Pareto Smoothed Importance Sampling*

Aki Vehtari† Andrew Gelman‡ Jonah Gabry‡

22 September 2016

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.

Keywords: importance sampling, Monte Carlo, Bayesian computation

1. Introduction

Importance sampling is a simple correction that is used when we can more easily obtain samples from some approximating distribution than directly from the target distribution. 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 bias but also provides a natural diagnostic for gauging the reliability of the estimate.

After presenting some background material in Section 2, we present our proposed method in Section 3, toy examples in Section 4, and pratical examples in Section 5, concluding in Section 6 with a brief discussion. Some of the content in Sections 2 and 3 is covered in Vehtari et al. (2016b) in relation to approximate leave-one-out cross-validation. Here we present it again in more general form, supplemented with additional details and more general commentary not tied to any specific application of the algorithm.

*We thank Juho Piironen for help with R, Viljami Aittomäki for help with gene expression data, and the Alfred P. Sloan Foundation, U.S. National Science Foundation, Institute for Education Sciences, and Office of Naval Research for partial support of this research.

†Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland.

‡Department of Statistics, Columbia University, New York.

2. Importance sampling

Suppose we want to estimate an integral,

$$\int h(\theta)p(\theta)d\theta, \tag{1}$$

where $h(\theta)$ is a function and $p(\theta)$ is a probability density which we cannot easily or cheaply draw from directly. Instead there is an approximating density $g(\theta)$ from which we can easily generate random draws. The integral (1) can be rewritten,

$$\int h(\theta)p(\theta) = \frac{\int [h(\theta)p(\theta)/g(\theta)]\, g(\theta)d\theta}{\int [p(\theta)/g(\theta)]\, g(\theta)d\theta}, \tag{2}$$

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

$$\frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)r(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} r(\theta^s)}, \tag{3}$$

where the factors

$$r(\theta^s) = \frac{p(\theta^s)}{g(\theta^s)} \tag{4}$$

are called *importance ratios*.

If p is a normalized probability density, the denominator of (2) 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 (3), for which only the relative values of the importance ratios are needed.

In Bayesian analysis, proposal distributions g are often recommended based on simple approximations, 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. Another 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.

2.1. From importance ratios to importance weights

Geweke (1989) shows that if the variance of the distribution of importance ratios is finite, then the central limit theorem holds for the convergence of the estimate in (3). Chen and Shao (2004) show further that the rate of convergence to normality is faster when higher moments exist. In simple 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). In general, however, we would like a procedure that works based on the importance ratios alone, without requiring additional analysis of the distributions.

If the variance of the distribution of importance ratios does not exist, then in general the importance weighted estimate cannot be trusted, and it makes sense to replace the importance ratios by more stable weights. Thus, (3) is replaced by

$$\frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)w(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} w(\theta^s)}, \tag{5}$$

where the *importance weights* w are some function of the importance ratios r from (4). Two extremes are $w \propto r$ (raw importance weighting) and $w \propto 1$ (identity weights, equivalent to just using the approximating distribution g).

Approaches that stabilize the weights can be interpreted as compromising between an unstable estimate of p and a stable computation using g. It would be tempting to characterize this as a bias-variance tradeoff, but that would not be quite correct because raw importance weighting (3) is a ratio estimate and thus is itself biased, and this bias can be considerable if the distribution of the importance ratios is long-tailed.

2.2. Truncated importance sampling

Truncatated 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

$$w_s = \min\left(r_s, \sqrt{S}\bar{r}\right), \tag{6}$$

where $\bar{r}$ is the average of the original S importance ratios. Ionides (2008) not only proves that the distribution of these weights is guaranteed to have finite variance, but also shows that the resulting importance sampling estimate of interest has a mean square error close to that of an estimate obtained using a case-specific optimal truncation point. Unfortunately, while this truncation method can greatly improve stability, it comes at the expense of bias. And, as our examples will demonstrate, this bias can be large.

2.3. Sample based estimate using the generalized Pareto distribution

To make a sample based estimate of the existing moments, Koopman et al. (2009) suggest analyzing the properties of a generalized Pareto distribution fit to the upper tail of the distribution of the importance ratios $r(\theta^s)$. 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,

$$p(y|u,\sigma,k) = \begin{cases} \frac{1}{\sigma}\left(1 + k\left(\frac{y-u}{\sigma}\right)\right)^{-\frac{1}{k}-1}, & k \neq 0 \\ \frac{1}{\sigma}\exp\left(\frac{y-u}{\sigma}\right), & k = 0, \end{cases} \tag{7}$$

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.

Koopman et al. (2009) set the lower bound u to the chosen threshold and use a maximum likelihood estimate for (σ, k). Because the generalized Pareto distribution has the property that when $k > 0$ the number of existing moments is less than $\lfloor 1/k \rfloor$, they then form a statistical test of the hypothesis $k < 1/2$, from which an inference can be drawn about whether the underlying distribution has a finite variance.

Pickands (1975) notes that to obtain asymptotic consistency, the threshold u should be chosen so that the sample size M in the tail should increase to infinity while the proportion of simulation draws in the tail, M/S, goes to zero.

Zhang and Stephens (2009) propose a quick empirical Bayes flavored estimation method for the parameters of the generalized Pareto distribution, which 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, and is both simple and fast to compute. It is likely that an even better estimate could be obtained from a fully Bayesian approach, but since speed is important in many applications we use their fast estimate for now.

3. Pareto smoothed importance sampling

We propose a novel importance sampling estimate that has the finite-variance property of truncated importance sampling while also reducing bias by fitting the generalized Pareto distribution to the upper tail of the weight distribution.

Generalized Pareto distribution fit to the tail. We start by fitting a generalized Pareto distribution (7) to the importance weight values above a threshold u. As mentioned in Section 2.3, the shape parameter k of the generalized Pareto distribution can be used to characterize the thickness of the of the tail of the importance weight distribution and determine the existence of moments.

- If $k < \frac{1}{2}$ then the distribution of importance ratios has finite variance. In this case the central limit theorem holds and we can observe fast convergence of the estimate. Given reasonably large S, the estimate is not sensitive to the largest weight values.

- If $\frac{1}{2} \leq k < 1$ then the variance is infinite, but the mean exists. As the generalized central limit theorem for stable distributions holds, the distribution of the estimate converges to a stable distribution, although the convergence of the raw importance sampling estimate is slower and the estimate is sensitive to rare large weights in the long tail.

- If $k \geq 1$ then neither the variance nor the mean exists.

In our experiments we choose the threshold u so that proportion of the samples in the tail, M/S, is between 10% and 20%, depending on how large S is. As the value of M increases, the uncertainty in the estimate shrinks, but the bias is reduced when the ratio M/S is smaller. The experimental results were not sensitive to the exact value of u.

Smoothing of the 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. As the largest weight is $F^{-1}\left(\frac{M-1/2}{M}\right)$, the weights are truncated and the variance of the estimate is finite. Compared to the simple truncation by Ionides (2008), the bias is reduced as the largest weights are not all truncated to the same value, but rather spread according to the estimated tail shape.

Truncation of very large weights. When the shape parameter k is close to or larger than 1, small changes in the estimate of k have a big effect on the largest values of the inverse-CDF. To reduce this variation, we use an additional truncation of the largest weights despite the potential for increased the bias. Since the large weights have been smoothed, we need less truncation than in the truncated importance sampling by Ionides (2008). Based on simulations we recommend truncation with $S^{3/4}\bar{w}$, where $\bar{w}$ is the average of the smoothed weights.

Diagnostics. Previous research has focused on identifying whether the variance is finite or infinite (Peruggia, 1997; Epifani et al., 2008; Koopman et al., 2009), but we demonstrate in Section 4that it can often be more useful to look at the continuous $\hat{k}$ values than the discrete number of moments. Based on our experiments, Pareto smoothed importance sampling is able to provide estimates with relatively small variance and bias when $k < 0.7$. With $k > 0.7$ the estimator can have high variance and high bias, and if the method is implemented in a software package we recommend reporting to the user if $\hat{k} > 0.7$. Depending on the application a lower threshold could also be used.

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

1. Set $M = 0.2S$ and set u to the 80th percentile of the values of r_s.
2. Fit the generalized Pareto distribution (7) 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).

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 value $S^{3/4}\bar{w}$, where $\bar{w} = \frac{1}{S}\sum_{s=1}^{S} w_s$ is the average weight value.

5. If the estimated shape parameter $\hat{k}$ exceeds 0.5 (or 0.7, as suggested above), report a warning that the resulting importance sampling estimates might be unstable.

This method has been implemented in an R function called `psislw` which is included in the `loo` R package (Vehtari et al., 2016a). 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`.

4. Toy examples

In the following toy examples we know the true target value and vary the proposal distribution. This allows us to study how $\hat{k}$ works as diagnostic and how bad the approximate distribution has to be before the importance sampling estimates break down. In each of the examples we simulate $S = 16000$ draws from the proposal distribution. We repeated the experiments with $S = 4000$ and $S = 1000$ and obtained similar results, but the differences between the methods were clearer with larger S.

4.1. Exponential, approximating distribution is too narrow

In the first toy example, we demonstrate why it is useful to look at the continuous $\hat{k}$ value, instead of discrete number of moments. The proposal distribution is exponential with rate 1 and the target distribution is exponential with varying value of the rate parameter, $\theta = 1.3, 1.5, 1.9, 2.0, 2.1, 3.0,$ or 10.0. The target function is $h \equiv \theta$, that is, we are estimating the mean of the target distribution. In this simple case it is possible to compute analytically that the variance is finite only if $\theta < 2$ (Robert and Casella, 2004).

Figure 1 shows the comparison of regular importance sampling (IS), truncated importance sampling (TIS), and Pareto smoothed importance sampling (PSIS). The vertical axis is the ratio of the estimated mean to the true mean (i.e., values close to 1 are good). For each case the mean of the estimated Pareto shape values $\hat{k}$ is shown along the horizontal axis. The figure confirms that when the variance is finite the errors are smaller, but the Pareto shape value $\hat{k}$ gives additional information about the distribution of errors in both the finite and infinite variance cases (since we are using a finite number of samples, the sample variance is finite). Compared to the other methods, Pareto smoothing reduces the variance of the estimate when $k < 1/2$; PSIS has lower variance than both IS (6–80%) and TIS (6–40%). Using PSIS we can get low variance estimates also when $k \geq 1/2$ (and IS now has infinite variance). Also notable is that the bias of PSIS is relatively small for k less than about 0.7, and the bias of TIS is larger than the bias of PSIS in all cases. The results of the other

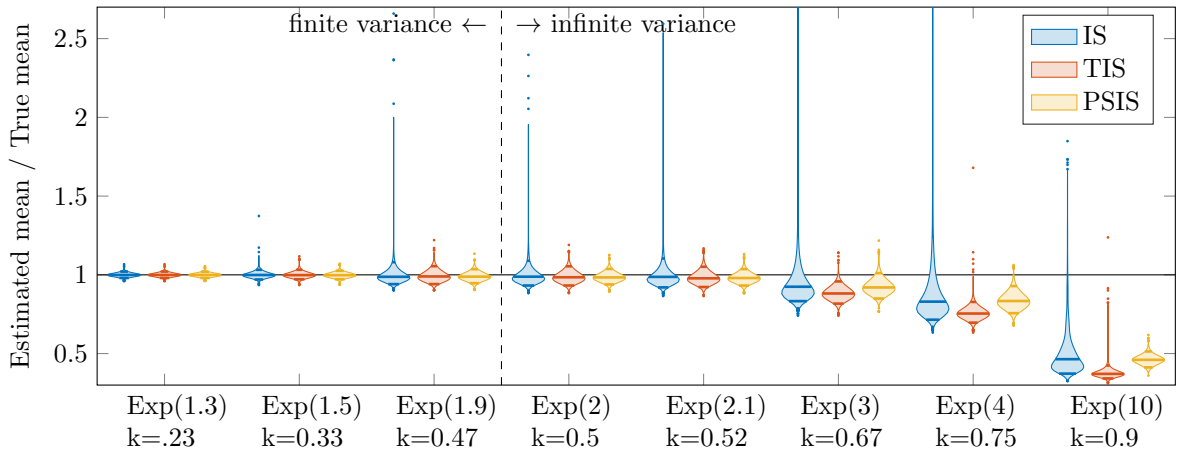

Figure 1: *Toy example 1: Violin plot for regular importance sampling (blue), truncated importance sampling (red), and Pareto smoothed importance sampling (yellow) estimates of* $\mathrm{E}(\theta)/\theta_0$ *in 10000 repeated simulations with different values of* θ_0*, in which the proposal distribution is exponential with unit rate, and the target distribution is exponential with different choices of rate parameter. The true mean in this example is 1 in each case. The vertical axis has been truncated to 2.8, cutting off two simulations (out of 10000) where the estimate was up to 5 times larger than the true value.*

experiments reported below show similar behavior of the compared methods with respect to the Pareto shape value k.

4.2. Univariate normal, approximating distribution is too narrow.

Our next example is the same as the example used by Ionides (2008) to demonstrate the performance of TIS when the proposal distribution is narrower than the target. The target distribution is $p(\theta) = N(\theta \,|\, 0, 1)$ and the proposal is $g(\theta) = N(\theta \,|\, 0, \sigma^2)$, with simulations performed for $\sigma = 0.1, 0.2, \ldots, 0.8$. In this and all subsequent examples the target function is $h \equiv 1$; that is, we are implicitly estimating the ratio of the normalizing constants of the target and approximate distributions.

We plot the estimate $\widehat{E(h)}$ on the log scale as it improves the clarity of the figures. The true value for the logarithm of the integral is 0. Figure 2 shows comparison of IS, TIS, and PSIS for the second toy example. TIS and PSIS have much smaller variance than IS, and PSIS has a slightly smaller bias than TIS. All methods fail if the proposal distribution is much narrower than the target distribution. Figure 3 shows the PSIS estimate of $E(h)$ versus the estimated tail shape k over 1000 repeated simulations. Both the variance and bias increase considerably after $\hat{k} > 1/2$.

4.3. Univariate t, approximating distribution is shifted and slightly too narrow.

In the previous example, the means of the target and the proposal distributions were the same. Next we demonstrate the performance of the various methods when the mean of the

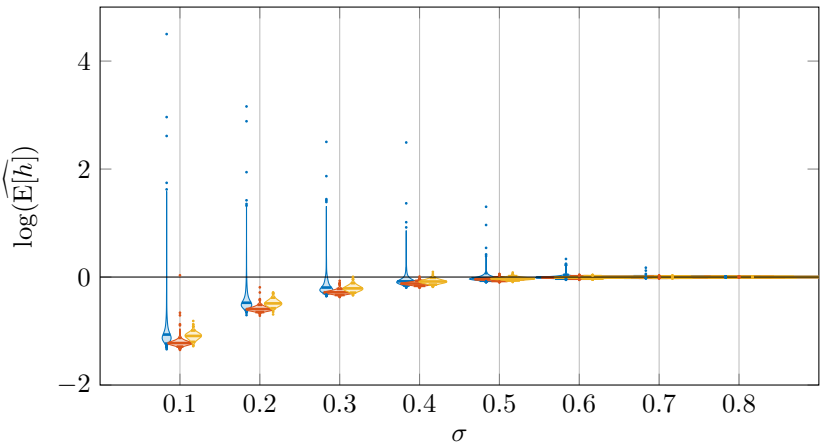

Figure 2: *Toy example 2: Violin plot for regular importance sampling (blue), truncated importance sampling (red), and Pareto smoothed importance sampling (yellow) estimates of* $\log \mathrm{E}(h)$ *in 1000 repeated simulations with different values of* $\sigma \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$. *The true value in this example is 0 in each case.*

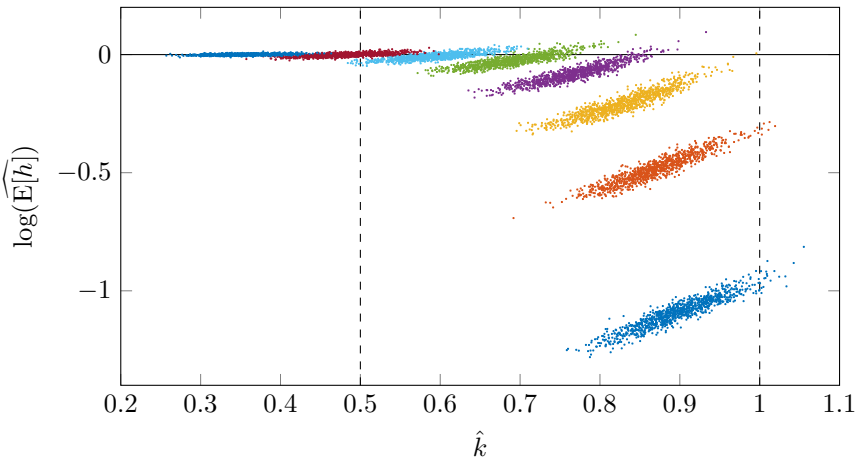

Figure 3: *Toy example 2: Scatterplot for Pareto smoothed importance sampling estimates of* $\mathrm{E}(h)$ *versus the estimated tail shape* $\hat{k}$ *in 1000 repeated simulations. The different colors separate the points by value of* $\sigma \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)$. *Larger* $\hat{k}$ *values correspond to smaller values of* σ.

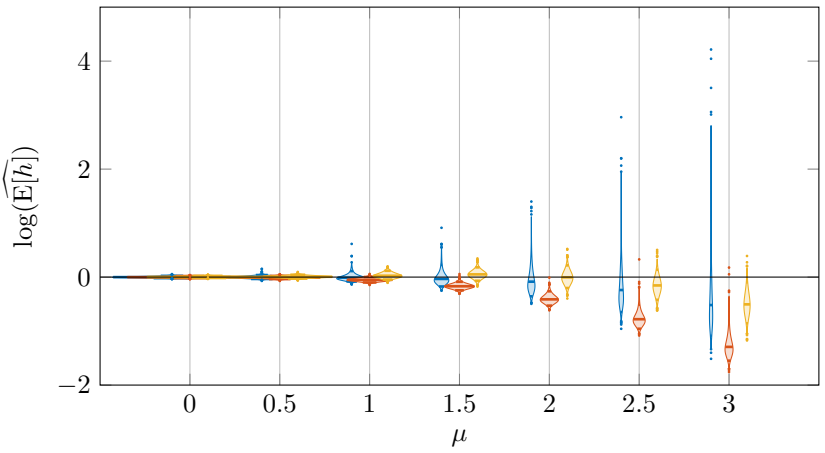

Figure 4: *Toy example 3: Violin plot for raw importance sampling (blue), truncated importance sampling (red), and Pareto smoothed importance sampling (yellow) estimates of* $\mathrm{E}(h)$ *in 1000 repeated simulations with the shift parameter* μ *set to the increasingly challenging values,* $0, 0.5, \ldots, 2.5, 3$*. The true value for* $\log \mathrm{E}(y)$ *is 0 here, as in our other toy examples.*

proposal distribution is varied. The target is $p(\theta) = t_{21}(\theta \,|\, 0, 1)$ and the proposal distribution is $g(\theta) = t_{21}(\theta \,|\, \mu, 1 - \frac{1}{21})$, with simulations performed for $\mu = 0, 0.5, \ldots, 2.5, 3.0$. We use the t distribution in this simulation to better match the distributions we might encounter in Bayesian inference, and the use of $1 - \frac{1}{21}$ represents the scenario from leave-one-out cross-validation in which the approximating distribution includes all the data and thus tends to have slightly lower variance than the target, which excludes one of the factors from the likelihood.

Figure 4 shows the comparison of the methods for the third toy example. TIS and PSIS have much smaller variance than IS, and PSIS has much smaller bias than TIS. All methods eventually fail when the proposal distribution is far away from the essential mass of the target distribution. Figure 5 shows the variation in the estimated tail shape parameter $\hat{k}$ and the PSIS estimate. It can again be seen that the variance and bias increase after $\hat{k} > 1/2$, but even if $\hat{k} > 1$ the bias from PSIS is still small in this example.

4.4. Multivariate normal, approximating distribution is shifted and slightly too narrow.

In our final toy example we compare the performance of the various importance sampling methods as we increase the number of dimensions of the vector θ. The target distribution is $p(\theta) = N(\theta \,|\, 0, I)$ and the proposal is $g(\theta) = t_{21}(\theta \,|\, 0.4 \times \mathbf{1}, 0.8I)$, and we examine performance in one dimension and then for dimensions $p = 20, 40, \ldots, 100$. As with our other examples we have purposely chosen an approximating distribution that is narrower than the target so as to make the problem more difficult.

Figure 6 shows the sampling distributions of the estimated integral for the three methods, for each value of p. Again PSIS and TIS have much smaller variance than raw importance sampling, and PSIS has much smaller bias than TIS. All methods eventually fail as the number of dimensions increases and even a small difference in the distributions is amplified.

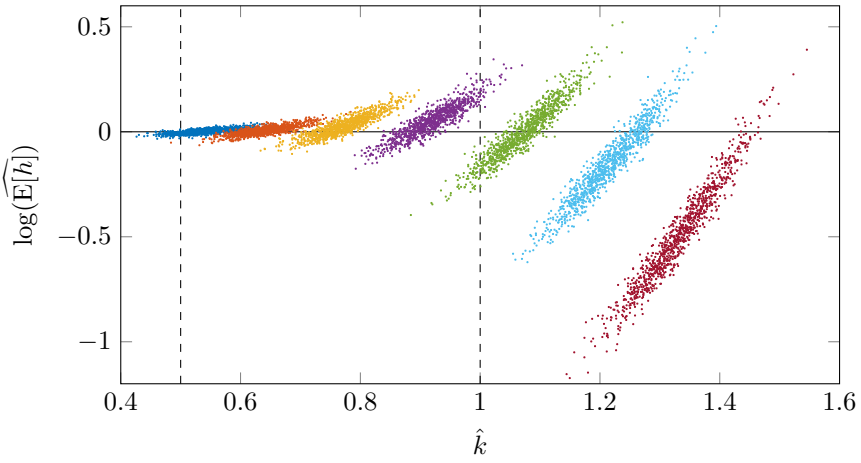

Figure 5: *Toy example 3: Scatterplot for Pareto smoothed importance sampling estimates of* $\mathrm{E}(h)$ *versus estimated tail shape* $\hat{k}$ *in 1000 repeated simulations, with different colors corresponding to the shift parameter* μ *set to* $0, 0.5, \ldots, 2.5, 3$.

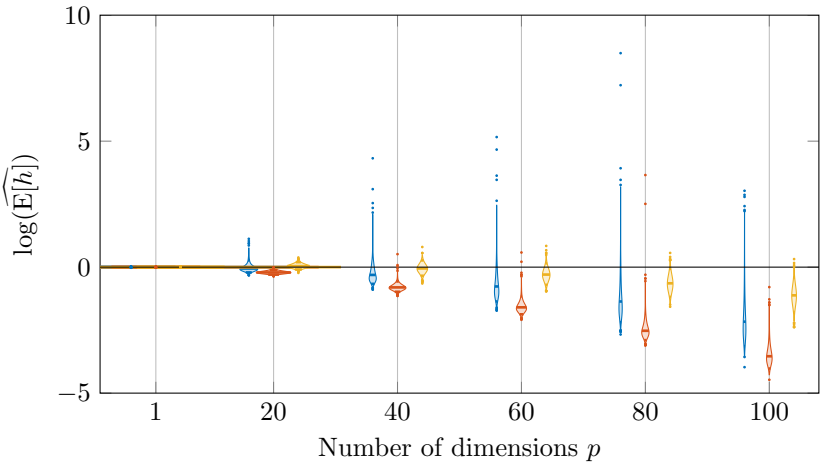

Figure 6: *Toy example 4: Violin plot for raw importance sampling (blue), truncated importance sampling (red), and Pareto smoothed importance sampling (yellow) estimates of* $\mathrm{E}(h)$ *in 1000 repeated simulations (each based on 16,000 draws from the approximate density). We make the conditions increasingly more challenging by setting the dimensionality parameter first to 1, then to 20, 40, . . . , 100.*

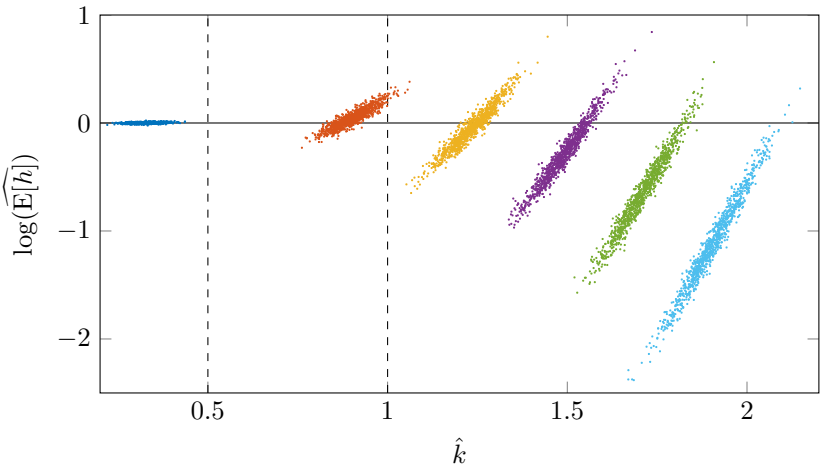

Figure 7: *Toy example 4: Scatterplot for Pareto smoothed importance sampling estimates of* $\mathrm{E}(h)$ *versus estimated tail shape* $\hat{k}$ *in 1000 repeated simulations with different colors corresponding to the numbers of dimensions* $p \in (1, 20, 40, 60, 80, 100)$.

Figure 7 shows the variation in the estimated tail shape parameter $\hat{k}$ and in the PSIS estimate. The variance and bias increases after $\hat{k} > 1/2$, but with PSIS it is possible to get only a small bias even when $\hat{k}$ is a bit larger than 1.

5. Practical examples

In this section we present two practical examples where the Pareto shape estimate $\hat{k}$ is a useful diagnostic and Pareto smoothed importance sampling clearly improves the estimates. In the first example PSIS is used to improve distributional approximation (split-normal) of the posterior of logistic Gaussian process density estimation model. In the second example PSIS is used for fast and reliable approximate leave-one-out cross-validation.

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 importance sampling.

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 $\mathcal{V}$ 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,

$$p(x) = \frac{\exp(f(x))}{\int_{\mathcal{V}} \exp(f(s))ds}, \tag{8}$$

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 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.

Riihimäki and Vehtari (2014) use importance sampling with a multivariate split Gaussian density as an approximation, following Geweke (1989). 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 Kong et al. (1994) estimate is based on the estimate of the variance of the weights and is not valid if the variance is infinite. Here we propose to use the Pareto shape parameter diagnostic instead and to use PSIS to stabilize the weights.

We repeated the density estimation of Galaxy data set[1] 4000 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. We obtain 8000 draws from the proposal distribution. As a baseline we used Markov chain Monte Carlo.

Figure 8 shows the log density estimate compared to MCMC. The plain split-normal without importance sampling performs significantly worse than MCMC. On average, the split-normal with importance sampling performs similarly to MCMC, but the computation takes only 1.3s compared to about half an hour with MCMC (a laptop with Intel Core i5-4300U CPU @ 1.90GHz x 4). However, IS sometimes produces worse results and occasionally even worse results than without IS. TIS works better but has one clearly deviating case. PSIS gives the most stable performance. In the experiment the average $\hat{k}$ was 0.52, indicating that the variance of the raw weights is infinite. For comparison, for a simple normal approximation without split scaling the average $\hat{k}$ was 0.58, illustrating that the $\hat{k}$ diagnostic can be also be used to evaluate the quality of the distributional 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

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

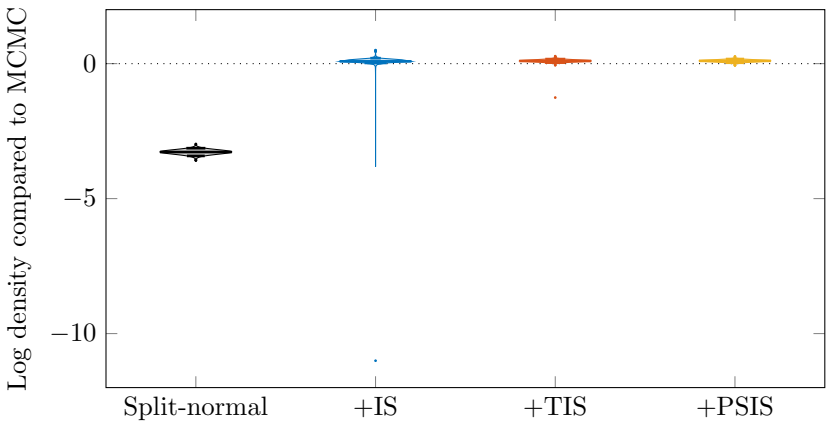

Figure 8: *Split-normal importance sampling posterior approximation for the logistic Gaussian process density estimate. Violin plot of split-normal without importance sampling (gray), regular importance sampling (blue), truncated importance sampling (red), and Pareto smoothed importance sampling (yellow) estimates of the log density in 4000 repeated simulations with different random seeds. The baseline is the log density computed with MCMC.*

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. 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 compute the necessary and 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.

We demonstrate the benefit of fast importance sampling leave-one-out cross-validation 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 and used Stan (Stan Development Team, 2015) to fit the model. As there are more covariates (411) than observations (283) we use a sparsifying hierarchical shrinkage prior (Piironen and Vehtari, 2015). 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

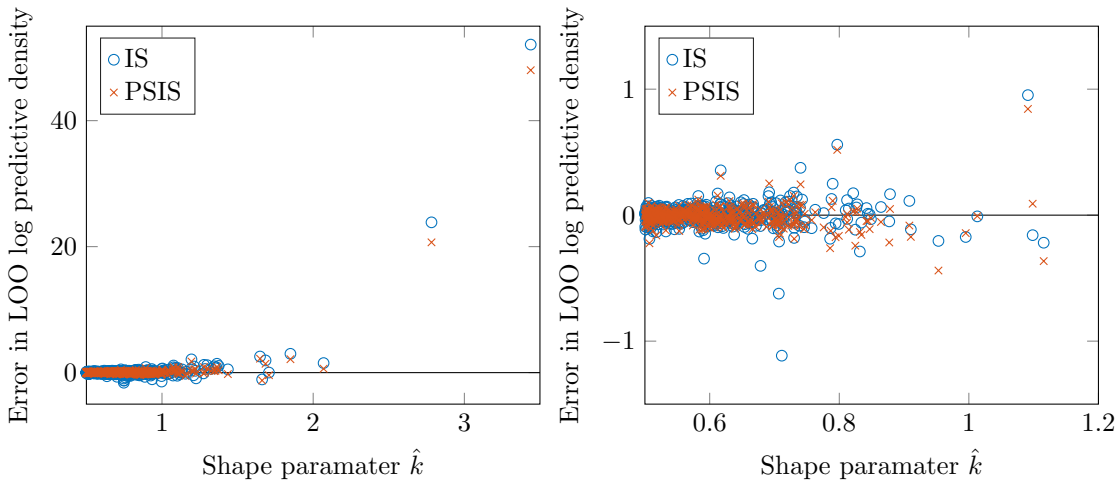

Figure 9: *Importance-sampling leave-one-out cross-validation: Error in LOO log predictive density (compared to exact LOO) from regular importance sampling (blue circles) and Pareto smoothed importance sampling (red crosses), and corresponding* $\hat{k} > 0.5$ *values. The plot on the left is for the Gaussian models and the plot on the right is for the Student-t models.*

minutes (desktop Intel Xeon CPU E3-1231 v3 @ 3.40GHz x 8), which is reasonable speed. For all 105 genes the computation then 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 $\hat{k} > 0.7$. Figure 9 shows the $\hat{k}$ values and error for IS-LOO and PSIS-LOO compared to exact LOO when $\hat{k} > 0.5$. We see that the magnitude the of errors increases as $\hat{k}$ increases and that the errors are small for $\hat{k} < 0.7$. For $0.5 < \hat{k} < 1$, compared to IS-LOO, PSIS-LOO has a root mean square error 38% smaller for Gaussian models and 27% smaller for Student-t models.

As is it more sensitive to outliers, the Gaussian model has many high $\hat{k}$ values and thus LOO distributions can be highly different from full posterior distributions. To improve upon PSIS-LOO we can make the exact LOO computations for any points corresponding to $\hat{k} > 0.7$. In this example there were 330 such cases for the Gaussian models and 80 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. In this case, reasonable results could also have been obtained by making the exact computation only when $\hat{k} > 1$, which would have reduced the time to less than 6 hours for all models.

Figure 10 shows the final results from PSIS-LOO+ (PSIS-LOO with exact computation for cases with $\hat{k} > 0.7$), comparing the Gaussian and Student-t model for each gene. It is easy to see that there are several genes with strong outlier observations in the data. The Student-t model is never worse and the estimated effects of microRNAs using the Gaussian model are likely to be bad for many of the genes in the data.

The previously mentioned `loo` R package provides an implementation of PSIS-LOO and

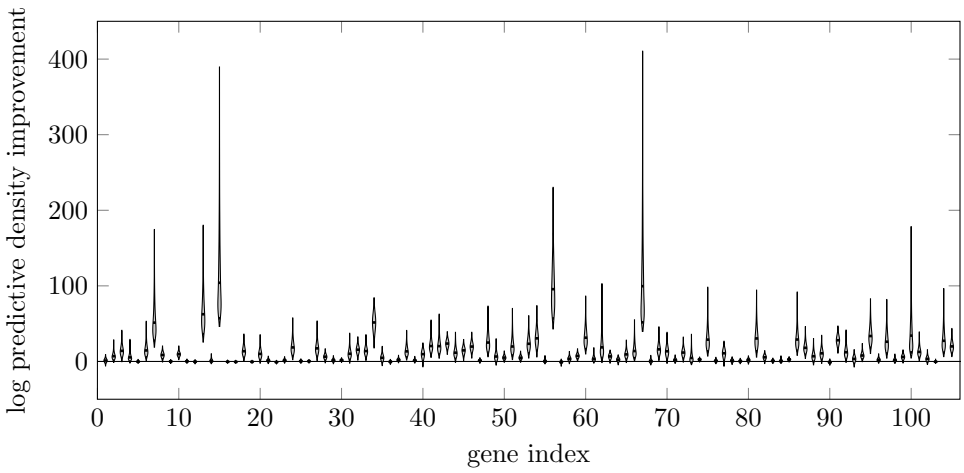

Figure 10: *Importance-sampling leave-one-out cross-validation: Violin plots for estimated log predictive density improvement of the Student-t model over the Gaussian model for all 105 genes using PSIS-LOO+.*

many examples are reported in Vehtari et al. (2016b), which focuses strictly on the use of PSIS for approximate leave-one-out cross-validation.

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 for big-data problems, and then there is a desire to adjust approximations to better match target distributions. It is well known that importance-weighted estimates are unstable if the weights have high variance.

We have found that one can reduce both bias and variance of importance sampling estimates using a stabilizing transformation that we call Pareto smoothed importance sampling (PSIS) in which the highest weights are replaced by expected quantiles from a generalized Pareto distribution fit to the data (that is, to the weights from the available simulations). 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 autodifferentiated 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.

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

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.

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 (with discussion). 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.

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

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

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*.

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.

Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using Stan+R. *arXiv:1508.02502*.

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*.

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.

Stan Development Team (2015). *Stan modeling language: User's guide and reference manual*. Version 2.9.0, `http://mc-stan.org/`.

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. (2016a). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, R package version 0.1.6. `https://github.com/stan-dev/loo`.

Vehtari, A., Gelman, A., and Gabry, J. (2016b). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. First online doi:10.1007/s11222-016-9696-4.

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.