2}u} c_2\exp\left(-\frac{1}{2}v\right)\left|\frac{\partial(u,v)}{\partial(x_1,x_2)}\right|^{-1}\mathrm{d}v\mathrm{d}u, && \text{(A-12)} \\ &= \int\int_{2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{w}{c_1}\right)<v\le 2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{w}{c_1}\right),\,2\log\left(\frac{w}{c_1}\right)\le u\le\frac{\epsilon_1}{1+\epsilon_1}v} c_2\exp\left(-\frac{1}{2}v\right)\left|\frac{\partial(u,v)}{\partial(x_1,x_2)}\right|^{-1}\mathrm{d}u\mathrm{d}v \\ &\quad+\int\int_{v>2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{w}{c_1}\right),\,\frac{\epsilon_2}{1+\epsilon_2}v\le u\le\frac{\epsilon_1}{1+\epsilon_1}v} c_2\exp\left(-\frac{1}{2}v\right)\left|\frac{\partial(u,v)}{\partial(x_1,x_2)}\right|^{-1}\mathrm{d}u\mathrm{d}v && \text{(A-13)} \\ &\equiv \mathcal{P}_1(w)+\mathcal{P}_2(w). && \text{(A-14)} \end{aligned}$$

Solving equation (A-11), we obtain

$$\begin{aligned} x_1^2 &= \frac{(1+\epsilon_2)\,u-\epsilon_2 v}{\epsilon_1-\epsilon_2}, x_2^2 = \frac{(1+\epsilon_1)\,u-\epsilon_1 v}{\epsilon_2-\epsilon_1}, \\ x_1 x_2 &= \frac{\sqrt{[(1+\epsilon_2)\,u-\epsilon_2 v]\,[\epsilon_1 v-(1+\epsilon_1)\,u]}}{\epsilon_1-\epsilon_2}. \end{aligned}$$

Now we compute $\left|\frac{\partial(u,v)}{\partial(x_1,x_2)}\right|$,:

$$\begin{aligned} \left|\frac{\partial\,(u,v)}{\partial\,(x_1,x_2)}\right| &= \left|\begin{pmatrix} 2\epsilon_1 x_1 & 2\epsilon_2 x_2 \\ 2\,(1+\epsilon_1)\,x_1 & 2\,(1+\epsilon_2)\,x_2 \end{pmatrix}\right| = 4\,(\epsilon_1-\epsilon_2)\,x_1 x_2 \\ &= 4\sqrt{[(1+\epsilon_2)\,u-\epsilon_2 v]\,[\epsilon_1 v-(1+\epsilon_1)\,u]}. \end{aligned}$$

Substitute this into equation (A-12), we have

$$\begin{aligned} \mathcal{P}_1(w) &= \int_{2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{w}{c_1}\right)<v\leq 2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{w}{c_1}\right)} \int_{2\log\left(\frac{w}{c_1}\right)\leq u\leq \frac{\epsilon_1}{1+\epsilon_1}v} \frac{c_2\exp\left(-\frac{1}{2}v\right)}{4\sqrt{[(1+\epsilon_2)\,u-\epsilon_2 v]\,[\epsilon_1 v-(1+\epsilon_1)\,u]}}\mathrm{d}u\mathrm{d}v, \\ \mathcal{P}_2(w) &= \int_{v>2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{w}{c_1}\right)} \int_{\frac{\epsilon_2}{1+\epsilon_2}v\leq u\leq \frac{\epsilon_1}{1+\epsilon_1}v} \frac{c_2\exp\left(-\frac{1}{2}v\right)}{4\sqrt{[(1+\epsilon_2)\,u-\epsilon_2 v]\,[\epsilon_1 v-(1+\epsilon_1)\,u]}}\mathrm{d}u\mathrm{d}v. \end{aligned}$$

We assume for now that L'Hopital's Rule can be applied, and establish $\lim_{w\to\infty}\frac{\mathbb{P}(W_i>aw)}{\mathbb{P}(W_i>w)}$ as follows:

$$\lim_{w\to\infty}\frac{\mathbb{P}(W_i>aw)}{\mathbb{P}(W_i>w)} = \lim_{w\to\infty}\frac{\partial\mathbb{P}(W_i>aw)/\partial w}{\partial\mathbb{P}(W_i>w)/\partial w}. \tag{A-15}$$

First

$$
\begin{aligned}
\frac{\partial \mathbb{P}\left(W_i > aw\right)}{\partial w} &= \frac{\partial \mathcal{P}_1\left(aw\right)}{\partial w} + \frac{\partial \mathcal{P}_2\left(aw\right)}{\partial w} \\
&= \int\limits_{2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{aw}{c_1}\right) < v \leq 2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{aw}{c_1}\right)} \left\{ \frac{-1}{2w} \right. \\
&\qquad \left. \frac{c_2 \exp\left(-\frac{1}{2}v\right)}{4\sqrt{\left[\left(1+\epsilon_2\right)2\log\left(\frac{aw}{c_1}\right) - \epsilon_2 v\right]\left[\epsilon_1 v - \left(1+\epsilon_1\right)2\log\left(\frac{aw}{c_1}\right)\right]}} \right\} \mathrm{d}v, \\
&= \left(\int\limits_{2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{aw}{c_1}\right)+h(w) < v \leq 2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{aw}{c_1}\right)-h(w)} + \int\limits_{\text{The rest}} \right) \frac{-1}{2w} \\
&\qquad \left. \frac{c_2 \exp\left(-\frac{1}{2}v\right)}{4\sqrt{\left[\left(1+\epsilon_2\right)2\log\left(\frac{aw}{c_1}\right) - \epsilon_2 v\right]\left[\epsilon_1 v - \left(1+\epsilon_1\right)2\log\left(\frac{aw}{c_1}\right)\right]}} \right\} \mathrm{d}v, \\
&\equiv \mathcal{J}\left(aw\right) + \mathcal{R}\left(aw\right),
\end{aligned}
$$

where $h\left(w\right)$ is a function that only depends on w and both $h\left(w\right)$ and $h'\left(w\right)$ are very small such that

$$
\mathcal{J}\left(aw\right) + \mathcal{R}\left(aw\right) = \mathcal{J}\left(aw\right)\left(1 + o\left(1\right)\right),
$$

for any fixed $a > 1$. Such $h\left(w\right)$ can be e^{-w} for example.[4] Use this result,

$$
\lim_{w\to\infty} \frac{\partial \mathbb{P}\left(W_i > aw\right)/\partial w}{\partial \mathbb{P}\left(W_i > w\right)/\partial w} = \lim_{w\to\infty} \frac{\mathcal{J}\left(aw\right)\left(1+o\left(1\right)\right)}{\mathcal{J}\left(w\right)\left(1+o\left(1\right)\right)} = \lim_{w\to\infty} \frac{\mathcal{J}\left(aw\right)}{\mathcal{J}\left(w\right)}. \tag{A-16}
$$

[4] This is possible because we restrict it to a fixed $a > 1$, the term inside the integral is much bigger than ce^{-w} for any fixed $a > 1$ and a constant c, and

$$
\int\limits_{2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{aw}{c_1}\right) < v \leq 2\frac{1+\epsilon_2}{\epsilon_2}\log\left(\frac{aw}{c_1}\right)} \frac{1}{\sqrt{\left[\left(1+\epsilon_2\right)2\log\left(\frac{aw}{c_1}\right) - \epsilon_2 v\right]\left[\epsilon_1 v - \left(1+\epsilon_1\right)2\log\left(\frac{aw}{c_1}\right)\right]}} \mathrm{d}v
$$

is a constant.

Further

$$
\begin{aligned}
&\frac{\partial \mathcal{J}(aw)}{\partial w} \\
= &\frac{1}{2w}\left(2\frac{1+\epsilon_1}{\epsilon_1 w}+h'(w)\right)\frac{c_2\exp\left(-\frac{1}{2}\left[2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{aw}{c_1}\right)+h(w)\right]\right)}{4\sqrt{\left[(1+\epsilon_2)\,2\log\left(\frac{aw}{c_1}\right)-\epsilon_2\left(2\frac{1+\epsilon_1}{\epsilon_1}\log\left(\frac{aw}{c_1}\right)+h(w)\right)\right]\epsilon_1 h(w)}} \\
&-\frac{1}{2w}\left(2\frac{1+\epsilon_2}{\epsilon_2 w}-h'(w)\right)\frac{c_2\exp\left(-\frac{1}{2}\left[2\frac{1+\epsilon_2}{\epsilon 2}\log\left(\frac{aw}{c_1}\right)-h(w)\right]\right)}{4\sqrt{\epsilon_2 h(w)\left[\epsilon_1\left(2\frac{1+\epsilon_2}{\epsilon 2}\log\left(\frac{aw}{c_1}\right)-h(w)\right)-(1+\epsilon_1)\,2\log\left(\frac{aw}{c_1}\right)\right]}}.
\end{aligned}
$$

Use the above result and note that $\lim_{w\to\infty}\log(aw)/\log(w)=1$ and both $h(w)$ and $h'(w)$ are very small, we have

$$
\lim_{w\to\infty}\frac{\partial \mathcal{J}(aw)/\partial w}{\partial \mathcal{J}(w)/\partial w}=\lim_{w\to\infty}\frac{\frac{1+\epsilon_1}{\epsilon_1}\left(\frac{aw}{c_1}\right)^{-\frac{1+\epsilon_1}{\epsilon_1}}-\frac{1+\epsilon_2}{\epsilon_2}\left(\frac{aw}{c_1}\right)^{-\frac{1+\epsilon_2}{\epsilon_2}}}{\frac{1+\epsilon_1}{\epsilon_1}\left(\frac{w}{c_1}\right)^{-\frac{1+\epsilon_1}{\epsilon_1}}-\frac{1+\epsilon_2}{\epsilon_2}\left(\frac{w}{c_1}\right)^{-\frac{1+\epsilon_2}{\epsilon_2}}}=a^{-\frac{1+\epsilon_1}{\epsilon_1}}. \tag{A-17}
$$

Since $\lim_{w\to\infty}\frac{\partial\mathcal{J}(aw)/\partial w}{\partial\mathcal{J}(w)/\partial w}$ exits and both numerators and denominators in equations (A-15) and (A-16) go to zero, L'Hopital's Rule can be applied in equations (A-15) and (A-16). Equations (A-15) (A-16) and (A-17) show that

$$
\lim_{w\to\infty}\frac{\mathbb{P}(W_i>aw)}{\mathbb{P}(W_i>w)}=a^{-\frac{1+\epsilon_1}{\epsilon_1}}.
$$

For the case when $\epsilon_1=\epsilon_2=\epsilon>0$, we can show

$$
\lim_{w\to\infty}\frac{\mathbb{P}(W_i>aw)}{\mathbb{P}(W_i>w)}=a^{-\frac{1+\epsilon}{\epsilon}},
$$

where $\mathbb{P}(W_i>w)$ can be calculated by the following change of variables:

$$
u=x_1^2+x_2^2,\qquad v=x_1/x_2.
$$

For the general $d>2$ case, similar but much more tedious analysis leads to:

$$
\lim_{w\to\infty}\frac{\mathbb{P}(W_i>aw)}{\mathbb{P}(W_i>w)}=a^{-\frac{1+\max\left\{\epsilon_j\right\}_{j=1}^{d}}{\max\left\{\epsilon_j\right\}_{j=1}^{d}}},
$$

and Assumption 1′ is satisfied with $\alpha = \left(1 + \max \{\epsilon_j\}_{j=1}^d\right) \Big/ \max \{\epsilon_j\}_{j=1}^d$.

Appendix B A Simple Alternative of Estimating the Bias Term without Using $\widehat{\alpha}$

The convergence speed of $\widehat{\alpha}$ can be very slow, because we only use observations on tails for estimation. $\widehat{\alpha}$ may be volatile and this might lead to a volatile bias corrected estimator. In this section, we suppose X is a scalar and $W(X)$ is monotonic in X. In this case (though rather restricted), we propose a simple alternative where we do not use $\widehat{\alpha}$ for the bias correction.

The idea is to estimate the bias term B_n by another importance sampling estimator. Recall that $B_n = \mathbb{E}\left[W_i I\left(W_i \geq W_{(k_n)}\right)\right] = \int_{H\frac{f}{g} \geq W_{(k_n)}} H(x) f(x) \, \mathrm{d}x$. Then, we have

$$B_n = \mathbb{E}\left[H(X) \frac{f(X)}{g\left(X \mid W \geq W_{(k_n)}\right)}\right] \text{ and } X \sim g\left(x \mid W \geq W_{(k_n)}\right).$$

However, $g(x \mid W \geq W_{(k_n)})$ is very hard to generate unless we assume a condition like the following:

Assumption 5 *Suppose X is a scalar. $W(x)$ is strictly monotonic increasing for $x > c_{r_1} > 0$, and there exists a $c_{r_2} \geq c_{r_1}$ such that* $\sup_{x \leq c_r} W(x) < W(c_{r_2})$.

Under this assumption, $W_{(1)}, ..., W_{(k_n)}$ is the same as $W\left(X_{(1)}\right), ..., W\left(X_{(k_n)}\right)$, for any $k_n/n \to 0$, and

$$g\left(X \mid W \geq W_{(k_n)}\right) = g\left(X \mid X \geq X_{(k_n)}\right).$$

A truncated X distributed as $g\left(X \mid X \geq X_{(k_n)}\right)$ can be generated without much difficulty. We propose estimating B_n using

$$\tilde{B}_n = \frac{1}{n_B} \sum_{i=1}^{n_B} H(X_i) \frac{f(X_i)}{g\left(X_i | X_i \geq X_{(k_n)}\right)},$$

and $X_i \sim g\left(x \mid x \geq X_{(k_n)}\right)$, $n_B \to \infty$.

Remark 5 One may estimate B_n by a simpler estimator without Assumption 5:

$$\tilde{B}_n = \frac{1}{n_B} \sum_{i=1}^{n_B} W_i I\left(W_i \geq W_{(k_n)}\right), \tag{B-18}$$

where $n_B \gg n$. However, this estimator is extremely inefficient, because we use a tiny potion of data to estimate B_n.

Remark 6 Assumption 5 enables us to generate a truncated X to estimate B_n using all generated data. To serve the same purpose, one can impose similar assumptions such that we can easily generate X from $\left\{X \mid W(X) \geq W_{(k_n)}\right\}$. One can also impose similar yet weaker assumptions such that we can a get a region of X that contains $\left\{X \mid W(X) \geq W_{(k_n)}\right\}$ but is only slightly larger than $\left\{X \mid W(X) \geq W_{(k_n)}\right\}$. We will not lose much efficiency in this case. An extreme situation is discussed in Remark 5 where we do not restrict X at all and the resulting estimator is highly inefficient.

The following theorem shows that the alternative estimator is asymptotically equivalent to the original bias-corrected estimator.

Theorem B.1 *Suppose* $n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* - \mathbb{E}\widehat{\Psi}_n^*\right) \xrightarrow{d} \mathcal{N}(0,1)$ *and* $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = o_P(1)$. *Then, we have*

$$n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* + \tilde{B}_n - \Psi\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The conditions $n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* - \mathbb{E}\widehat{\Psi}_n^*\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ and $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = o_P(1)$ can be verified by the following low level assumptions. We note that, unlike the previous method which relies critically on Assumption $1'$ and $\widehat{\alpha}$, one can potentially come up many other low level assumptions such that this requirement is satisfied.

Lemma B.2 *Suppose Assumptions* $1'$ *and* 5 *hold and* $k_n \to \infty, n_B \to \infty, k_n/n \to 0$, $n_B/\max\left\{k_n^{0.5\alpha(\alpha-1)^{-1}}, n\right\} \to \infty$, *then we have* $n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* - \mathbb{E}\widehat{\Psi}_n^*\right) \xrightarrow{d} \mathcal{N}(0,1)$ *and* $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = o_P(1)$.

The proof is in Appendix C.

Remark 7 This alternative method does not work for the case when $\alpha = 1$, and it becomes quite demanding if α is close to 1. We need an estimate of α to guide the choice of n_B, but we do not need it directly to estimate B_n. In practice, one can let $n_B = \log(n)\max\left\{k_n^{0.5\widehat{\alpha}(\widehat{\alpha}-1)^{-1}}, n\right\}$. From the proof of the above lemma, we need to have a much bigger n_B, namely $n_B \Big/ \max\left\{nk_n^{0.5(2-\alpha)(\alpha-1)^{-1}}, n\right\} \to \infty$, if we adopt $\tilde{B}_n$ from equation (B-18).

Appendix C Main Proofs

Below we apply the results in Hill (2013) to prove Theorem 2.2 and Theorem 2.3.

Proof of Theorem 2.2: We apply the same reasoning and derivations of Theorem 1.2 in Hill (2013). Using the notations in his proof, replace α, Ψ, B_n with κ, $S, \alpha^{-1}\mathbb{E}[y_t I(y_t < -l_n)]$, respectively. Also note that our standard deviation is denoted by S_n while his is $\alpha^{-1}\mathcal{S}_n$.

Since Hill (2013) deals with the more general dependent case—whereas we focus on the independent case— Assumption N in Hill (2013) is automatically satisfied. Assumption D and Assumption T required in Hill (2013) can be verified by using our Assumption 3 and Assumption $1'$, respectively. *Q.E.D.*

Proof of Theorem 2.3: We again apply Theorem 2.1 in Hill (2013) by replacing α, Ψ, B_n with κ, $S, \alpha^{-1}\mathbb{E}[y_t I(y_t < -l_n)]$, respectively. Here our standard deviation is denoted by V_n, whereas his is $\alpha^{-1}\mathcal{V}_n$. By the same reasoning as above, Assumption N' in Hill (2013) is automatically satisfied. Assumption D, Assumption T' and (5)-(6) required in Hill (2013) can verified by using our Assumption 3, Assumption 1, Assumption 4, and Assumption 2, respectively. *Q.E.D.*

Proof of Lemma B.2: $n^{1/2}S_n^{-1}\left(\widehat{\Psi}_n^* - E\left(\widehat{\Psi}_n^*\right)\right) \xrightarrow{d} \mathcal{N}(0,1)$ is a direct result from previous theorems. We now prove $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = o_P(1)$

We choose n_B such that it goes to infinity much faster than n, i.e., $n_B/n \to \infty$. For a fixed n, the tail index of $H(X_i)\frac{f(X_i)}{g\left(X_i|X_i \geq X_{(k_n)}\right)}$ is also α, because we can treat $\mathbb{P}\left(X_i \geq X_{(k_n)}\right)$ as a constant for a fixed n. Given this,

$$\begin{aligned}
\tilde{B}_n &= \frac{1}{n_B}\sum_{i=1}^{n_B} H(X_i)\frac{f(X_i)}{g\left(X_i \,|\, X_i \geq X_{(k_n)}\right)} \\
&= \frac{1}{n_B/G^-\left(X_{(k_n)}\right)}\sum_{i=1}^{n_B/G^-\left(X_{(k_n)}\right)} H(X_i)\frac{f(X_i)}{g\left(X_i \,|\, X_i \geq X_{(k_n)}\right)} I\left(\tilde{X}_i \geq X_{(k_n)}\right),
\end{aligned}$$

where $\tilde{X}_i$ is distributed as g, those $\tilde{X}_i$ that $\tilde{X}_i \geq X_{(k_n)}$ are the same as X_i. The second equality holds by hypothetically filling in those untruncated observations. To get n_B dropped observations (larger than $X_{(k_n)}$), we need $n_B/G^-\left(X_{(k_n)}\right)$ observations of X_i

coming from the original g. The result from Linton and Xiao (2013) indicates[5] that

$$\tilde{B}_n - B_n = O_P\left(\left[n_B/G^-\left(X_{(k_n)}\right)\right]^{-(\alpha-1)/\alpha}\right).$$

Therefore $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = O_P\left(n^{1/2}S_n^{-1}\left[n_B/G^-\left(X_{(k_n)}\right)\right]^{(\alpha-1)/\alpha}\right)$ and $n^{1/2}S_n^{-1}\left(\tilde{B}_n - B_n\right) = o_P(1)$ if $n^{1/2}S_n^{-1}\left[n_B/G^-\left(X_{(k_n)}\right)\right]^{(\alpha-1)/\alpha} \to 0$. Since $k_n/n \approx G^-\left(X_{(k_n)}\right)$ ($\approx$ denotes equal by first order approximation), we need $n_B/k_n^{0.5\alpha(\alpha-1)^{-1}} \to \infty$ to ensure this condition holds. *Q.E.D.*

[5] Linton and Xiao (2013) show the expected shortfall in time series framework. This result can be applied to our iid case.

Appendix D Additional Results: Robustness Checks

In this appendix we redo the Monte Carlo study in Section 3 by using an empirical rule based on the alternative Hill plot to choose m_n and k_n discussed in Section 2.3. The sampling distributions of $\widehat{\Psi}_n$ and $\widehat{\Psi}_n^{(b)}$ for the cases $\epsilon = 0.5$, $\epsilon = 1$ and $\epsilon = 3$ are reported below. The results are very similar to those reported in Section 3 using the baseline choice of m_n and k_n.

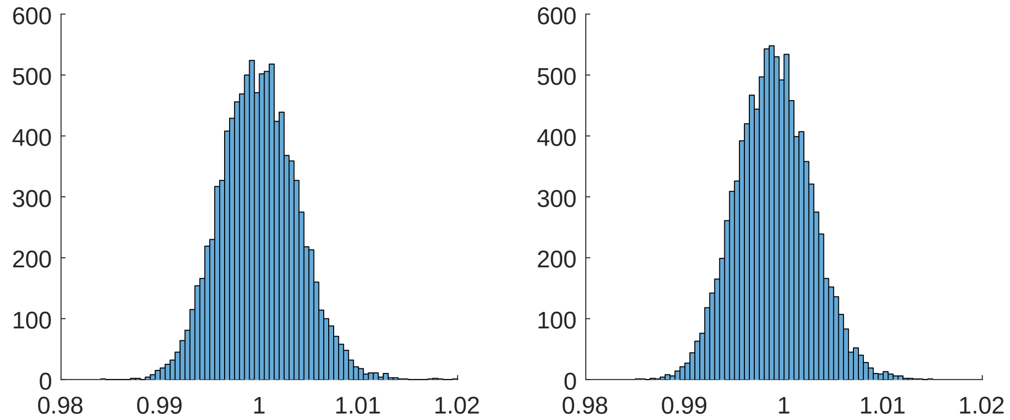

Figure 7: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 0.5$.

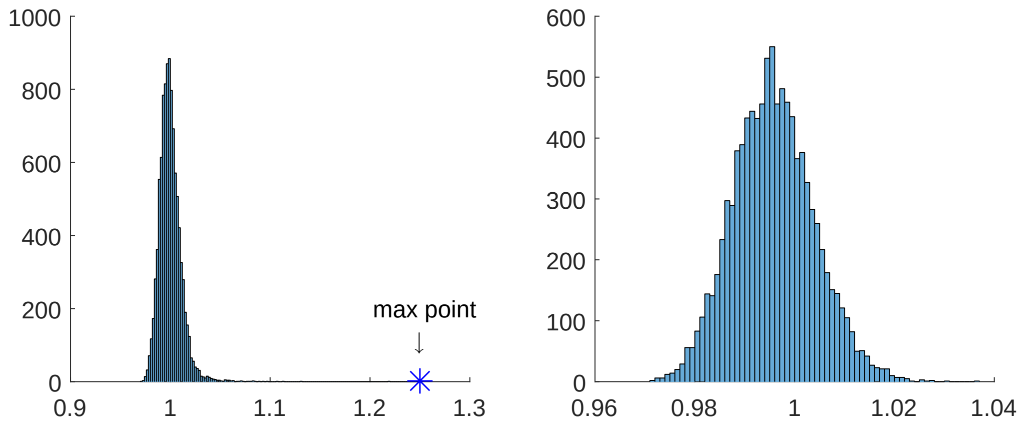

Figure 8: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 1$.

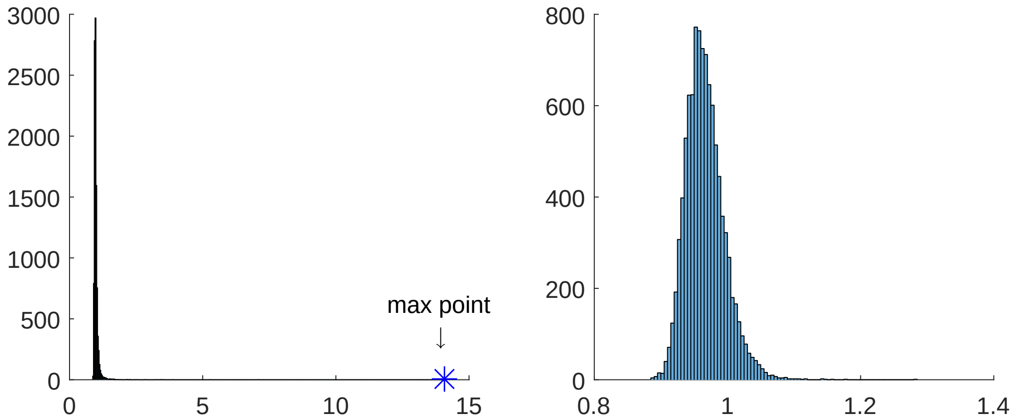

Figure 9: Sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ (left panel) and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$ (right panel) with a simulation size of $n = 10000$; $\epsilon = 3$.

Table 2: Quantiles of the sampling distributions of the importance sampling estimator $\widehat{\Psi}_n$ and the proposed bias-corrected tail-trimmed estimator $\widehat{\Psi}_n^{(b)}$.

	$n = 10000$					$n = 100000$				
$\epsilon = 0.5$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.984	0.991	1.000	1.009	1.020	0.996	0.997	1.000	1.003	1.007
$\widehat{\Psi}_n^{(b)}$	0.985	0.991	0.999	1.008	1.015	0.995	0.997	1.000	1.003	1.005
$\epsilon = 1$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.972	0.979	0.999	1.036	1.250	0.989	0.993	1.000	1.011	1.076
$\widehat{\Psi}_n^{(b)}$	0.971	0.978	0.996	1.016	1.032	0.990	0.993	0.999	1.007	1.012
$\epsilon = 3$	min	1%	50%	99%	max	min	1%	50%	99%	max
$\widehat{\Psi}_n$	0.883	0.908	0.973	1.358	14.05	0.935	0.952	0.985	1.207	13.23
$\widehat{\Psi}_n^{(b)}$	0.885	0.912	0.964	1.049	1.280	0.940	0.961	0.989	1.037	1.103

References

CHAN, J. C. C., (2017): "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling," *Journal of Business and Economic Statistics*, 35, 1, 17-28.

CHAN, J. C. C. AND E. EISENSTAT, (2015): "Marginal Likelihood Estimation with the Cross-Entropy Method," *Econometric Reviews*, 34, 3, 256-285.

CHAN, J. C. C. AND A. L. GRANT, (2016): "On the Observed-Data Deviance Information Criterion for Volatility Modeling," *Journal of Financial Econometrics*, 14, 4, 772-802.

CHAN, J. C. C. AND I. JELIAZKOV, (2009): "Efficient Simulation and Integrated Likelihood Estimation in State Space Models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 1/2, 101-120.

DANIELSSON, J., (1994): " Stochastic Volatility in Asset Prices Estimation with Simulated Maximum Likelihood," *Journal of Econometrics*, 64,1, 375-400.

DANIELSSON, J., L. DE HAAN, L. PENG, AND C. G. DE VRIES, (2001): " Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation," *Journal of Multivariate Analysis*, 76, 226-248.

DAOUIA, A., J. P. FLORENS AND L. SIMAR, (2010): " Frontier Estimation and Extreme Value Theory," *Bernoulli*, 16, 4, 1039- 1063.

DREES, H., L. DE HAAN, AND S. I. RESNICK, (2000): "How to Make a Hill Plot," *Annals of Statistics*, 28, 254-274.

DURBIN, J., AND S. J. KOOPMAN, (1997): "Monte Carlo Maximum Likelihood Estimation for NonGaussian State Space Models," *Biometrika*, 84, 669-684.

FRÜHWIRTH-SCHNATTER, S., (1995): "Bayesian Model Discrimination and Bayes Factors for Linear Gaussian State Space Models," *Journal of the Royal Statistical Society Series B*, 57, 237-246.

FRÜHWIRTH-SCHNATTER, S. AND H. WAGNER, (2008): "Marginal Likelihoods for Non-Gaussian Models Using Auxiliary Mixture Sampling," *Computational Statistics and Data Analysis*, 52, 4608-4624.

GEWEKE, J., (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317-1339.

Geweke, J., M. Keane and D. Runkle, (1994): "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, 609-632.

Gomes, M. I. and O. Oliviera, (2001): "The Bootstrap Methodology in Statistics of Extremes — Choice of the Optimal Sample Fraction," *Extremes,* 4, 331-358.

Haeusler, E., and J. L. Teugels, (1985): "On Asymptotic Normality of Hill's Estimator for the Exponent of Regular Variation," *Annals of Statistics*, 13, 743–756.

Hajivassiliou, V. A. and D. L. McFadden, (1998): "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica*, 66, 863-896.

Hall, P., (1990): "Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems," *Journal of Multivariate Analysis*, 32, 177-203.

Hill, B. M., (1975): "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics,* 3, 1163–1174.

Hill, J. B., (2010): "On Tail Index Estimation for Dependent Heterogeneous Data," *Econometric Theory*, 26, 1398–1436.

Hill, J. B., (2013): "Expected Shortfall Estimation and Gaussian Inference for Infinite Variance Time Series," *Journal of Financial Econometrics*, 13, 1, 1–44.

Hill, J. and A. Shneyerov, (2013): "Are There Common Values in First-Price Auctions? A Tail-Index Nonparametric Test," *Journal of Econometrics,* 174, 144-164.

Hoogerheide, L., A. Opschoor and H. K. Van Dijk, (2012): "A Class of Adaptive Importance Sampling Weighted EM Algorithms for Efficient and Robust Posterior and Predictive Simulation," *Journal of Econometrics,* 171(2), 101-120.

Kahn, H., and A. W. Marshall, (1953): "Methods of Reducing Sample Size in Monte Carlo Computations," *Journal of the Operational Research Society of America,* 1, 263–271.

Kim, S., N. Shepherd and S. Chib, (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies*, 65, 3, 361–393.

Kloek, T. and H. K. Van Dijk, (1978): "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1, 1-19.

Koopman, S. J. and E. Hol Uspensky, (2002): "The Stochastic Volatility in Mean Model: Empirical Evidence from International Stock Markets," *Journal of Applied Econometrics*, 17, 667-689.

Koopman, S. J., N. Shephard, and D. Creal, (2009): "Testing the Assumptions behind Importance Sample," *Journal of Econometrics*, 149, 2–11.

Kroese, D. P., T. Taimre, and Z. I. Botev, (2013): *Handbook of Monte Carlo Methods*, John Wiley & Sons, Hoboken.

Leland, W. E. M. S. Taqqu, W.Willinger, and D. V.Wilson, (1994): "On the Self-similar Nature of Ethernet Traffic (extended version)," *IEEE/ACM Trans. Networking*, 2-1, 1–15.

Linton, O. and Z. Xiao, (2013): "Estimation and Inference about the Expected Shortfall for Time Series with Infinite Variance," *Econometric Theory*, 29, 771-807.

Marshall, A. W., (1956): *The Use of Multi-Stage Sampling Schemes in Monte Carlo Computations*, in Meyer, M. (Ed.), Symposium on Monte Carlo Methods. Wiley, New York, pp. 123–140.

McNeil, A., (1997): "Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory," *Astin Bull*, 27, 125-139.

Monahan, J. F., (1993): *Testing the Behavior of Importance Sampling Weights*, in: Computer Science and Statistics: Proceedings of the 25th Annual Symposium on the Interface. 112–117.

Monahan, J. F., (2001): *Numerical Methods of Statistics*, Cambridge University Press, Cambridge.

Peng, L., (2001): "Estimating the Mean of a Heavy Tailed Distribution," *Statistics and Probability Letters*, 52, 255–264.

Pitt, M. K., R. Silva, P. Giordani, and R. Kohn, (2012): "On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter," *Journal of Econometrics*, 171, 134–151.

Resnick, S. I., (1987): Extreme Values, Regular Variation and Point Processes, SpringerVerlag: New York.

Resnick, S. I., (1997): "Discussion of the Danish Data on Large Fire Insurance Losses," *Astin Bull*, 27, 139-151.

Resnick, S. I. and C. Starica, (1997): “Smoothing the Hill Estimator,” *Advances in Applied Probability*, 29, 271-293.

Sandmann, G. and S. J. Koopman, (1998): “Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood,” *Journal of Econometrics*, 87, 271-301.

Shephard, N. and M. K. Pitt, (1997): “Likelihood Analysis of Non-Gaussian Measurement Time Series,” *Biometrika*, 84, 3, 653-667.

Tran, M. N., M. Scharth, M. K. Pitt, and R. Kohn, (2014): “Importance Sampling Squared for Bayesian Inference in Latent Variable Models.”