Stable Limit Laws for Marginal Probabilities from MCMC Streams: Acceleration of Convergence

Robert L. Wolpert*
Institute of Statistics and Decision Sciences
Duke University, Durham NC 27708-0251
02-00†

Revised April 12, 2002

Abstract

In the Bayesian paradigm the marginal probability density function at the observed data vector is the key ingredient needed to compute Bayes factors and posterior probabilities of models and hypotheses. Although Markov chain Monte Carlo methods have simplified many calculations needed for the practical application of Bayesian methods, the problem of evaluating this marginal probability remains difficult. Newton and Raftery discovered that the harmonic mean of the likelihood function along an MCMC stream converges almost-surely but very slowly to the required marginal probability density. In this paper examples are shown to illustrate that these harmonic means converge in distribution to a one-sided stable law with index between one and two. Methods are proposed and illustrated for evaluating the required marginal probability density of the data from the limiting stable distribution, offering a dramatic acceleration in convergence over existing methods.

Key words: Bayes factors, domain of attraction, harmonic mean, Markov chain Monte Carlo, model mixing.

*Supported by U.S. E.P.A. grant CR822047-01-0
†Available at url: `ftp://ftp.isds.duke.edu/pub/WorkingPapers/02-00.ps`

1 Introduction

In the Bayesian paradigm hypotheses are tested and models are selected by evaluating their posterior probabilities or, when opinions differ about what should be the hypotheses' or models' prior probabilities, Bayes factors are used to represent the relative support offered by the data for one hypothesis or model against another. In Bayesian model mixing, predictions and inference are made amid uncertainty about the underlying model by weighting different models' predictions by their posterior probabilities. In all these cases a key ingredient is the marginal probability density function $f_m(x)$ at the observed data vector x, for each model m under consideration.

Although Markov chain Monte Carlo methods have broadened dramatically the class of problems that can be solved numerically using Bayesian methods, the problem of evaluating $f_m(x)$ remains difficult. Newton and Raftery (1994) discovered that the sample means of the inverse likelihood function $f_m(x|\theta_i)^{-1}$ along an MCMC stream $\{\theta_i\}$ will converge almost surely to the inverse $f_m(x)^{-1}$, so that $f_m(x)$ itself can be approximated by the harmonic mean of the likelihoods $f_m(x|\theta_i)$; unfortunately, as they also discovered, the partial sums S_n of the $f_m(x|\theta_i)^{-1}$ often do not obey a Central Limit Theorem, and in those cases S_n/n does not converge quickly. Many others have discovered the same phenomenon.

In this paper examples are shown to illustrate that $f_m(x|\theta_i)^{-1}$ will lie in the domain of attraction of a one-sided stable law of index $\alpha > 1$. Only in problems with precise prior information and diffuse likelihoods is $\alpha = 2$, where the Central Limit Theorem applies and the S_n have a limiting Gaussian distribution, with sample means converging at rate $n^{-1/2}$; with more sample information (or less prior information) the limit law is stable of index close to one, and slow convergence at rate $n^{-1/\alpha}$.

2 Stable Laws

Early this century Paul Lévy (1925) proved that the only possible limiting distributions for recentered and rescaled partial sums $S_n = \sum_{j \le n} Y_j$ of independent identically-distributed random variables are the *stable laws* $(S_n - a_n)/b_n \Rightarrow Z$ with characteristic functions (see Cheng and Liu, 1997)

$$\mathsf{E}\left[e^{i\omega Z}\right] = \exp\left(i\delta\omega - |\gamma\omega|^\alpha + i\beta \tan\frac{\pi\alpha}{2}\left\{|\gamma\omega|^\alpha \operatorname{sgn}\omega - \gamma\omega\right\}\right) \qquad (1)$$

for some index $\alpha \in (0,1) \cup (1,2]$, skewness $-1 \le \beta \le 1$, scale $\gamma > 0$, and location $-\infty < \delta < \infty$ (a similar formula holds for $\alpha = 1$); for $1 < \alpha < 2$ and $\beta = 1$, the cases of interest to us below, Z is integrable and the characteristic function can be written in Lévy-Khinchine form as

$$\mathsf{E}\,[e^{i\omega Z}] = \exp\big(i\delta\omega + C\int_0^\infty \big\{e^{i\omega u} - 1 - i\omega u\big\} u^{-1-\alpha}\,du\big) \tag{2}$$

with $C = 2\gamma^\alpha \sin\frac{\pi\alpha}{2}\Gamma(\alpha+1)/\pi$. We will face the problem of estimating the mean $\mathsf{E}\,[Z] = \delta - \beta\gamma\tan\frac{\pi\alpha}{2} \approx \delta + \frac{2\beta\gamma}{\pi(\alpha-1)}$ for $\alpha \approx 1$ from data.

This location/scale family has a density function $\gamma^{-1} f_{\alpha,\beta}\big(\frac{x-\delta}{\gamma}\big)$, but only in a few special cases is $f_{\alpha,\beta}(x)$ known in closed form; still it and its distribution function can be evaluated numerically by inverting the Fourier transform,

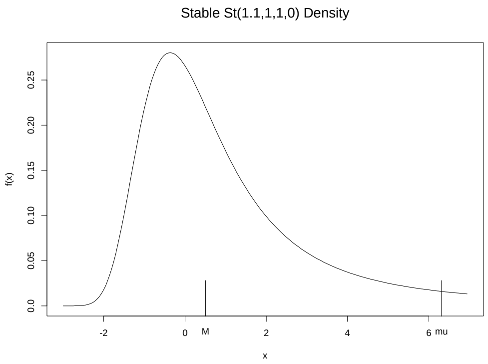
$$\begin{aligned} f_{\alpha,\beta}(x) &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\omega x - |\omega|^\alpha + i\beta\tan\frac{\pi\alpha}{2}[|\omega|^\alpha \operatorname{sgn}\omega - \omega]}\,d\omega \\ &= \frac{1}{\pi}\int_0^\infty e^{-\omega^\alpha}\cos\big(\omega x - \beta\tan\frac{\pi\alpha}{2}[\omega^\alpha - \omega]\big)\,d\omega \\ F_{\alpha,\beta}(x) &= c + \frac{1}{\pi}\int_0^\infty e^{-\omega^\alpha}\sin\big(\omega x - \beta\tan\frac{\pi\alpha}{2}[\omega^\alpha - \omega]\big)\omega^{-1}\,d\omega. \end{aligned} \tag{3}$$

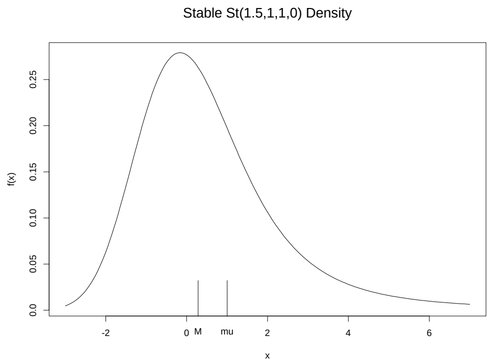
If the distribution of the $\{Y_j\}$ has finite variance (or tail probabilities that fall off fast enough that $y^2\,\mathsf{P}[|Y_j| > y] \to 0$) the central limit theorem applies and the limit must be normal. The limit will be stable of index $\alpha < 2$ if instead $\mathsf{P}[|Y_j| > y] \sim k y^{-\alpha}$ as $y \to \infty$ for some $0 < \alpha < 2$; it will be one-sided stable if also $\mathsf{P}[Y_j < -y]/\mathsf{P}[Y_j > y] \to 0$, whereupon $\beta = 1$ and the density function is given by

$$f_{\alpha,1,\gamma,\delta}(x) = \frac{1}{\gamma\pi}\int_0^\infty e^{-\omega^\alpha}\cos\left[\frac{\omega(x-\delta)}{\gamma} - \tan\frac{\pi\alpha}{2}(\omega^\alpha - \omega)\right]d\omega \tag{4}$$

$$= \frac{1}{\gamma\pi}\int_0^\infty e^{-\omega}\cos\left[\frac{\omega(x-\delta)}{\gamma} + \frac{2}{\pi}\,\omega\log\omega\right]d\omega \quad \text{if } \alpha = 1, \tag{5}$$

with approximate median $F^\alpha_{50} \approx \delta$ and quartiles F^α_{25}, $F^\alpha_{75} \approx \delta \mp \gamma$. For example, here are the densities for $\alpha = 1.1$ and $\alpha = 1.5$, with unit scale $\gamma = 1$ and zero location $\delta = 0$:





Each has the median (near $x = 0$) and the mean (further to the right) labeled (with "M" and "mu", respectively) and indicated with short vertical strokes; evidently the right tails are quite heavy, more so for $\alpha \approx 1$, making it difficult to estimate δ or $\mathsf{E}\,[Y_j]$ from sample averages. In principle it would be possible to estimate α, γ and δ by constructing averages Z_i of sufficiently many of the Y_j that the $\{Z_j\}$ are approximately independent one-sided stable variables of index α and center δ, then applying nonlinear optimization methods to minimize the negative log likelihood

$$-\log L(\alpha,\gamma,\delta) = \sum_{i=1}^{m} -\log f_{\alpha,1,\gamma,\delta}(z_i),$$

evaluated by applying numerical integration to (4); this computation of the maximum likelihood estimators $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\delta}$ entails an enormous computational burdon for sample sizes m large enough to provide clear evidence about δ, since the stable distribution has no nontrivial sufficient statistics or simple formulas for the MLE's. Instead we follow the following prescription of McCulloch (1986), based on the distributional quantiles x_p^α of the standard fully-skewed stable distribution of index α, satisfying $\mathsf{P}[X \le x_p^\alpha] = p$, and the sample quantiles $\hat{x}_p$ from the data, for $p \in \{.05, .25, .50, .75, .95\}$:

- Find the index α for which

  $$\frac{\hat{x}_{.95} - \hat{x}_{.05}}{\hat{x}_{.75} - \hat{x}_{.25}} \equiv \hat{\nu}_\alpha = \nu_\alpha \equiv \frac{x_{.95}^\alpha - x_{.05}^\alpha}{x_{.75}^\alpha - x_{.25}^\alpha};$$

  note that the quantity ν_α depends only on the kurtosis (and hence on α) but not on either scale or location (hence γ or δ).

- For this α, find the scale $\gamma = (\hat{x}_{.75} - \hat{x}_{.25})/(x^\alpha_{.75} - x^\alpha_{.25})$ for which

$$\frac{\hat{x}_{.75} - \hat{x}_{.25}}{\gamma} \equiv \hat{\nu}_\gamma = \nu_\gamma \equiv \frac{x^\alpha_{.75} - x^\alpha_{.25}}{1};$$

  note that the quantity $\nu_\gamma = (x^\alpha_{.75} - x^\alpha_{.25})$ depends only on the kurtosis and scale (hence on α and γ) but not on location (hence δ).

- For this α and γ, find the location $\delta = \hat{x}_{.50} - \gamma x^\alpha_{.50}$ for which

$$\frac{\delta - \hat{x}_{.50}}{\gamma} \equiv \hat{\nu}_\delta = \nu_\delta \equiv \frac{0 - x^\alpha_{.50}}{1}.$$

Thus $\mathsf{E}\,[Y] = \delta - \beta\gamma \tan\frac{\pi\alpha}{2}$ may be estimated from the sample quantiles. A partial table of the necessary indices (along with $\zeta \equiv \nu_\delta + \tan\frac{\pi\alpha}{2}$) is given in (McCulloch, 1986).

2.1 Example 1

Let $X_j \sim \mathsf{Ga}(a, \lambda)$ be independent draws from a Gamma distribution with shape parameter a and rate parameter λ, and set $Y_j \equiv \exp(X_j)$; then Y_j satisfies

$$\begin{aligned}
\mathsf{E}\,[Y_j{}^p] &= (1 - p/\lambda)^{-a} < \infty \qquad \text{if } p < \lambda; \\
\mathsf{P}[Y_j > y] &= \mathsf{P}[\lambda X_j > \lambda \log y] \\
&= \Gamma(a, \lambda \log y)/\Gamma(a) \\
&\approx (\lambda \log y)^{a-1} \exp(-\lambda \log y)[1 + O(1/\lambda \log y)]/\Gamma(a) \\
&\sim ky^{-\lambda} \qquad \text{as } y \to \infty,
\end{aligned}$$

where $\Gamma(a, x)$ denotes the incomplete Gamma function (Abramowitz and Stegun, 1964, §6.5.32). If $\lambda > 2$ then Y_j has finite variance and lies in the normal domain of attraction, while for $\lambda < 2$ the limit is one-sided stable of index $\alpha = \lambda$.

2.2 Example 2

Now let Z_j be independent normally-distributed random variables with mean $\mu \in \mathbb{R}$ and variance $V > 0$, and set $Y_j = \exp(c\,Z_j{}^2)$. If $\mu = 0$ then $c\,Z_j{}^2 \sim c\,V\chi^2_1$ is Gamma distributed with $a = 1/2$ and $\lambda = 1/2cV$, so for $V > 1/4c$

the limiting distribution is again one-sided stable of index $\alpha = 1/2cV$; even for $\mu \neq 0$ the same limit follows from the calculation

$$\begin{aligned}
\mathsf{P}[Y_j > y] &= \mathsf{P}\big[|Z_j| > \sqrt{(\log y)/c}\big] \qquad \Big(\text{set } \eta \equiv \sqrt{(\log y)/c}\Big) \\
&= \Phi\big(\frac{-\eta - \mu}{\sigma}\big) + \Phi\big(\frac{-\eta + \mu}{\sigma}\big) \\
&\approx \frac{\sigma}{\sqrt{2\pi}} \left[\frac{\exp\big(-(\eta+\mu)^2/2V\big)}{\eta+\mu} + \frac{\exp\big(-(\eta-\mu)^2/2V\big)}{\eta-\mu} \right] \\
&\sim k\, e^{-\eta^2/2V} = k\, y^{-1/2cV} \qquad \text{as } y \to \infty,
\end{aligned}$$

where $\Phi(z)$ is the cumulative distribution function for the standard normal distribution (Abramowitz and Stegun, 1964, §26.2.13), so again Y_j lies in the domain of attraction of the one-sided stable distribution of index $\alpha = 1/2cV$.

2.3 Example 3 (Main Example)

Now let $X_k \overset{\text{iid}}{\sim} \mathsf{No}(\theta, \sigma^2)$ with known variance $\sigma^2 > 0$ but uncertain mean θ. Two models are entertained: M_0, under which $\theta \sim \mathsf{No}(\mu_0, \tau_0^2)$, and M_1, under which $\theta \sim \mathsf{No}(\mu_1, \tau_1^2)$ (the point null hypothesis with $\tau_0 = 0$ is included). Thus the joint and marginal densities for the sufficient statistic $\bar{x}$ from a vector of n observations $\{X_k\}$ under model $m \in \{0, 1\}$ are:

$$\begin{aligned}
\pi_m(\theta, \bar{x}) &= (2\pi\sigma^2/n)^{-1/2}(2\pi\tau_m^2)^{-1/2} e^{-n(\bar{x}-\theta)^2/2\sigma^2 - (\theta-\mu_m)^2/2\tau_m^2} \\
f_m(\bar{x}) &= [2\pi(\sigma^2/n + \tau_m^2)]^{-1/2} e^{-(\bar{x}-\mu_m)^2/2(\sigma^2+\tau_m^2)}
\end{aligned}$$

and the posterior probability of model M_0 and the posterior odds against M_0, under prior probabilities $\pi[M_0] = \pi_0$ and $\pi[M_1] = \pi_1$, are

$$\mathsf{P}[M_0 \mid \vec{x}] = \frac{\pi_0 f_0(\bar{x})}{\pi_0 f_0(\bar{x}) + \pi_1 f_1(\bar{x})} \qquad \frac{\mathsf{P}[M_1 \mid \vec{x}]}{\mathsf{P}[M_0 \mid \vec{x}]} = \frac{\pi_1}{\pi_0}\frac{f_1(\bar{x})}{f_0(\bar{x})}.$$

Thus the key for making inference and for computing the Bayes factor $B = f_1(\bar{x})/f_0(\bar{x})$ is computing the marginal density function

$$f(\bar{x}) = [2\pi(\sigma^2/n + \tau^2)]^{-1/2} e^{-(\bar{x}-\mu)^2/2(\sigma^2/n+\tau^2)}$$

at the observed data point $\bar{x}$.

At equilibrium any Markov Chain Monte Carlo (MCMC) procedure will generate random variables $\theta_j \sim \pi(\theta|\bar{x})$ from the posterior distribution $\pi(\theta|\bar{x}) = \pi(\theta,\bar{x})/f(\bar{x})$, so for any posterior-integrable function $g(\theta)$ the ergodic theorem ensures

$$\mathsf{E}\,[g(\theta) \mid \bar{x}] = \frac{\int g(\theta)\, f(\bar{x} \mid \theta)\, \pi(\theta)\, d\theta}{f(\bar{x})} = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} g(\theta_j).$$

Newton and Raftery (1994) reasoned that the posterior mean of the random variables $Y_j \equiv 1/f(\bar{x} \mid \theta_j)$ would be

$$\mathsf{E}\,\left[f(\bar{x} \mid \theta)^{-1} \mid \bar{x}\right] = \frac{\int \pi(\theta)\, d\theta}{f(\bar{x})} = \frac{1}{f(\bar{x})},$$

leading to the Harmonic Mean Estimator

$$f(\bar{x}) = \lim_{n\to\infty} \frac{1}{\bar{Y}_n} = \lim_{n\to\infty} \frac{n}{\sum_{j=1}^{n} 1/f(\bar{\theta}_j \mid \bar{x})}. \tag{6}$$

For any proper prior the limit in (6) converges almost surely. It is our goal to show that the convergence can be *very* slow. Indeed,

$$Y_j = 1/f(\bar{x} \mid \theta_j) = (2\pi\sigma^2/n)^{1/2} e^{n(\bar{x}-\theta_j)^2/2\sigma^2} \propto \exp\left(\frac{n}{2\sigma^2}\,(\theta_j - \bar{x})^2\right)$$

is the same as Example 2 above, with $c = n/2\sigma^2$ and V the conditional variance of θ_j given $\bar{x}$, $V = (n/\sigma^2 + 1/\tau^2)^{-1}$, so the limiting distribution is normal and the central limit theorem applies if

$$\alpha = \frac{1}{2cV} = \frac{1}{2(n/2\sigma^2)(n/\sigma^2 + 1/\tau^2)^{-1}} = \left(1 + \sigma^2/n\tau^2\right)$$

exceeds 2, *i.e.*, if the prior variance τ^2 is less than the sampling variance σ^2/n). Otherwise, if $\sigma^2/n < \tau^2$, then the limiting distribution of S_n is one-sided stable with index $\alpha \in (1,2)$. As the sample size n increases, so that the data contain substantially more information than the prior, then we are driven inexorably to the stable limit, with index α just slightly above one.

Since $\mathsf{E}\,[Y_j] = 1/f(x)$ Lévy's limit theorem asserts that

$$\frac{S_n - n/f(x)}{n^{1/\alpha}} \Rightarrow Z \tag{7}$$

converges in distribution to a stable of index $\alpha = 1 + \sigma^2/n\tau^2$, $\beta = 1$, and mean 0, whence

$$S_n/n \;\approx\; 1/f(x) + Z\, n^{1/\alpha - 1};$$

evidently $S_n/n \to 1/f(x)$ as $n \to \infty$, but the convergence is only at rate $n^{-\epsilon}$ for $\epsilon = 1 - 1/\alpha = (1 + n\tau^2/\sigma^2)^{-1}$ and moreover the errors have thick-tailed distributions with infinite moments of all orders $p \geq \alpha$.

2.4 Example 4 (Bernstein/von Mises)

Under suitable regularity conditions *every* posterior distribution is asymptotically normally distributed, and every likelihood function is asymptotically normal, so the stable limiting behaviour of the preceeding section can be expected in nearly all efforts to apply the Harmonic Mean Estimator to compute Bayes factors for large sample sizes and relatively vague prior information.

3 Improving the Estimate

Instead of estimating $1/f(x) \approx S_n/n$ directly from ergodic averages, we may try to estimate the parameters $\alpha, \gamma_n, \delta_n$ for the fully-skewed stable $\bar{Y}_n = S_n/n$ using the quantile-based method of McCulloch (1986) or the recent Bayesian approach of [check citation], whereupon we can estimate

$$1/f(x) = \mathsf{E}\,[\bar{Y}_n] = \delta_n - \gamma_n \tan\frac{\pi\alpha}{2}.$$

References

Abramowitz, M. and Stegun, I. A., eds. (1964) *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, vol. 55 of *Applied Mathematics Series*. Washington, D.C.: National Bureau of Standards.

Cheng, R. C. H. and Liu, W. B. (1997) A continuous representation of the family of stable law distributions. *Journal of the Royal Statistical Society, Series B (Methodological)*, **59**, 137–145.

Lévy, P. (1925) *Calcul des Probabilités*. Paris: Gauthier-Villars.

McCulloch, J. H. (1986) Simple consistent estimators of stable distribution parameters. *Communications in Statistics (Part B)–Simulation and Computation*, **15**, 1109–1136.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B (Methodological)*, **56**, 3–48.