Pareto Smoothed Importance Sampling[*]

Aki Vehtari[†]   Andrew Gelman[‡]   Jonah Gabry[‡]

21 October 2017

Abstract

Importance weighting is a convenient general way to adjust for draws from the wrong distribution, but the resulting ratio estimate can be noisy when the importance weights have a heavy right tail, as routinely occurs when there are aspects of the target distribution not well captured by the approximating distribution. More stable estimates can be obtained by truncating the importance ratios. Here we present a new method for stabilizing importance weights using a generalized Pareto distribution fit to the upper tail of the distribution of the simulated importance ratios. The method includes stabilized effective sample estimates, Monte Carlo error estimates and convergence diagnostics.

Keywords: importance sampling, Monte Carlo, Bayesian computation

1. Introduction

Importance sampling is a simple procedure for computing expectations that is used when we can more easily obtain samples from some approximating distribution $g(\theta)$ than directly from the target distribution $p(\theta)$. Expectations with respect to the target distribution can be estimated by weighting the samples by the ratio of the densities. But when the approximating distribution is narrower than the target distribution—or, more generally, when the approximating distribution is a poor fit—the distribution of importance ratios can have a heavy right tail, which can lead to unstable importance weighted estimates, sometimes with infinite variance.

Ionides (2008) introduced a truncation scheme for importance ratios in which the truncation point depends on the number of simulation draws so that the resulting importance-weighted estimates have finite variance and are simulation consistent. In this paper we take the truncation scheme of Ionides and add to it the idea of fitting a generalized Pareto distribution to the right tail of the distribution of importance ratios. Our method, which we call Pareto smoothed importance sampling (PSIS), not only reduces mean square error relative to plain importance sampling (IS) and truncated importance sampling (TIS), but also provides improved Monte Carlo error estimates and natural diagnostics for gauging the reliability of the estimates. Based on extensive simulation studies, we also present empirical convergence rate results matching closely known theoretical results.

In this paper we focus on the case with a fixed proposal distribution, but the presented method can also be used as part of sequential, iterative and adaptive importance sampling (e.g. Kong et al., 1994; Owen and Zhou, 2000; Raftery and Bao, 2010; Bornn et al., 2010; Cornuet et al., 2012; Owen, 2013).

After presenting some background material in Section 2, we present our proposed method in Section 3, simulated examples in Section 4, and practical examples in Section 5, concluding in Section 6 with a brief discussion. Some of the content in Sections 2 and 3 is also presented

in Vehtari et al. (2017b) in relation to approximate leave-one-out cross-validation. Here we present it in more general form, supplemented with a significantly refined method, additional details and more general commentary not tied to any specific application of the algorithm.

2. Importance sampling

Suppose we want to estimate an integral

$$\mu = \mathrm{E}_p[h(\theta)] = \int h(\theta)p(\theta)d\theta, \tag{1}$$

where $h(\theta)$ is a function and $p(\theta)$ is a probability density. Given draws θ^s from $p(\theta)$, this can be approximated as

$$\hat{\mu} = \frac{1}{S}\sum_{s=1}^{S} h(\theta^s). \tag{2}$$

Often we cannot easily or cheaply draw directly from the *target distribution* $p(\theta)$, but there is an approximating *proposal distribution* $g(\theta)$ from which we can easily generate random draws. The integral (1) can be rewritten,

$$\int h(\theta)p(\theta) = \frac{\int [h(\theta)p(\theta)/g(\theta)]\, g(\theta)d\theta}{\int [p(\theta)/g(\theta)]\, g(\theta)d\theta}, \tag{3}$$

and it can then be estimated using S draws $\theta^1, \ldots, \theta^S$ from $g(\theta)$ by computing

$$\tilde{\mu} = \frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)r(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} r(\theta^s)}, \tag{4}$$

where the factors

$$r(\theta^s) = \frac{p(\theta^s)}{g(\theta^s)} \tag{5}$$

are called *importance ratios*. As θ are drawn from the proposal distribution $g(\theta)$, the ratios will have a distribution, too.

If p is a normalized probability density, the denominator of (3) is 1. However, in general p might only be known up to a normalizing constant, as is common in Bayesian inference where p might represent the posterior density of interest (with the dependence on data suppressed in our notation). It is therefore standard to use the ratio estimate (4) also known as self-normalized importance sampling, for which only the relative values of the importance ratios are needed. The self-normalized estimator induces a small bias of $O(1/n)$ but it remains consistent (Kong, 1992).

In Bayesian analysis, a proposal distribution g is often recommended based on simple approximation, for example, normal, split normal, or split-t fit at the mode (Geweke, 1989), or mixtures of multivariate normals, or approximate distributions obtained by variational inference or expectation propagation. One important application of importance sampling is for leave-one-out cross-validation, in which case the approximating distribution g is the full posterior and the target distribution p is the cross-validated posterior, excluding the likelihood for one observation (Gelfand et al., 1992; Gelfand, 1996), with the entire computation repeated for each data point, hence the need for quick computations.
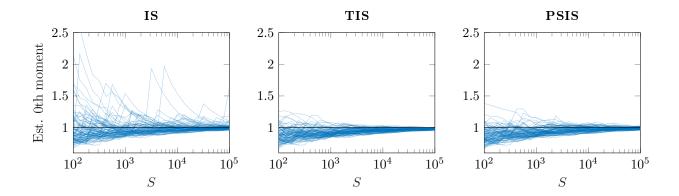
Figure 1: *Each graph has 100 blue lines, and each line is the trajectory of the estimated normalization term as the number of simulation draws increases. Target and proposal distributions are exponential distributions with rate parameters 1 and 1/3 respectively, chosen because (a) the distribution of the importance ratios has infinite variance, but (b) this would not generally be considered a pathological family of distributions.*

2.1. From importance ratios to regularized importance weights

Geweke (1989) shows for i.i.d. draws that if the variances of $r(\theta)$ and $h(\theta)r(\theta)$ are finite, then the central limit theorem holds for the convergence of the estimate in (4). Tierney (1994) extends the results to draws from from a uniformly ergodic Markov chain. Chen and Shao (2004) show further that the rate of convergence to normality is faster when higher moments exist. In simple restricted cases, the existence of the variance and higher moments can be checked analytically (Peruggia, 1997; Epifani et al., 2008; Robert and Casella, 2004; Pitt et al., 2013). However, we would like a generic diagnostic procedure to check the reliability of the importance sampling that works based on the observed importance ratios $r(\theta^s)$ alone, without requiring additional analysis of the distributions p and g.

If the variance of the distribution of importance ratios is very large or infinite, the estimate is still asymptotically consistent, but in general the simple importance weighted estimate with a finite number of draws cannot be trusted, and it makes sense to replace the importance ratios by more stable weights. Thus, (4) is replaced by

$$\tilde{\mu} = \frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)w(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} w(\theta^s)}, \tag{6}$$

where the *importance weights* w are some function of the importance ratios r from (5). Two extremes are $w \propto r$ (raw importance weighting) and $w \propto 1$ (identity weights, equivalent to just using the approximating distribution g).

Figure 1 illustrates the stability of estimating the 0th moment (normalization term) $\frac{1}{S}\sum_{s=1}^{S} w(\theta^s)$ in self-normalized importance sampling when the variance of the weights is infinite. Even if the estimate is still asymptotically consistent, plain importance sampling (IS) has big jumps in the estimates even with 10^5 draws. Truncated importance sampling (TIS) and Pareto smoothed importance sampling (PSIS) using stabilized weights, which are described later, produce more stable estimates.

Figure 2 shows the 100 largest weights from the same 100 simulations of importance weights given $S = 100$ or $S = 10000$ draws, illustrating that with a finite number of draws the empirical variance of the plain IS raw ratios is finite but greatly increasing as the number of draws increases.
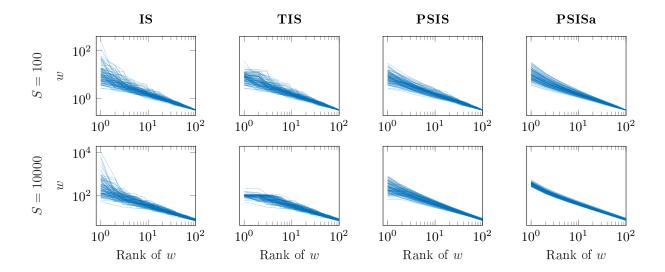
3

Figure 2: *Each graph has 100 blue lines, and each line is the largest weights sorted and plotted by rank. The top row shows the weights for $S = 100$ draws, and the bottom row shows only the 100 largest weights for $S = 10000$ draws (note the different scale of y-axis). Target and proposal distributions are exponential distributions with rate parameters 1 and 1/3 respectively, as in Figure 1.*

2.2. Truncated importance sampling

Truncated importance sampling is the same as standard importance sampling but using weights obtained by truncating the raw ratios. Ionides (2008) proposes a scheme in which the truncation point depends on the sample size S, and each individual weight w_s is obtained from the corresponding ratio r_s by taking

$$w_s = \min\left(r_s, \sqrt{S}\bar{r}\right), \tag{7}$$

where $\bar{r}$ is the average of the original S importance ratios. Ionides (2008) proves that the distribution of these weights is guaranteed to have finite variance and the estimator is asymptotically mean square consistent under weak conditions. Unfortunately, while this truncation method can greatly improve stability, as shown in Figure 1, it comes at the expense of downward bias in the largest weights, as illustrated in Figure 2. This causes bias both in the estimate and in the Monte Carlo error estimate as demonstrated in Sections 4 and 5.

2.3. Sample based diagnostic using the generalized Pareto distribution

Before introducing Pareto smoothing of the importance weights, we review and discuss the use of the generalized Pareto distribution for extreme value analysis.

Pickands (1975) proves that, if the unknown distribution function lies in the domain of attraction of some extremal distribution function, then, as the sample size increases and a threshold for the tail is allowed to increase, the upper tail of an unknown distribution is well approximated by the three-parameter generalized Pareto distribution,

$$p(y|u, \sigma, k) = \begin{cases} \frac{1}{\sigma}\left(1 + k\left(\frac{y-u}{\sigma}\right)\right)^{-\frac{1}{k}-1}, & k \neq 0 \\ \frac{1}{\sigma}\exp\left(\frac{y-u}{\sigma}\right), & k = 0, \end{cases} \tag{8}$$

where u is a lower bound parameter, y is restricted to the range (u, ∞), σ is a scale parameter, and k is a shape parameter. The generalized Pareto distribution has the property that when

$k > 0$ the number of existing moments is less than $\lfloor 1/k \rfloor$, and thus we can infer the number of existing moments of the weight distribution by focusing on k.

Pickands (1975) notes that "most 'textbook' continuous distribution functions" (Gumbel, 1958) lie in the domain of attraction of some extremal distribution function. But only for special cases can we identify the analytic form of the distribution of w_s, and verify that it belongs to that set of distributions referred by Pickands.

According to Koopman et al. (2009), Monahan (1993) uses Hill's (1975) tail-index estimator for k. Koopman et al. (2009) use a maximum likelihood estimate for (σ, k). In both cases a statistical hypothesis test is used to infer whether $k < 1/2$ and thus whether the underlying distribution has a finite variance.

We propose to use a Bayes-flavored method of Zhang and Stephens (2009) to estimate the parameters of the generalized Pareto distribution; this method has lower bias and higher efficiency than the maximum likelihood estimate. For this method, the distribution is reparameterized as (b, k), where $b = k/\sigma$. The parameters b and k are highly correlated and the likelihood is replaced by a profile likelihood where k is chosen to maximize the likelihood given b. The profile likelihood is combined with a weakly informative prior for b and the posterior mean $\hat{b}$ is computed numerically. Finally $\hat{k}$ is obtained by maximizing the likelihood given $\hat{b}$, and $\hat{\sigma}$ is set to $\hat{k}/\hat{b}$. Zhang and Stephens (2009) show that this estimate has a small bias, is highly efficient (having MSE close to the Cramér-Rao lower bound of the variance), and is both simple and fast to compute.

We compared Zhang and Stephens' (2009) method to a fully Bayesian approach with Markov chain Monte Carlo inference implemented in Stan (Stan Development Team, 2017), and did not observe any practical difference between $\hat{k}$ and the posterior mean given by Stan. It is also possible to get the approximate marginal posterior distribution of k from the intermediate results of Zhang and Stephens' method, which also matched accurately with the marginal posterior distribution obtained with Stan. To reduce the variance of the posterior for small S, we added an additional prior for k, corresponding approximately to weight of 10 observations from the tail shrinking towards 0.5 ($\hat{k} = (M\hat{k} + 10 \cdot 0.5)/(M + 10)$), where M is the sample size in the tail. Based on our simulation results, this reduced the variance and RMSE of the PSIS estimates with small S (S less than about 1000), without introducing significant bias or affecting the RMSE for larger S.

Use of the marginal posterior distribution of k would allow computing the probability $p(k < 1/2)$, that is the probability that the variance of weights is finite, but we have found it more useful to examine the continuous value of $\hat{k}$. Chen and Shao (2004) show that the convergence speed towards normality is higher with more moments, and thus continuous values of $\hat{k}$ are more informative than binarization at $\hat{k} < 1/2$. Furthermore, we show that we can get useful estimates also with $\hat{k} > 1/2$, although with decreasing convergence speed as $\hat{k}$ increases. In Section 4 we show several results of the practical convergence rates.

Pickands (1975) notes that to obtain asymptotic consistency, the threshold u should be chosen so that the sample size M in the tail increases to infinity while the proportion of simulation draws in the tail M/S goes to zero. There are many simple approaches which fulfill this asymptotic requirement, but choosing a good threshold in the finite sample case is non-trivial. With a lower threshold u, the variance of the fit is smaller but at the expense of potentially higher bias.

Pickands (1975) suggests to choose u by minimizing the absolute distance between empirical and generalized Pareto distribution fit, but we found this to be a very noisy approach. Koopman et al. (2009) set the lower bound u by graphing the maximum likelihood estimates of k along with confidence bands for a large number of thresholds, but we find this impractical for automated uses of importance sampling. Scarrot and MacDonald (2012) review tens of different approaches for selecting u, including various graphical diagnostics

of data, graphical diagnostics of the tail shape estimate given varying u (as used, e.g, by Koopman et al. (2009)), optimization of the fit of the tail estimate, bootstrapping, and mixture model approaches. Many of the reviewed approaches require human interpretation or intensive computation, which may be applicable in extreme value analysis for a single dataset but not feasible for automated importance sampling.

We tested several quick automated approaches, but found them to have high variance, increasing greatly the variance of the shape parameter estimate $\hat{k}$. Based on extensive computational experiments we chose u so that the sample size

$$M = \min\left(S/5, 3\sqrt{S}\right). \tag{9}$$

For $S \leq 225$, $M = S/5$ provides enough samples for practical estimation accuracy. For $S > 225$, $3\sqrt{S}$ satisfies the asymptotic consistency requirements. $\sqrt{S}$ has been used, for example, by Ferreira et al. (2003) and the multiplier 3 was chosen based on extensive simulation studies balancing the variance and bias. The majority of the results are not sensitive to the choice of M (as demonstrated by the performance of the fixed threshold $M = S/5$ used previously by Vehtari et al. (2017b)).

3. Pareto smoothed importance sampling

In addition of using the generalized Pareto distribution for diagnostics, we propose a novel importance sampling estimate that has the finite-variance property of truncated importance sampling while also reducing bias by replacing the largest weights with ordered statistics of the generalized Pareto distribution fitted to the upper tail of the weight distribution.

Smoothing weights using the generalized Pareto distribution. We stabilize the importance weights by replacing the M largest weights above the threshold u by the expected values of the order statistics of the fitted generalized Pareto distribution

$$F^{-1}\left(\frac{z - 1/2}{M}\right), \quad z = 1, \ldots, M,$$

where F^{-1} is the inverse-CDF of the generalized Pareto distribution. This reduces the variation in the largest weights, and thus typically reduces the variance of the importance sampling estimate. Figure 2 illustrates the effect of the Pareto smoothing of the weights: the variability among the simulations is greatly reduced by shrinking the weights towards the expected magnitude of the weight based on the generalized Pareto distribution fit. Compared to the simple truncation by Ionides (2008), the bias in PSIS is reduced as the largest weights are not all truncated to the same value, but rather spread according to the estimated tail shape. Some variability remains as the generalized Pareto distribution is fitted to a finite sample. The example used in Figure 2 has the special property that the whole importance ratio distribution is distributed as a generalized Pareto distribution and the rightmost plot illustrates the case of maximal Pareto smoothing with $M = S$. As discussed and demonstrated in Section 4, in some cases with increasing S there can be a fast change in the tail shape. To reduce the bias, instead of reducing M which would increase variance, we additionally truncate the smoothed weights to the largest raw ratio, which works well as demonstrated with different examples in Sections 4 and 5.

Summary of method. Given importance ratios $r_s, s = 1, \ldots, S$, our method proceeds as follows.

1. Set $M = \min\left(S/5, 3\sqrt{S}\right)$ and set u accordingly.

2. Fit the generalized Pareto distribution (8) to the sample consisting of the M highest importance ratios, with the lower bound parameter u as just chosen, estimating the parameters k and σ using the method from Zhang and Stephens (2009) and the additional prior mentioned in Section 2.3.

3. Replace the M highest importance ratios by the expected values of their order statistics of the generalized Pareto distribution given the estimated parameters from the previous step. The values below u are unchanged. We now refer to the S values as weights and label them $w_s, s = 1, \ldots, S$.

4. Truncate the weights at the maximum raw weight value $\max(r_s)$.

5. If the estimated shape parameter $\hat{k}$ exceeds 0.7 (see Section 3.1), report a warning that the resulting importance sampling estimates are likely to be unstable.

6. Report the estimated effective sample size S_{eff} and Monte Carlo error for the desired quantities as discussed in Section 3.3.

This method has been implemented in an R function called `psislw` which is included in the `loo` R package (Vehtari et al., 2017a). The package is available from CRAN and the source code can be found at https://github.com/stan-dev/loo. Python and Matlab/Octave implementations are available at https://github.com/avehtari/PSIS.

3.1. Diagnostics

Previous research has focused on identifying whether the variance of the raw ratios is finite or infinite (Peruggia, 1997; Epifani et al., 2008; Koopman et al., 2009), but we demonstrate in Sections 4 and 5 that it is more useful to look at the continuous $\hat{k}$ values than the discrete number of moments. Based on theory and simulation results (in Sections 4 and 5), we have

- If $k < \frac{1}{3}$ the Berry-Esseen theorem (Chen and Shao, 2004; Koopman et al., 2009) states faster convergence rate to normality. If $\hat{k} < \frac{1}{3}$, we observe that importance sampling is very stable and all IS, TIS and PSIS work well.

- If $k < \frac{1}{2}$ then the distribution of importance ratios has finite variance and the central limit theorem holds (Geweke, 1989). However with $k \to \frac{1}{2}$, the root mean square error (RMSE) of plain IS increases significantly while TIS and PSIS have lower RMSE.

- If $\frac{1}{2} \leq k < 1$ then the variance is infinite, but the mean exists. TIS and PSIS estimates have finite variance by accepting some bias. The convergence becomes slower with increasing k. If $0.5 \leq \hat{k} \lesssim 0.7$ we observe practically useful convergence rates and Monte Carlo error estimates with PSIS. The bias of TIS increases faster than the bias of PSIS. If $\hat{k} > 0.7$ we observe impractical convergence rates and unreliable Monte Carlo error estimates.

- If $k \geq 1$ then neither the variance nor the mean of raw ratios exists. The convergence rate is close to zero and bias can be large with practical sample sizes S.

If the method is implemented in a software package we recommend reporting to the user if $\hat{k} > 0.7$. Depending on the application and need for a faster convergence rate, a lower threshold could also be used.

In addition to examining $\hat{k}$ for the distribution of the ratios $r(\theta)$, it is useful to examine h specific diagnostics, that is, $\hat{k}$ for the distribution of $h(\theta)r(\theta)$. We follow Epifani et al. (2008), and if h can be equal to zero, then instead of $h(\theta)r(\theta)$ we use $\sqrt{1 + h(\theta)^2}r(\theta)$.

3.2. MCMC

In the case that we obtain the samples from the proposal distribution via Markov chain Monte Carlo, the S samples are dependent. The convergence of the importance sampling estimate holds for draws from a uniformly ergodic Markov chain (Tierney, 1994) and extreme value theory holds also for correlated sequences (see, e.g., Hill, 2010). We adjust the algorithm in the following way to take into account the smaller effective sample size of dependent MCMC draws. First, common convergence diagnostics are used to check that the MCMC draws are likely to be representative of the typical set (see, e.g., Gelman et al., 2013). Then we compute the effective sample size $S_{\text{eff,MCMC}}$ for $p(y_i|\theta)$ using the split-chain effective sample size estimate method (Gelman et al., 2013, Ch 11.). To make the following equations simpler we define relative efficiency as $R_{\text{eff,MCMC}} = S_{\text{eff,MCMC}}/S$. Due to correlated MCMC draws, we use more weights in the tail to keep the variance of $\hat{k}$ in control, and we adjust the number of tail weights used as $M = \min\left(S/5, 3\sqrt{S/R_{\text{eff,MCMC}}}\right)$.

3.3. Effective sample size and error estimates

We can also compute estimates for the Monte Carlo error and the effective sample size for importance sampling. In Sections 4 and 5 we demonstrate that these estimates are more accurate for PSIS than for IS and TIS.

Kong (1992) defines a measure of relative efficiency between sampling directly from $p(\theta)$ and sampling from $g(\theta)$ as

$$\frac{\text{Var}_g[\tilde{\mu}]}{\text{Var}_p[\hat{\mu}]} \approx 1 + \text{Var}_f[w]. \tag{10}$$

The approximation drops the dependency on h, which may introduce a substantial error, but it is a useful generic measure of the efficiency when there are many different potentially interesting target functions h. Kong et al. (1994) define effective sample size of importance sampling as

$$S_{\text{eff}} = \frac{S}{1 + \text{Var}_f[w]}, \tag{11}$$

which, after rearranging terms, can be written as

$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} w_s^2}. \tag{12}$$

As S_{eff} is based on the variance of weights, we may assume it to be unreliable when $\hat{k} > 1/2$. In Section 4, we demonstrate that when $\hat{k} > 1/2$, S_{eff} is highly unreliable for IS, strongly underestimated for TIS and useful for PSIS if $\hat{k} < 0.7$ (approximately).

Using the delta method used also by Kong (1992), it is possible to derive h specific Monte Carlo error estimates (Owen, 2013, Ch. 9) as

$$\widetilde{Var}(\tilde{\mu}) = \sum_{s=1}^{S} w_s^2 (h(\theta_s) - \tilde{\mu})^2. \tag{13}$$

If the draws have been obtained via MCMC we adjust the above error estimates by using the relative efficiency of the MCMC sample as

$$S_{\text{eff}} = \frac{R_{\text{eff,MCMC}}}{\sum_{s=1}^{S} w_s^2} \quad \text{and} \quad \widehat{Var}(\tilde{\mu}) = \sum_{s=1}^{S} w_s^2 (h(\theta_s) - \tilde{\mu})^2 / R_{\text{eff,MCMC}}. \tag{14}$$

Ionides (2008) proposes an Monte Carlo error estimate which also takes the bias into account. As the bias is estimated relative to plain IS, we found it to have too high variance for practical purposes, and the above Monte Carlo error estimate with Pareto smoothed weights has much smaller mean square error.

4. Simulated examples

In the following simulated examples we know the true target value and vary the proposal distribution. This allows us to study how $\hat{k}$ functions as a diagnostic and how bad the approximating distribution has to be before the importance sampling estimates break down. To diagnose the performance with respect to the number of draws S, in each of the examples we vary the number of draws from $S = 10^2$ to $S = 10^5$. We examine the estimates for the normalization term (0th moment), which also has a special role later in importance sampling examples, and estimates for $\text{E}[h(\theta)]$, where $h(\theta) = \theta$ (1st moment) or $h(\theta) = \theta^2$ (2nd moment). Although in these examples the normalization terms of both p and g are available, all experiments have been made assuming that normalization terms are unknown and self-normalized importance sampling is used.

4.1. Exponential target and proposal

In the first simulated example the target and proposal distributions are exponential distributions with rate parameter 1 and $1/\theta$, respectively. In this case, it is possible to compute the distribution of the importance ratios in closed form. The variance is infinite when $\theta > 2$ (Robert and Casella, 2004), and we know that the distribution of the importance ratios has the form of a generalized Pareto distribution.

Figure 1 shows 100 simulations of estimating the normalization term using a proposal distribution with rate parameter $1/3$, leading to $k \approx 0.66$ which illustrates the typical behavior of IS, TIS and PSIS when $0.5 < k < 0.7$. IS has high variability, while TIS and PSIS produce more stable estimates. For the same 100 simulations, Figure 2 shows the 100 largest IS raw ratios and modified weights from TIS, PSIS and PSISa. PSISa uses all the raw ratios, that is $M = S$, to fit the generalized Pareto distribution and estimate $\hat{k}$. IS has high variability, TIS truncates the largest weights downwards, while PSIS reduces the variability without biasing the largest weights. PSISa uses more samples to estimate $\hat{k}$ and has smaller variability than PSIS.

Figure 3 shows the bias and standard deviation of the 0th, 1st, and 2nd moment estimates with respect to the number of draws S computed from 1000 simulations. This illustrates the typical differences between the methods. IS (yellow) has the smallest bias but the largest deviation. PSISa (dashed blue) has the smallest deviation, and similar bias as PSIS. TIS (red) has the largest bias and with large S similar deviation as PSISa. PSIS (solid blue) has similar bias as PSISa, but slightly larger deviation than PSISa.

Figures 4, 5, and 6 show the RMSE and mean of the Monte Carlo error estimates for 0th, 1st and 2nd moment estimates with varying $\theta \in (1.3, 1.5, 2, 3, 4, 10)$. We can see that convergence rate depends *continuously* on k. There is no abrupt jump at $k = \frac{1}{2}$. The results for the 1st and 2nd moment agree well with the theoretical results by Epifani et al.
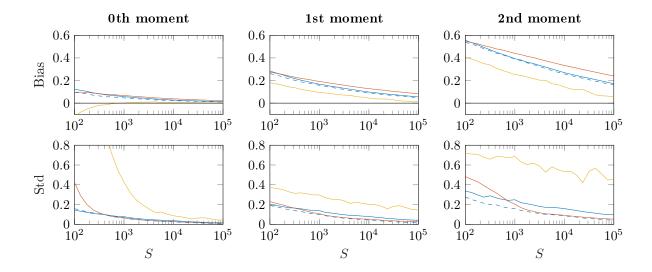
Figure 3: *Bias and standard deviation of the 0th, 1st, and 2nd moment estimates with respect to the number of draws S computed from 1000 simulations. Target and proposal distributions are exponential distributions with rate parameters 1 and 1/3 respectively, leading to k ≈ 0.66. IS is yellow, TIS is red, PSIS is blue and PSISa is blue dashed.*

(2008), that is, the conditions for the existence of the moments of the distribution of $r(\theta)$ and $h(\theta)r(\theta)$. We may assume that similar results would be obtained with $h(\theta)$ corresponding to higher moments. All methods have impractical convergence rates in the case of larger k. With smaller k values the errors are similar, but already at the borderline case $k = \frac{1}{2}$ IS has larger RMSE. TIS has larger RMSE than PSIS and PSISa for large k values. There is not much difference in RMSE between PSIS and PSISa, demonstrating the diminishing benefit of using more samples to estimate k. The Monte Carlo error estimates are unbiased for IS, but have very high variance for larger k (not shown here). The Monte Carlo error estimates are unbiased for TIS when $k < \frac{1}{2}$, but the error is underestimated for $k > \frac{1}{2}$. The Monte Carlo estimates for PSIS and PSISa are useful for $k \lesssim 0.7$, with PSISa yielding slightly more accurate error estimates. The results from PSISa show that using larger M (in PSISa $M = S$) would give better estimates, but this works only if the M largest weights are well approximated by the generalized Pareto distribution. For easier comparison of RMSEs between IS, TIS and PSIS, Figure 21 in Appendix A shows the relative RMSEs.

Figure 7 shows how the practical convergence rate depends on h specific $\hat{k}$ estimates (see Section 3.1) . For direct independent draws from $p(\theta)$ the variance of the Monte Carlo would decrease as S^{-1}. For PSIS we observe convergence rates $S^{-\alpha}$, where $0 < \alpha \leq 1$ depends on k and can be estimated from $\hat{k}$. Here $\hat{k}$ is estimated using the mean from 1000 simulations, with $S = 10^5$ for each simulation. Convergence rates are estimated by a linear fit to the RMSE results illustrated in Figures 4, 5, 6. Previously in the literature the focus has been mostly on the binary decision between $k < \frac{1}{2}$ and $k \geq \frac{1}{2}$, although the Berry-Esseen theorem states that the convergence rate to normality is faster with decreasing k. Here we can see that in practice the convergence rate starts to decrease for PSIS when $k > \frac{1}{4}$ and for PSISa when $k > \frac{1}{3}$. In this and the other experiments in the paper the convergence rate at $k = \frac{1}{2}$ is approximately with $\alpha \approx 0.87$, and at $k = 0.7$ is approximately with $\alpha \approx 0.6$ (marked in the plots). In this experiment we added results for $\theta \in (1.9, 2.1)$ which have k just below and above 0.5, illustrating that there is no sharp transition. A dashed line has been drawn from $k = 0.5, \alpha = 1$ to $k = 0, \alpha = 0$, which matches well the observed behavior. We assume that for the 1st and 2nd moment estimates the h specific $\hat{k}$ estimates are not as good as $\hat{k}$
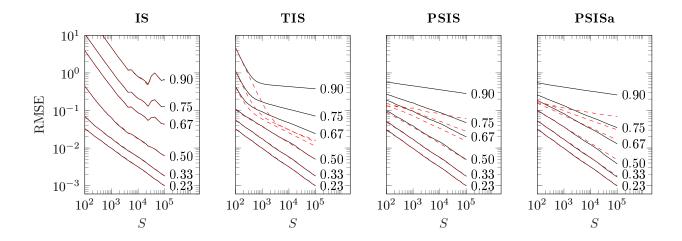
Figure 4: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. Target and proposal distributions are exponential distributions with rate parameters 1 and $1/\theta$ respectively, with $\theta \in (1.3, 1.5, 2, 3, 4, 10)$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*
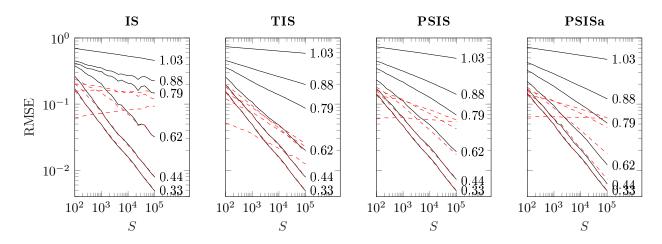

Figure 5: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. Target and proposal distributions are exponential distributions with rate parameters 1 and $1/\theta$ respectively, with $\theta \in (1.3, 1.5, 2, 3, 4, 10)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*

estimates for weights (0th moment), which would explain why the corresponding convergence rate curves go slightly beyond this limit line. PSISa is able to produce better convergence rates, but the difference is small and PSISa using $M = S$ is not usually applicable.

4.2. Univariate normal and Student's t

The previous example with exponential target and proposal is especially suited for PSIS, as the whole importance ratio distribution is well fitted with the generalized Pareto distribution. In this section we consider various univariate target and proposal distribution combinations to show the behavior in the case of different target-proposal tail combinations. In the next section we examine the corresponding multivariate cases.
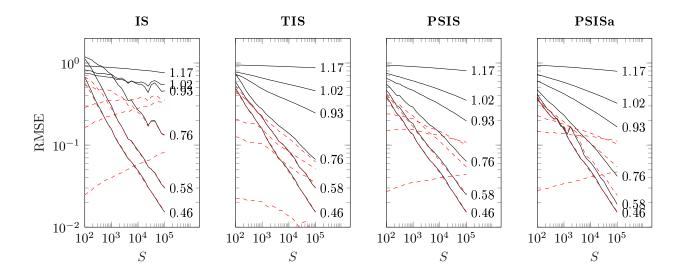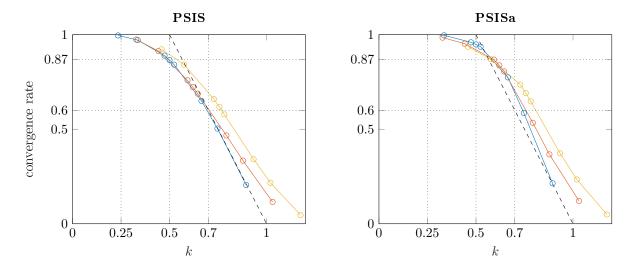
Figure 6: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 2nd moment estimate. Target and proposal distributions are exponential distributions with rate parameters 1 and $1/\theta$ respectively, with $\theta \in (1.3, 1.5, 2, 3, 4, 10)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*

Figure 7: *The practical relative convergence rates for PSIS and PSISa estimating 0th (blue), 1st (red) and 2nd (yellow) moments. Target and proposal distributions are exponential distributions with rate parameters 1 and $1/\theta$ respectively, with $\theta \in (1.3, 1.5, 1.9, 2, 2.1, 3, 4, 10)$. The circles in each line correspond to the results with different θ, with higher θ values leading to lower convergence rates and higher $\hat{k}$ values. $\hat{k}$ estimates in case of 1st and 2nd moments are h specific. Dotted horizontal lines at 0.87 and 0.6 show the typical convergence rates at $\hat{k} = 0.5$ and $\hat{k} = 0.7$ for several different experiments in this paper.*

We do not use the simulated example by Ionides (2008) having the target $p(\theta) = N(\theta \,|\, 0, 1)$ and the proposal $g(\theta) = N(\theta \,|\, 0, \sigma)$, as both distributions have the same mean and thus when estimating the 1st moment the lowest RMSE would be obtained by using identity weights and $\sigma \to 0$.

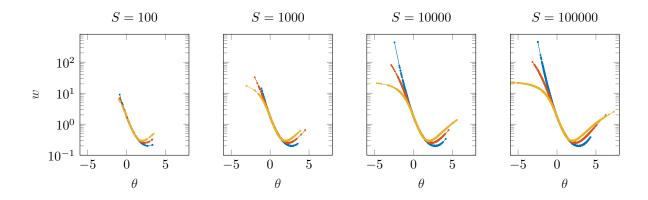To test the performance for the 0th, 1st and 2nd moments we choose the proposal

Figure 8: *Plain IS weights plotted by θ for different target-proposal pairs: $p(\theta) = \mathrm{N}(0,1)$, $g(\theta) = \mathrm{N}(1.5, 0.8)$ (blue), $p(\theta) = t_{20}(0,1)$, $g(\theta) = t_{21}(1.5, 0.8)$ (red), $p(\theta) = t_7(0,1)$, $g(\theta) = t_8(1.5, 0.8)$ (yellow), and different sample sizes S.*

distributions to have different mean and scale than the target distribution. The tested pairs are

1. $p(\theta) = \mathrm{N}(0,1)$, $g(\theta) = \mathrm{N}(\mu, 0.8)$: This is a special case with the matching tail shape of the target and proposal leading to a case which is favorable for PSIS.

2. $p(\theta) = t_{20}(0,1)$, $g(\theta) = \mathrm{N}(\mu, 0.9)$: This resembles an applied case of using a Gaussian posterior approximation when the target has a thicker tail.

3. $p(\theta) = t_{20}(0,1)$, $g(\theta) = t_{21}(\mu, 0.8)$: This resembles an applied case of leave-one-out importance sampling where the proposal has a slightly thinner tail.

4. $p(\theta) = t_7(0,1)$, $g(\theta) = t_8(\mu, 0.8)$: This resembles an applied case of leave-one-out importance sampling where the proposal has a slightly thinner tail, but both having thicker tails than in case 3.

Here we have left out easy examples where the proposal would have a thicker tail than the target. To vary how well the proposal matches the target, the mean of the proposal μ is varied.

Figure 8 shows plain IS weights for different θ, different target-proposal pairs (1, 3 and 4 in the above list with $\mu = 1.5$), and different sample sizes S. With increasing sample size S we get more draws from the tails and the differences between the weight functions become more apparent. To better illustrate how the tail shape of the empirical weight distributions change when the sample size S increases, Figure 9 shows the same weights sorted and plotted by rank. With $S = 100$ the weight distributions look very similar and the corresponding $\hat{k}$'s are 0.66, 0.66, and 0.64. As S increases the distributions of the weights eventually look very different and the corresponding $\hat{k}$'s are 0.66, 0.47, and -0.28. This shows that using only a small portion of the raw weights in the tail allows the $\hat{k}$ diagnostic to adapt to the empirically observed tail shape. Figure 9 also illustrates the motivation to truncate the weights at the maximum raw weight value $\max(r_s)$. This will allow the use of larger M for the Pareto fit to reduce the variance, while being able to adapt when the magnitude of the weights is saturating with increasing S.

Based on the above $\hat{k}$ values we may assume that PSIS is beneficial for small S, and for cases 3 and 4 when S grows eventually plain IS will also work well. Figure 10 shows the mean RMSE from 1000 simulations for the four different target-proposal pairs as listed above with hand-picked μ values to illustrate the typical behavior of IS, TIS and PSIS when
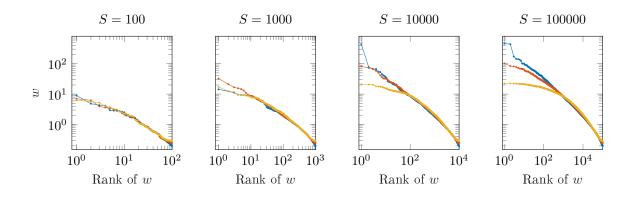
Figure 9: *Sorted plain IS weights plotted by rank for different target-proposal pairs: $p(\theta) = \mathrm{N}(0,1)$, $g(\theta) = \mathrm{N}(1.5, 0.8)$ (blue), $p(\theta) = t_{20}(0,1)$, $g(\theta) = t_{21}(1.5, 0.8)$ (red), $p(\theta) = t_7(0,1)$, $g(\theta) = t_8(1.5, 0.8)$ (yellow), and different sample sizes S.*
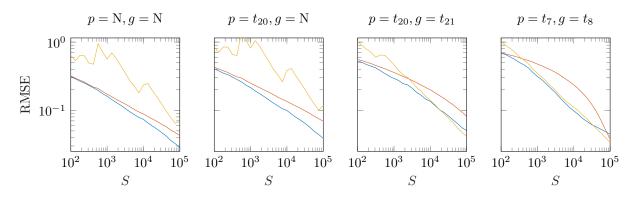

Figure 10: *The mean RMSE from thousand simulations for the different target-proposal pairs with μ from left to right being 1.5, 2, 2.25 and 3. IS is yellow, TIS is red and PSIS is blue.*

$0.5 \geq \hat{k} < 0.7$ (for very small $\hat{k}$ there are no differences and for very high values all methods fail). PSIS is able to adapt well in all cases and has the smallest RMSE in almost all cases. When the proposal distribution has a thin tail (e.g. the normal in cases 1 and 2), IS has high variance and the variance remains high with increasing S. If the tail of the proposal is thick (e.g. Student's t) and asymptotically the weight distribution has a short tail (small k), PSIS performs better than IS for small S. Eventually the RMSE of IS can get as small as the RMSE of PSIS. Most of the time TIS has a larger RMSE than PSIS. For thick tailed proposals TIS is not able to adapt well to the changing shape of the empirical weight distribution.

Figure 11 shows as a representative example of the RMSE and Monte Carlo error estimates for 0th moment estimated with IS, TIS and PSIS in case of $p(\theta) = t_{20}(0,1)$, $g(\theta) = N(\mu, 0.9)$. PSIS is more stable, has smaller RMSE than IS, and has more accurate Monte Carlo error estimates than TIS. Figures 22–33 in Appendix A show the corresponding information for 0th, 1st, and 2nd moments and for all four target-proposal pairs listed above. Figures 34–37 in Appendix A show the corresponding relative differences in RMSE between IS, TIS and PSIS.

Figure 12 shows the practical convergence rate with respect to h specific $\hat{k}$ estimates. We observe similar behavior as in in the exponential distribution example. For thick tail proposal distributions we tend to overestimate $\hat{k}$ values more than for the thin tailed proposal distributions, but the convergence rates stay good. Given estimated $\hat{k}$ values, we can use the dashed diagonal line drawn from $k = 0.5, \alpha = 1$ to $k = 0, \alpha = 0$ to provide a conservative convergence rate estimate.
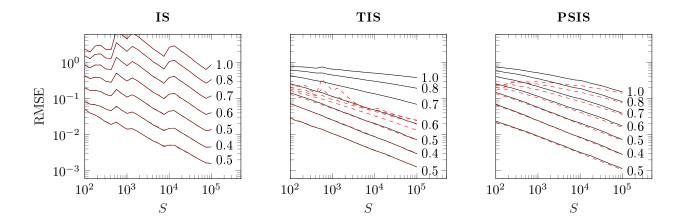
14

Figure 11: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_{20}(0, 1)$ and the proposal distribution is $N(\mu, 0.9)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*

Figure 12: *The practical relative convergence rates for PSIS estimating 0th (blue), 1st (red) and 2nd (yellow) moments. Different target-proposal distribution pairs are used in different subplots. $\hat{k}$ estimates in case of 1st and 2nd moments are h specific. Dotted horizontal lines at 0.87 and 0.6 show the typical convergence rates at $\hat{k} = 0.5$ and $\hat{k} = 0.7$ for several different experiments in this paper.*

4.3. Multivariate normal and Student's t

In this section we consider the isotropic multivariate versions of the four target-proposal pairs as the number of dimensions increases. In addition, we also examine a case where the proposal has thicker tails than the target, and show that with an increasing number dimensions this will also lead to increasing variability of the weights. The compared target-proposal pairs are

1. $p(\theta) = N(\mathbf{0}, I)$, $g(\theta) = N(0.4 \cdot \mathbf{1}, 0.8I)$

2. $p(\theta) = t_{20}(\mathbf{0}, I)$, $g(\theta) = N(0.4 \cdot \mathbf{1}, 0.9I)$

3. $p(\theta) = t_{20}(\mathbf{0}, I)$, $g(\theta) = t_{21}(0.4 \cdot \mathbf{1}, 0.8I)$

4. $p(\theta) = t_7(\mathbf{0}, I)$, $g(\theta) = t_8(0.4 \cdot \mathbf{1}, 0.8I)$

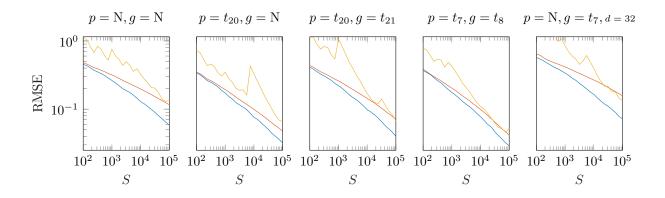5. $p(\theta) = N(\mathbf{0}, I)$, $g(\theta) = t_7(0.4 \cdot \mathbf{1}, 0.8I)$

15

Figure 13: *The mean RMSE from 1000 simulations for the different target-proposal pairs with $D = 16$ for all except the last one with $D = 32$. IS is yellow, TIS is red and PSIS is blue line.*

In all of these cases, the proposal distribution is just slightly displaced and with slightly narrower scale. The 5th proposal distribution has thicker tails than the target distribution, and we can then assume asymptotically finite variance. The number of dimensions is varied as $D \in (1, 2, 4, 8, 16, 32, 64)$.

Figure 13 shows the mean RMSE from 1000 simulations for the five different target-proposal pairs as listed above with $D = 16$ for all except the last case with $D = 32$. The plots illustrate the typical behavior of IS, TIS and PSIS when $0.5 \le \hat{k} < 0.7$ (for very small $\hat{k}$ there are no differences and for very high values all methods fail). PSIS has the smallest RMSE in all cases. TIS has a slower convergence rate than PSIS. IS is overall more unstable and has higher RMSE. Comparing the case with a thick tailed proposal distribution to the corresponding univariate case we see that IS requires larger S before beginning to stabilize. The rightmost subplot illustrates that even if the proposal distribution has thicker tails than the target distribution, the variability of the importance ratio distribution increases with the number of dimensions D.

Figure 14 shows a representative example of the RMSE and Monte Carlo estimates for 0th moment estimated with IS, TIS and PSIS with $p(\theta) = t_{20}(\mathbf{0}, I)$, $g(\theta) = N(0.4 \cdot \mathbf{1}, 0.9I)$. PSIS is more stable, has smaller RMSE than IS and TIS, and has more accurate Monte Carlo estimates than TIS. All methods eventually fail as the number of dimensions increases and even a small difference in the distributions is amplified. Figures 38–45 in Appendix A show the corresponding information for 0th and 1st moments, and for the first four target-proposal pairs listed above. Figures 46–49 in Appendix A show the corresponding relative differences in RMSE between IS, TIS and PSIS. In these multivariate examples, we observe sudden large jumps also for TIS. Truncation in TIS fails when there is one extremely large weight that causes the truncation level to rise so high that other large weights are not truncated. PSIS performs better in the same situation as one extreme large weight doesn't affect the generalized Pareto fit as much.

Figure 15 shows the practical convergence rate with respect to the h specific $\hat{k}$ estimates. We observe similar behavior as in the univariate example. As we saw before, the observed convergence rate with respect to $\hat{k}$ follows quite well the dashed diagonal line drawn from $k = 0.5, \alpha = 1$ to $k = 0, \alpha = 0$.

5. Practical examples

In this section we present three practical examples where Pareto smoothed importance sampling improves the estimates and where the Pareto shape estimate $\hat{k}$ is a useful diagnostic.
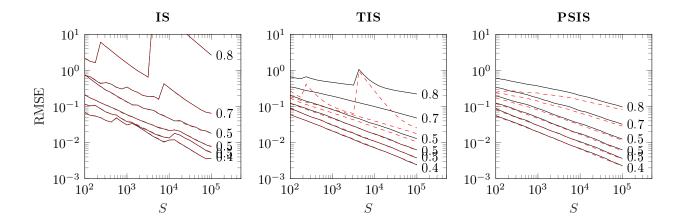
Figure 14: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_{20}(\mathbf{0}, I)$ and the proposal distribution is $\mathrm{N}(0.4\mathbf{1}, 0.9I)$, with the number of dimensions $D \in (1, 2, 4, 8, 16, 32, 64)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*

Figure 15: *The practical relative convergence rates for PSIS estimating 0th (blue), 1st (red) and 2nd (yellow) moments. Different target-proposal distribution pairs are used in different subplots. $\hat{k}$ estimates in case of 1st and 2nd moments are h specific. Dotted horizontal lines at 0.87 and 0.6 show the typical convergence rates at $\hat{k} = 0.5$ and $\hat{k} = 0.7$ for several different experiments in this paper.*

In the first example PSIS is used to improve the distributional approximation (split-normal) of the posterior of a logistic Gaussian process density estimation model. We then demonstrate the performance and reliability of PSIS for leave-one-out (LOO) cross-validation analysis of Bayesian predictive models for the canonical stacks data as well as for a recent breast cancer tumor dataset with 105 different protein expressions.

5.1. Improving distributional posterior approximation with importance sampling

For computational efficiency in Bayesian inference, posterior distributions are sometimes approximated using simpler parametric distributions. Typically these approximations can be further improved by using the distributional approximation as a proposal distribution in an importance sampling scheme. Here we demonstrate the benefit of using PSIS for improving the Laplace approximation of a logistic Gaussian process (LGP) for density estimation (Riihimäki and Vehtari, 2014).

LGP provides a flexible way to define the smoothness properties of density estimates via the prior covariance structure, but the computation is analytically intractable. Riihimäki and Vehtari (2014) propose a fast computation using discretization of the normalization term and Laplace's method for integration over the latent values.

Given n independently drawn d-dimensional data points $x_1, \ldots, x_n$ from an unknown distribution in a finite region (having a compact support) $\mathcal{V}$ of $\Re^d$, we want to estimate the density $p(x)$. To introduce the constraints that the density is non-negative and that its integral over $\mathcal{V}$ is equal to 1, Riihimäki and Vehtari (2014) employ the logistic density transform,

$$p(x) = \frac{\exp(f(x))}{\int_{\mathcal{V}} \exp(f(s))ds}, \tag{15}$$

where f is an unconstrained latent function. To smooth the density estimates, a Gaussian process prior is set for f, which allows for assumptions about the smoothness properties of the unknown density p to be expressed via the covariance structure of the GP prior. To make the computations feasible $\mathcal{V}$ is discretized into finite m subregions (or intervals if the problem is one-dimensional). Here we skip the details of the Laplace approximation and focus on the importance sampling.

Following Geweke (1989), Riihimäki and Vehtari (2014) use importance sampling with a multivariate split Gaussian density as an approximation. The approximation is based on the posterior mode and covariance, with the density adaptively scaled along principal component axes (in positive and negative directions separately) to better match the skewness of the target distribution (see also Villani and Larsson, 2006). To further improve the performance Riihimäki and Vehtari (2014) replace the discontinuous split Gaussian used by Geweke with a continuous version.

Riihimäki and Vehtari (2014) use an ad hoc soft thresholding of the importance weights if the estimated effective sample size as defined by Kong et al. (1994) is less than a specified threshold. The approach can be considered to be a soft truncated version of TIS, which Ionides (2008) also mentions as a possibility without further analysis. Here we propose to use PSIS to stabilize the weights.

We repeated the density estimation using the Galaxy data set[1] 1000 times with different random seeds. The model has 400 latent values, that is, the posterior is 400-dimensional, although due to a strong dependency imposed by the Gaussian process prior the effective dimensionality is smaller. Because of this it is sufficient that the split-normal is scaled only along the first 50 principal component axes. As a baseline we used Markov chain Monte Carlo as described in Riihimäki and Vehtari (2014). Computation time for MCMC inference was about half an hour and computation time for split-normal with importance sampling was about 1.3s (laptop with Intel Core i5-4300U CPU @ 1.90GHz x 4).

Figure 16 shows the Kullback-Leibler divergence from the density estimate using MCMC to the density estimates using the split-normal approximation with and without the importance sampling correction. The shaded areas show the envelope of the KL-divergence from all 1000 runs. The variability of the plain split-normal approximation (purple) diminishes as the number of draws S increases, but the KL-divergence does not decrease. IS (yellow) has high variability. PSIS (blue) performs very well, with a very small KL-divergence already when S is only 100. TIS results (not shown) were mostly similar to PSIS, with some rare worse results (similar jumps as in Figure 14). The mean estimate for $\hat{k}$ was 0.43 with $S = 100$ and 0.55 with $S = 10^4$, which explains the high variability of IS, the rare bad results from TIS, and the excellent performance of PSIS. These $\hat{k}$ values also signal that we can trust the PSIS results.

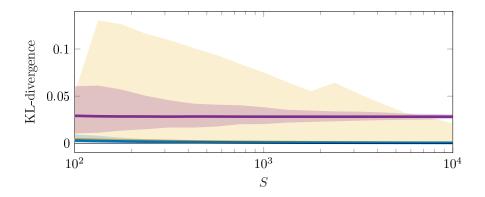[1]https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/galaxies.html

Figure 16: *Kullback-Leibler divergence from the density estimate using MCMC to the density estimates using the plain split-normal approximation (purple), IS (yellow), and PSIS (blue). The shaded areas show the envelope of the KL-divergence from all 1000 runs. The variability of plain split-normal approximation (purple) reduces with increasing number of draws S, but the KL-divergence doesn't decrease. IS (yellow) has high variability.*

The Pareto $\hat{k}$ diagnostic can also be used to compare the quality of the distributional approximations. In the case of a simple normal approximation without split scaling, the mean $\hat{k}$ with $S = 10^4$ was 0.60, and thus slightly higher variability and slower convergence can be assumed relative to the split-normal approximation.

The GPstuff toolbox (Vanhatalo et al., 2013) implementing logistic Gaussian process density estimation now uses PSIS for diagnostics and stabilization (code available at https://github.com/gstuff-dev/gpstuff). Another example of using PSIS to diagnose and stabilise importance sampling in Bayesian inference as part of an expectation propagation like algorithm can be found in Weber et al. (2016).

5.2. Importance-sampling leave-one-out cross-validation

We next demonstrate the use of Pareto smoothed importance sampling for leave-one-out (LOO) cross-validation approximation. The ith leave-one-out cross-validation predictive density can be approximated with

$$p(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} w_i(\theta^s) p(\tilde{y}_i|\theta^s)}{\sum_{s=1}^{S} w_i(\theta^s)}. \tag{16}$$

Importance sampling LOO was proposed by Gelfand et al. (1992), but it has not been widely used as the estimate is unreliable if the weights have infinite variance. For some simple models, such as linear and generalized linear models with specific priors, it is possible to analytically check the sufficient conditions for the variance of the importance weights in IS-LOO to be finite (Peruggia, 1997; Epifani et al., 2008), but this is not generally possible. This example also demonstrates the accuracy of Monte Carlo error estimates for the combination of MCMC and PSIS.

We first demonstrate properties of IS, TIS and PSIS with the stack loss data, which is known to have one observation producing infinite variance for LOO importance ratios. Then we demonstrate the speed and reliability of PSIS-LOO for performing model assessment and comparison for predictive regression models for 105 different protein expressions.

LOO for stack loss data The stack loss data has $n = 21$ daily observations on one response variable and three predictors pertaining to a plant for the oxidation of ammonia to
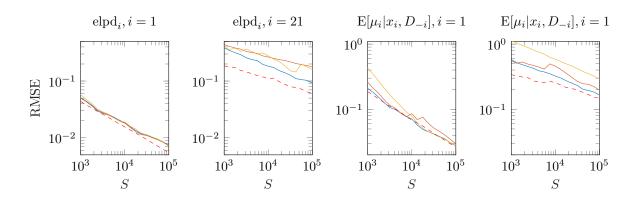
Figure 17: *RMSE with IS (yellow), TIS (red) and PSIS (blue), and the Monte Carlo error estimates with PSIS (red dashed) for the expected log predictive densities* $\text{elpd}_i = \log p(y_i|x_i, D_{-i})$ *and leave-one-out predictive mean* $\text{E}[\mu_i|x_i, D_{-i}]$. *Average h specific* $\hat{k}$'s, *using* $S = 10^5$ *draws from each of 100 runs, are 0.46, 0.79, 0.45, 0.81 (in the order of subplots).*

nitric acid. The model is a simple Gaussian linear regression. We fit the model using Stan (Stan Development Team, 2017) (the code is in the appendix). Peruggia (1997) showed that the importance ratios have an infinite variance when leaving out the 21st data point.

Figure 17 shows the RMSE and Monte Carlo estimate from 100 runs for the LOO estimated expected log predictive densities $\text{elpd}_i = \log p(y_i|x_i, D_{-i})$ and leave-one-out predictive mean $\text{E}[\mu_i|x_i, D_{-i}]$ (where D_{-i} denotes the data without ith observation) estimated with IS, TIS and PSIS when leaving out the 1st or 21st observation. Pareto smoothing and Monte Carlo error estimates were adjusted based on relative MCMC sample efficiency as discussed in Sections 3.2 and 3.3. The true values were computed by actually leaving out the ith observation and using very long MCMC chains to get a small Monte Carlo error. We see that PSIS gives the smallest RMSE, and the accuracy of the Monte Carlo error estimates are what we would expect based on h specific $\hat{k}$'s (i.e., error estimates are very accurate for $\hat{k} < 0.5$ and optimistic for $\hat{k} > 0.7$).

LOO for 105 protein expression data sets We demonstrate the benefit of fast importance sampling leave-one-out cross-validation and PSIS diagnostics with the example of a model for the combined effect of microRNA and mRNA expression on protein expression. The data were published by Aure et al. (2015) and are publicly available; we used the preprocessed data as described by Aittomäki (2016). Protein, mRNA, and microRNA expression were measured from 283 breast cancer tumor samples and when predicting the protein expression the corresponding gene expression and 410 microRNA expressions were used. We assumed a multivariate linear model for the effects with a Gaussian prior and used Stan (Stan Development Team, 2017) to fit the model. Initial analyses gave reason to suspect outlier observations; to verify this we compared Gaussian and Student-t observations models.

For 4000 posterior draws, the computation for one gene and one model takes about 9 minutes (desktop Intel Xeon CPU E3-1231 v3 @ 3.40GHz x 8), which is reasonable speed. For all 105 genes the computation takes about 30 hours. Exact regular LOO for all models would take 125 days, and 10-fold cross-validation for all models would take about 5 days. Pareto smoothed importance sampling LOO (PSIS-LOO) took less than one minute for all models. However, we do get several leave-one-out cases where $\hat{k} > 0.7$, which we should not trust based on our results above. Figure 18 shows $\hat{k} > 0.7$ values for 105 Gaussian and Student-t linear models, where each model may have several leave-one-out cases with $\hat{k} > 0.7$. Large $\hat{k}$ values arise when the proposal and the target distributions are very different, which
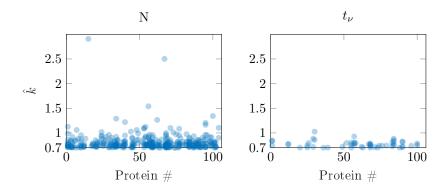
Figure 18: $\hat{k} > 0.7$ *values for 105 Gaussian and Student-t linear models predicting protein expression levels.*
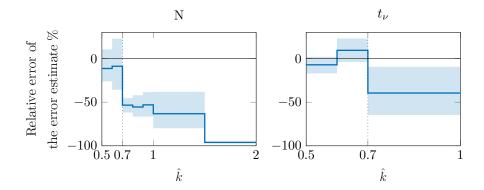

Figure 19: *Relative error of the PSIS Monte Carlo error estimates. True error was computed as RMSE of expected log predictive density* elpd_i *estimates with* $\hat{k}$ *on common interval (boundaries of intervals can be seen as steps in the plots).*

is typical when there are highly influential observations. Switching to the Student-t model reduces the number of high $\hat{k}$ values, as the outliers are less influential if they are far in the tail of the t-distribution. When working with many different models the Stan team has noticed that high $\hat{k}$ values are also a useful indicator that there is something wrong with the data or the model (personal communication).

Figure 19 shows the accuracy of PSIS Monte Carlo error estimates for the expected log predictive densities with respect to different $\hat{k}$ values (computed only for $\hat{k} > 0.5$). True values were computed by actually leaving out the ith observation and rerunning MCMC. We can see that $\hat{k}$ is a useful diagnostic and the Monte Carlo error estimates are accurate for $\hat{k} < 0.7$, as in the simulation experiments, and fail for $\hat{k} \geq 0.7$.

To improve upon PSIS-LOO we can make the exact LOO computations for any points corresponding to $\hat{k} > 0.7$ (for which we cannot trust the Monte Carlo error estimates). In this example there were 352 such cases for the Gaussian models and 53 for the Student-t models, and the computation for these took 42 hours. Although combining PSIS-LOO with exact LOO for certain points substantially increases the computation time in this example, it is still less than the time required for 10-fold-CV.

The left subplot in figure 20 shows comparison of PSIS-LOO and PSIS-LOO+ (PSIS-LOO with exact computation for cases with $\hat{k} > 0.7$) when comparing the difference of expected log predictive densities $\sum_{i=1}^{n} \left(\mathrm{elpd}_i(t_\nu) - \mathrm{elpd}_i(N) \right)$. We see that with high $\hat{k}$ values, the error of PSIS-LOO can be very large (the error would be large for IS and TIS, too). To trust the
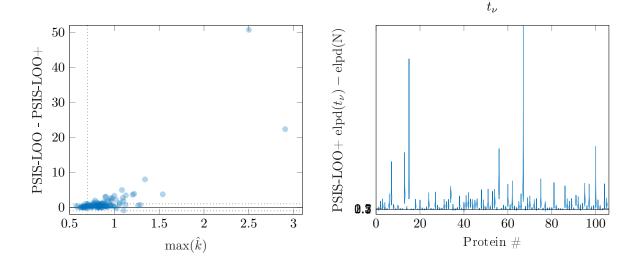
Figure 20: *The left plot shows comparison of PSIS-LOO and PSIS-LOO+ (PSIS-LOO with exact computation for cases with $\hat{k} > 0.7$) when comparing the difference of expected log predictive densities $\sum_{i=1}^{n}(\mathrm{elpd_i(t_\nu)} - \mathrm{elpd_i(N)})$. The right plot shows the PSIS-LOO+ estimated improvement of expected log predictive densities when switching from Gaussian model to Student-t model.*

model comparison we recommended using PSIS-LOO+ with exact computation for cases with $\hat{k} > 0.7$ or K-fold-CV. The right subplot shows the final model comparison results for all 105 models predicting protein expression levels. For most of the proteins, the student-t model is much better, and the Gaussian model is not significantly better for any of the proteins.

6.   Discussion

Importance weighting is a widely used tool in statistical computation. Even in the modern era of Markov chain Monte Carlo, approximate algorithms are often necessary, and then there is a desire to adjust approximations to better match target distributions. However, a challenge for practical applications of importance weighting is the well known fact that importance-weighted estimates are unstable if the weights have high variance.

In this paper we have shown that it is possible to reduce the mean square error of importance sampling estimates using a particular stabilizing transformation that we call Pareto smoothed importance sampling (PSIS). The key step is to replace the largest weights by expected quantiles from a generalized Pareto distribution. We have also demonstrated greatly improved Monte Carlo error estimates, natural diagnostics for gauging the reliability of the estimates, and empirical convergence rate results that closely follow known theoretical results.

We believe this method will be helpful in many cases where importance sampling is used. In addition to the examples in this paper, PSIS has been used to stabilize importance sampling as a part of the complex algorithm in Weber et al. (2016), and we are currently investigating its use in particle filtering, adaptive importance sampling, and as a diagnostic for auto-differentiated variational inference (Kucukelbir et al., 2014).

References

Aittomäki, V. (2016). MicroRNA regulation in breast cancer—a Bayesian analysis of expression data. Master's thesis, Aalto University.

Aure, M. R., Jernström, S., Krohn, M., Vollan, H. K., Due, E. U., Rødland, E., Kåresen, R., Ram, P., Lu, Y., Mills, G. B., Sahlberg, K. K., Børresen-Dale, A. L., Lingjærde, O. C., and Kristensen, V. N. (2015). Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome Medicine*, 7(1):21.

Bornn, L., Doucet, A., and Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64.

Chen, L. H. Y. and Shao, Q.-M. (2004). Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028.

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

Epifani, I., MacEachern, S. N., and Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2:774–806.

Ferreira, A., de Haan, L., and Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics*, 37(5):401–434.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 145–162. Chapman & Hall.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 147–167. Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174.

Hill, J. B. (2010). On tail index estimation for dependent, heterogeneous data. *Econometric Theory*, 26(5):1398–1436.

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

Kong, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, University of Chicago, Department of Statistics.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288.

Koopman, S. J., Shephard, N., and Creal, D. (2009). Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2014). Fully automatic variational inference of differentiable probability models. In *Proceedings of the NIPS Workshop on Probabilistic Programming*.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Owen, A. B. (2013). Monte Carlo theory, methods and examples. Online book at http://statweb.stanford.edu/~owen/mc/, accessed 2017-09-09.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, 92(437):199–207.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131.

Pitt, M. K., Tran, M.-N., Scharth, M., and Kohn, R. (2013). On the existence of moments for high dimensional importance sampling. *arXiv:1307.7975*.

Raftery, A. E. and Bao, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, 66(4):1162–1173.

Riihimäki and Vehtari, A. (2014). Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, second edition*. Springer.

Scarrot, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT – Statistical Journal*, 10(1):33–60.

Stan Development Team (2017). *Stan modeling language: User's guide and reference manual*. Version 2.16.0, http://mc-stan.org/.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179.

Vehtari, A., Gelman, A., and Gabry, J. (2017a). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, R package version 1.1.0. https://github.com/stan-dev/loo.

Vehtari, A., Gelman, A., and Gabry, J. (2017b). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Villani, M. and Larsson, R. (2006). The multivariate split normal distribution and asymmetric principal components analysis. *Communications in Statistics: Theory & Methods*, 35(6):1123–1140.

Weber, S., Gelman, A., Carpenter, B., Lee, D., Betancourt, M., Vehtari, A., and Racine, A. (2016). Hierarchical expectation propagation for Bayesian aggregation of average data. *arXiv:1602.02055*.

Zhang, J. and Stephens, M. A. (2009). A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, 51(3):316–325.
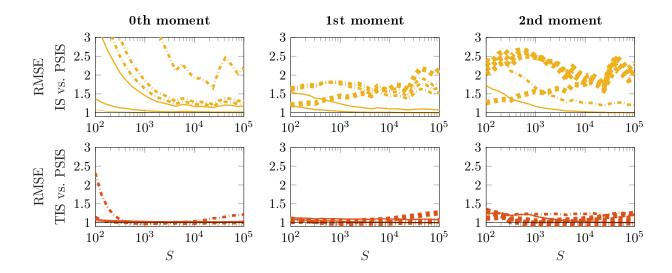
A.   Additional figures

Figure 21: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is* $\mathrm{Exp}(1)$ *and the proposal distribution is* $\mathrm{Exp}(1/\theta)$, *with* $\theta \in (1.3, 1.5, 2, 3, 4, 10)$. *The thinner continuous lines have* $k \leq \frac{1}{3}$, *and the thicker continuous lines have* $\frac{1}{3} < k \leq \frac{1}{2}$. *The dash-dotted lines have* $\frac{1}{2} < k \leq 0.7$ *and thick dashed lines have* $k > 0.7$. *For the 1st and 2nd moment* k *is the h specific version.*
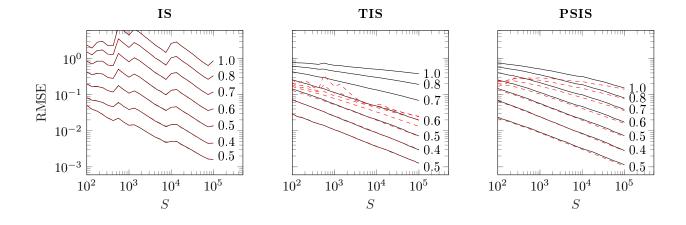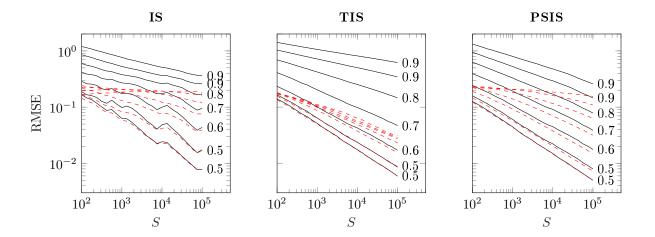
Figure 22: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is* $N(0, 1)$ *and the proposal distribution is* $N(\mu, 0.8)$, *with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. *For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high* θ *values leading to high RMSE and high* $\hat{k}$. *The numbers at the end of black lines are average of* $\hat{k}$ *values estimated when* $S = 10^5$.

Figure 23: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is* $N(0, 1)$ *and the proposal distribution is* $N(\mu, 0.8)$, *with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. *For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high* θ *values leading to high RMSE and high* $\hat{k}$. *The numbers at the end of black lines are average of h specific* $\hat{k}$ *values estimated when* $S = 10^5$.

27
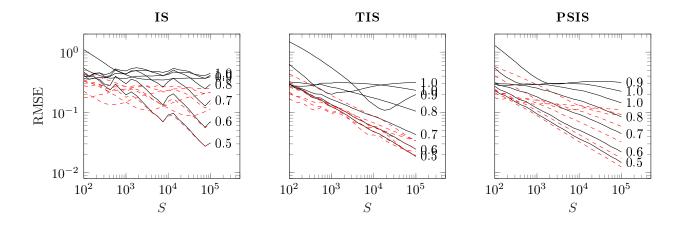
Figure 24: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 2nd moment estimate. The target distribution is* $\mathrm{N}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.8)$, *with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. *For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high* θ *values leading to high RMSE and high* $\hat{k}$. *The numbers at the end of black lines are average of h specific* $\hat{k}$ *values estimated when* $S = 10^5$.
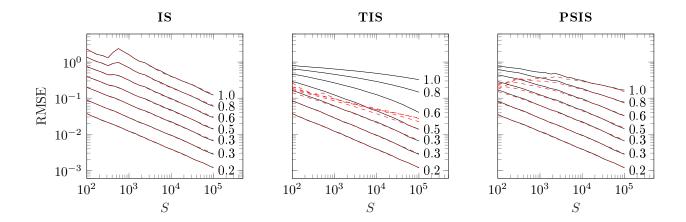

Figure 25: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is* $t_{20}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.9)$, *with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. *For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high* θ *values leading to high RMSE and high* $\hat{k}$. *The numbers at the end of black lines are average of* $\hat{k}$ *values estimated when* $S = 10^5$.
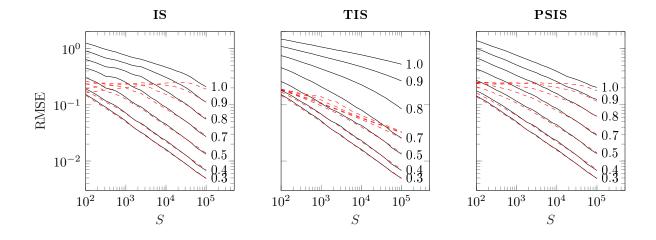
Figure 26: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $N(\mu, 0.9)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
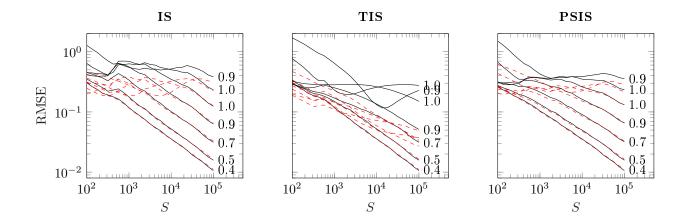

Figure 27: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 2nd moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $N(\mu, 0.9)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
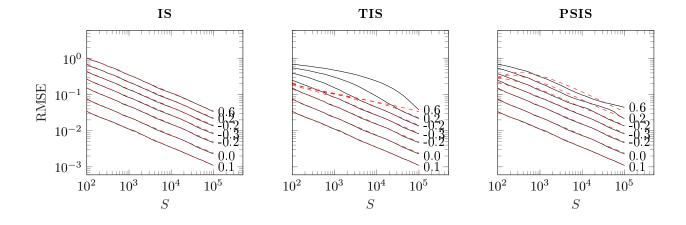
Figure 28: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_{20}(0, 1)$ and the proposal distribution is $t_{21}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*
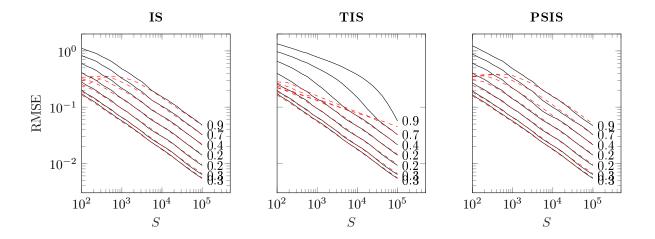

Figure 29: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_{20}(0, 1)$ and the proposal distribution is $t_{21}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*

Figure 30: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 2nd moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $t_{21}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
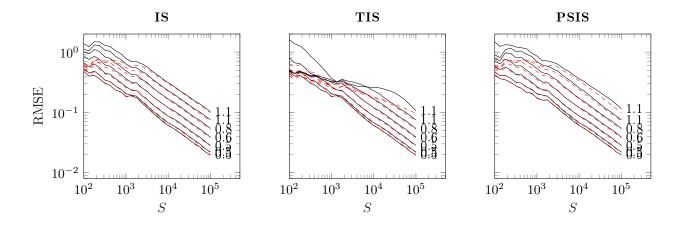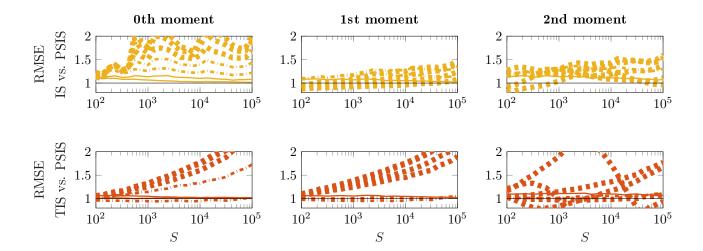

Figure 31: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*

Figure 32: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*

Figure 33: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 2nd moment estimate. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
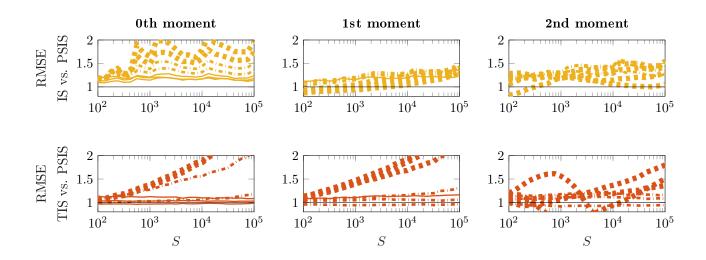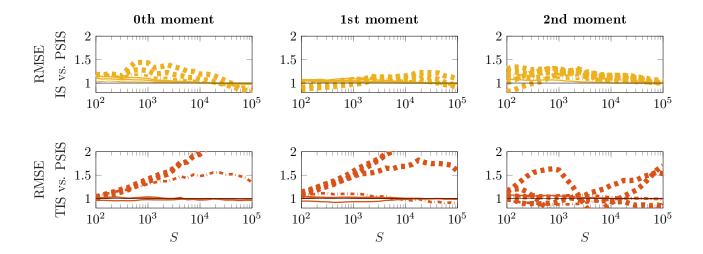
Figure 34: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is* $\mathrm{N}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.8)$*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. The thinner continuous lines have* $k \leq \frac{1}{3}$*, and the thicker continuous lines have* $\frac{1}{3} < k \leq \frac{1}{2}$*. The dash-dotted lines have* $\frac{1}{2} < k \leq 0.7$ *and thick dashed lines have* $k > 0.7$*. For the 1st and 2nd moment* k *is the h specific version.*
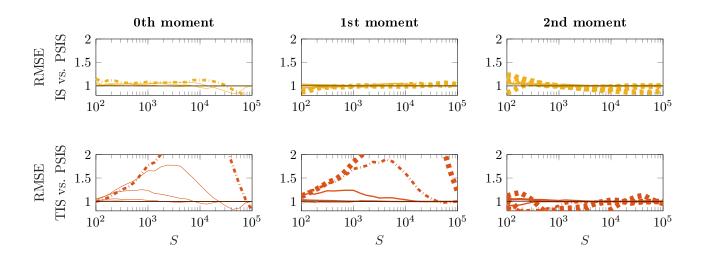

Figure 35: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is* $t_{20}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.9)$*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. The thinner continuous lines have* $k \leq \frac{1}{3}$*, and the thicker continuous lines have* $\frac{1}{3} < k \leq \frac{1}{2}$*. The dash-dotted lines have* $\frac{1}{2} < k \leq 0.7$ *and thick dashed lines have* $k > 0.7$*. For the 1st and 2nd moment* k *is the h specific version.*

Figure 36: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is $t_{21}(0,1)$ and the proposal distribution is $t_{22}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. The thinner continuous lines have $k \leq \frac{1}{3}$, and the thicker continuous lines have $\frac{1}{3} < k \leq \frac{1}{2}$. The dash-dotted lines have $\frac{1}{2} < k \leq 0.7$ and thick dashed lines have $k > 0.7$. For the 1st and 2nd moment k is the h specific version.*

Figure 37: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. The thinner continuous lines have $k \leq \frac{1}{3}$, and the thicker continuous lines have $\frac{1}{3} < k \leq \frac{1}{2}$. The dash-dotted lines have $\frac{1}{2} < k \leq 0.7$ and thick dashed lines have $k > 0.7$. For the 1st and 2nd moment k is the h specific version.*
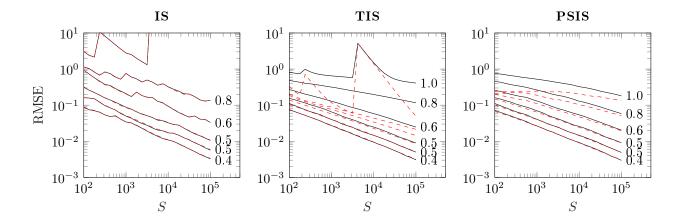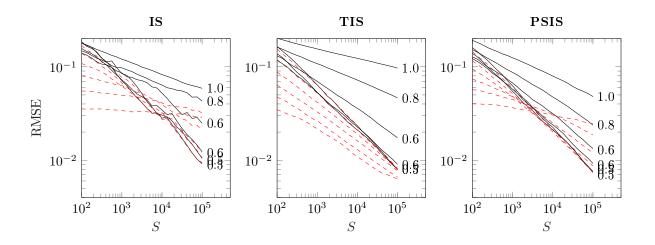
Figure 38: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is* N(0, 1) *and the proposal distribution is* N(μ, 0.8)*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when* $S = 10^5$*.*

Figure 39: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is* N(0, 1) *and the proposal distribution is* N(μ, 0.8)*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when* $S = 10^5$*.*
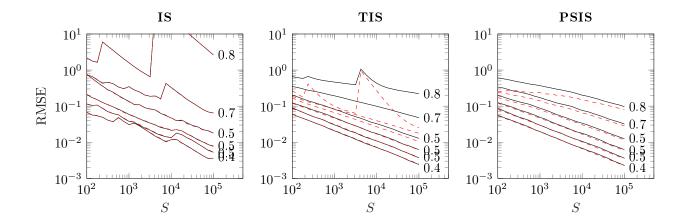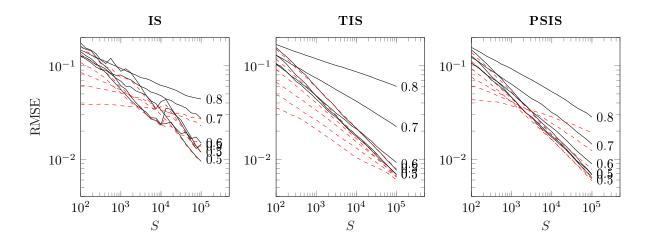
Figure 40: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $N(\mu, 0.9)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*
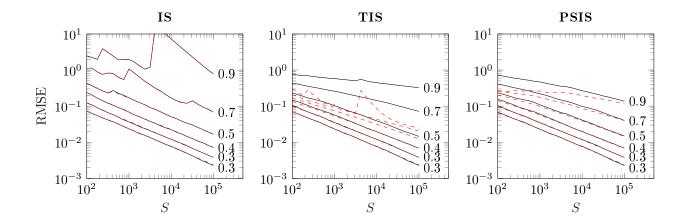

Figure 41: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $N(\mu, 0.9)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
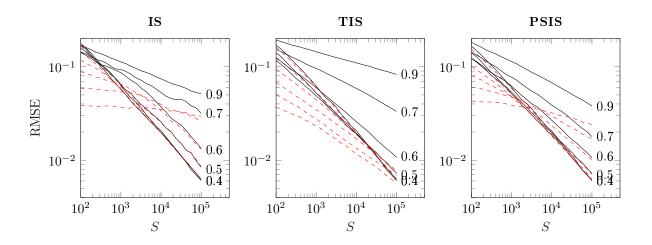
36

Figure 42: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $t_{21}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*

Figure 43: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_{20}(0,1)$ and the proposal distribution is $t_{21}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
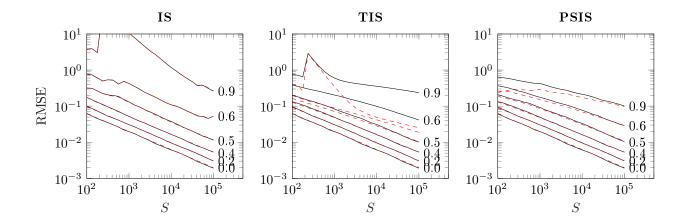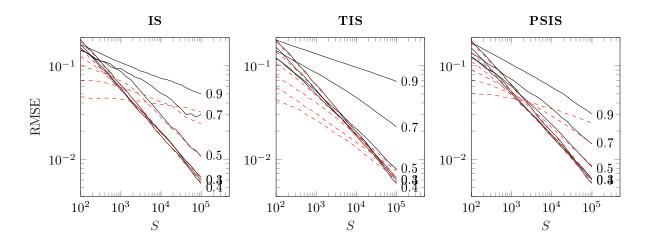
37

Figure 44: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 0th moment estimate. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of $\hat{k}$ values estimated when $S = 10^5$.*

Figure 45: *RMSE (black) and the mean of the Monte Carlo error estimates (red dashed) for the 1st moment estimate. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. For each graph, the lines are ordered with low values of theta at bottom and high values at top, with high θ values leading to high RMSE and high $\hat{k}$. The numbers at the end of black lines are average of h specific $\hat{k}$ values estimated when $S = 10^5$.*
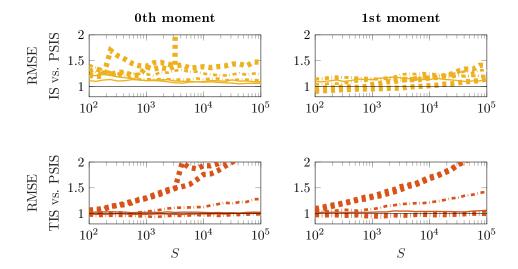
Figure 46: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is* $\mathrm{N}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.8)$*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. The thinner continuous lines have* $k \leq \frac{1}{3}$*, and the thicker continuous lines have* $\frac{1}{3} < k \leq \frac{1}{2}$*. The dash-dotted lines have* $\frac{1}{2} < k \leq 0.7$ *and thick dashed lines have* $k > 0.7$*. For the 1st and 2nd moment* k *is the h specific version.*
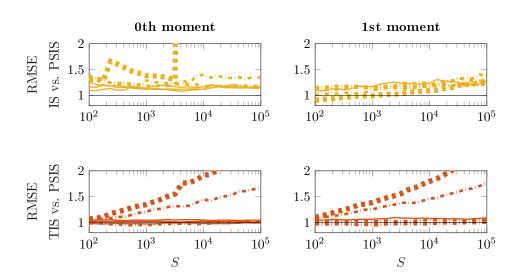

Figure 47: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is* $t_{20}(0,1)$ *and the proposal distribution is* $\mathrm{N}(\mu, 0.9)$*, with* $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$*. The thinner continuous lines have* $k \leq \frac{1}{3}$*, and the thicker continuous lines have* $\frac{1}{3} < k \leq \frac{1}{2}$*. The dash-dotted lines have* $\frac{1}{2} < k \leq 0.7$ *and thick dashed lines have* $k > 0.7$*. For the 1st and 2nd moment* k *is the h specific version.*
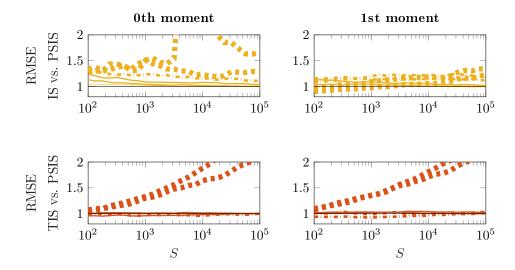
Figure 48: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is $t_{21}(0,1)$ and the proposal distribution is $t_{22}(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. The thinner continuous lines have $k \leq \frac{1}{3}$, and the thicker continuous lines have $\frac{1}{3} < k \leq \frac{1}{2}$. The dash-dotted lines have $\frac{1}{2} < k \leq 0.7$ and thick dashed lines have $k > 0.7$. For the 1st and 2nd moment k is the h specific version.*
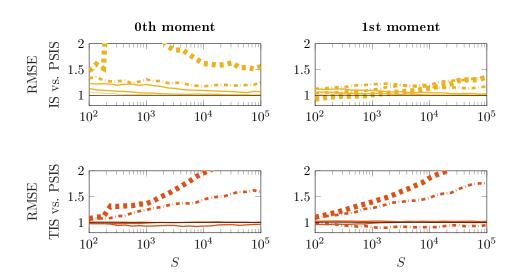

Figure 49: *Relative difference in RMSE between IS, TIS and PSIS for the 0th, 1st and 2nd moment estimates. The target distribution is $t_7(0,1)$ and the proposal distribution is $t_8(\mu, 0.8)$, with $\mu \in (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$. The thinner continuous lines have $k \leq \frac{1}{3}$, and the thicker continuous lines have $\frac{1}{3} < k \leq \frac{1}{2}$. The dash-dotted lines have $\frac{1}{2} < k \leq 0.7$ and thick dashed lines have $k > 0.7$. For the 1st and 2nd moment k is the h specific version.*

B. Stan Gaussian linear regression model for stack loss data

```
data {
  int<lower=0> N;
  int<lower=0> p;
  vector[N] y;
  matrix[N,p] x;
}
// to standardize the x's
transformed data {
  matrix[N,p] z;
  vector[p] mean_x;
  vector[p] sd_x;
  real sd_y;
  sd_y <- sd(y);
  for (j in 1:p) {
    mean_x[j] <- mean(col(x,j));
    sd_x[j] <- sd(col(x,j));
    for (i in 1:N)
      z[i,j] <- (x[i,j] - mean_x[j]) / sd_x[j];
  }
}
parameters {
  real beta0;
  vector[p] beta;
  real<lower=0> sigmasq;
  real<lower=0> phi;
}
transformed parameters {
  real<lower=0> sigma;
  vector[N] mu;
  sigma <- sqrt(sigmasq);
  mu <- beta0 + z * beta;
}
model {
  beta0 ~ normal(0, 100);
  phi ~ cauchy(0, sd_y);
  beta ~ normal(0, phi);
  sigmasq ~ inv_gamma(.1, .1);
  y ~ normal(mu, sigma);
}
generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] <- normal_log(y[i], mu[i], sigma);
}
```