MCMC and EM-based methods for inference in heavy-tailed processes with α-stable innovations

Simon Godsill*
Signal Processing Group, Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
sjg@eng.cam.ac.uk
tel. +44-1223-332604, fax +44-1223-332662

April, 1999

Abstract

In this paper we present both stochastic and deterministic iterative methods for inference about random processes with symmetric stable innovations. The proposed methods use a scale mixtures of normals (SMiN) representation of the symmetric stable law to express the processes in conditionally Gaussian form. This allows standard procedures for dealing with the Gaussian case to be re-used directly as part of the scheme. In contrast with other recently published work on the topic, we propose a novel hybrid rejection sampling method for simulating the scale parameters from their full conditional distributions, making use of asymptotic approximations for the tail of a positive stable distribution when rejection rates are too high. This hybrid approach potentially leads to improved performance compared with straightforward rejection sampling or Metropolis-Hastings (M-H) approaches. The methods can be applied to any model with symmetric stable terms, but we illustrate their application to linear models and present simulations for AR time series with stable innovations.

1 Introduction

In recent years there has been considerable interest in the Signal Processing community in α-stable distributions. There is now a large literature within signal processing, which we do not have space to list, but Nikias and Shao [1] and Kuruoglu [2] give a good review coverage. Theoretical details are well covered in Feller [3], Samorodnitsky and Taqqu [4] and Zolotarev [5]. There is physical justification for the occurence of heavy-tailed stable law distributions in fields as diverse as wireless communications, sonar and econometrics (where pioneering work was done by Mandelbrot).

*This work partially funded by the Nuffield Foundation - NUF-NAL

Exact inference for stable processes with small datasets is generally quite troublesome, owing largely to the non-analytic form of the density function. However, it is straightforward to draw random samples from stable law distributions (see section 2.3). This opens up the possibility for the use of simulation-based methods of inference. In this paper we consider Markov chain Monte Carlo (MCMC) and Monte Carlo Expectation-maximisation (MCEM) approaches to inference in the presence of symmetric stable noise. Relevant work in the MCMC field includes that of Buckle [6] who shows how to perform MCMC inference for all the parameters of stable noise, and Quiou and Ravishanker [7] who show how to extend Buckle's method to time series problems. Some very recently published material which has come to our attention from the econometrics literature by Tsionas [8] covers some of the same ground as our work (which first appeared as [9]), but the novel hybrid rejection sampling scheme and the Monte Carlo EM approach proposed here are new to the best of our knowledge and provide useful additional developments and insights.

2 Definitions: stable random variables

2.1 Characteristic function

A random variable X is drawn from a Stable Law distribution $f_{\alpha,\delta}(a,b)$ iff its characteristic function is given by [3, p.570, p.581], [4]:

$$\varphi(\omega) = \exp(a\,\psi_{\alpha,\delta}(\omega) + i\,b\omega)$$

where:

$$\psi_{\alpha,\delta}(\omega) = \begin{cases} -\left|\omega\right|^{\alpha}\left[1 - i\mathrm{sign}(\omega)\delta\tan\left(\frac{\pi\alpha}{2}\right)\right], & (\alpha \neq 1) \\ -\left|\omega\right|^{\alpha}\left[1 + i\mathrm{sign}(\omega)\delta\log\left(\left|\omega\right|\right)\right], & (\alpha = 1) \end{cases} \tag{1}$$

The parameters of the distribution are defined as follows:

- $\alpha \in (0, 2]$ is the *characteristic exponent*, which determines the tail thickness of the distribution
- $\delta \in [-1, 1]$ is a skewness parameter. $\delta = 0$ is the family of symmetric stable distributions. Well known special cases are $\alpha = 2$ (Gaussian) and $\alpha = 1$ (Cauchy). $\delta = \pm 1$ defines the positive/negative stable family. The only positive stable distribution whose density function can be written in closed form is for $\alpha = 1/2$, the Lévy distribution. Apart from these three cases and the equivalent negative random variable to the Lévy distribution no other densities from the stable family are currently known in closed form.
- $a > 0$ is a scale parameter and b is a location parameter.

2.2 Positive stable laws

The family of stable distributions with $\delta = 1$, $\alpha < 1$ define random variables with support $(0, \infty)$. The Laplace transform of all such distributions having $a = 1$ and $b = 0$ is $\exp\left(-\frac{1}{\cos(\pi\alpha/2)}\, s^{\alpha}\right)$ [3, p. 448], where the scale factor $\cos(\pi\alpha/2)$ ensures scale equivalence with (1). Laplace transform properties or Taylor expansions of the characteristic function show that tail behaviour of the density as $x \to \infty$ tends to $1/x^{\alpha+1}$ [3, p.448 & p.582]. This result will be used later in the hybrid rejection sampler scheme.

2.3 Random variate generation

It is straightforward to draw random samples from stable distributions using the Chambers-Mallows-Stuck (CMS) method [10, 11]. Here a $f_{\alpha,\delta}$ random variable X is generated via a non-linear transformation of a uniform variate V and an exponential variate W. The details are omitted for the sake of brevity.

2.4 Property - products of stable random variables

A property which will be used in the following sections is defined as follows (see [3, p. 176]).

> Let X and $Y > 0$ be independent random variables with $X \sim f_{\alpha,\delta}$ and $Y \sim f_{\beta,1}$. Then $XY^{\frac{1}{\alpha}}$ is stable with exponent $\alpha\beta$.

In other words, the product of an α-stable random variable X with a positive β-stable variable Y raised to the power $1/\alpha$, is itself stable with exponent $\alpha\beta$. We will utilise the case when $\alpha = 2$ (i.e. the Gaussian) and $\beta < 1$, which implies that a SMiN representation exists for the symmetric stable law. We will see how to use this result for inference in the next section.

3 Inference with symmetric stable distributions

Suppose a noise process $\{v_t\}$ is to be modelled as zero-location i.i.d. symmetric stable with scale parameter σ:

$$v_t/\sigma \sim f_{\alpha,0} \tag{2}$$

This is not straightforwardly incorporated into standard inference schemes, both because of the non-analytic form of the density function which will prevent likelihood functions from being evaluated, and because of the non-Gaussianity, which will make implementation more troublesome. The trick to solving this problem is to rewrite the noise model (2) in an exactly equivalent way, making use of the product property discussed above in section 2.4:

$$v_t \sim \mathrm{N}(0, \sigma^2\lambda_t), \qquad \lambda_t \sim f_{\alpha/2,1} \tag{3}$$

We have now included an auxiliary random variable λ_t in such a way that (2) is satisfied using the result in section 2.4. The reason why this will aid inference procedures is that, conditionally upon λ_t and σ, v_t is Gaussian. Hence it will be seen that iterative procedures such as MCMC and EM can readily be applied, incorporating any standard techniques previously developed for the Gaussian case.

3.1 Example: linear time series model

The methods proposed can be applied to any model for which MCMC methods exist in the Gaussian case. Take as an example the linear model which can be expressed at time t as $y_t = \mathbf{g}_t\boldsymbol{\theta} + v_t$ and in vector form as:

$$\mathbf{y} = \mathbf{G}\boldsymbol{\theta} + \mathbf{v}$$

where $\mathbf{y}$ is the observed data vector, $\boldsymbol{\theta}$ is a vector of unknown parameters, $\mathbf{G}$ is a fixed basis matrix and $\mathbf{v}$ is a noise or error vector whose elements are assumed i.i.d. $f_{\alpha,0}(\sigma, 0)$. Using the reparameterisation of (3) the likelihood is obtained as:

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \sigma^2) \propto \exp\left(1/2\sigma^2\, ||\Sigma(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})||_2^2\right) \tag{4}$$

where Σ is a diagonal matrix with diagonal elements $(1/\lambda_1^{1/2}, \ldots, 1/\lambda_N^{1/2})$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N]$.

We focus here on a Bayesian approach utilising MCMC and MCEM methods, although it should be stressed that both methods, and in particular the MCEM method, are readily adapted to ML estimation. Assume a conjugate prior framework for the unknowns $\boldsymbol{\theta}$ and σ:

$$p(\boldsymbol{\theta}, \sigma^2) = \mathrm{N}(\boldsymbol{\theta}|\mu, \mathbf{C})\mathrm{IG}(\sigma^2|\alpha, \beta)$$

where $\mathrm{IG}(\alpha, \beta)$ denotes the inverted Gamma distribution with parameters α and β. The full conditional posteriors for θ and σ^2 will be required in the following sections, and can be obtained as:

$$p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{\lambda}, \mathbf{y}) = \mathrm{N}(\boldsymbol{\mu}', \mathbf{C}'), \quad p(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y}) = \mathrm{IG}(\alpha', \beta') \tag{5}$$

where

$$\mathbf{C}' = (\mathbf{G}^T\mathbf{G} + \sigma^2\mathbf{C}^{-1})^{-1}, \quad \boldsymbol{\mu}' = \mathbf{C}'(\mathbf{G}^T\mathbf{y} + \sigma^2\mathbf{C}^{-1}\boldsymbol{\mu})$$
$$\alpha' = \alpha + N/2, \quad \beta' = \beta + ||\Sigma(\mathbf{y} - \mathbf{G}\boldsymbol{\theta})||_2^2/2$$

The full conditional for $\boldsymbol{\lambda}$ will also be required. Individual elements λ_t are easily shown to be conditionally independent of one another given $v_t = y_t - \mathbf{g}_t\boldsymbol{\theta}$ and σ; hence we obtain that $p(\boldsymbol{\lambda}|\boldsymbol{\theta}, \sigma^2, \mathbf{y}) = \prod_{t=1}^N p(\lambda_t|v_t, \sigma)$, where:

$$p(\lambda_t|v_t, \sigma) \propto \mathrm{N}(v_t|0, \lambda_t\sigma^2) \times f_{\alpha/2,1}(\lambda_t) \tag{6}$$

3.2 MCMC implementation

There are many possible MCMC strategies for problems such as this, but for simplicity we illustrate the proposed approach with a Gibbs sampler. At the ith iteration, for $i = 1, \ldots, M$ a random draw is made with replacement from each of the unknowns in turn:

$$\boldsymbol{\theta}^i \sim p(\boldsymbol{\theta}|\sigma^{i-1}, \boldsymbol{\lambda}^{i-1}, \mathbf{y}) \qquad \sigma^i \sim p(\sigma|\boldsymbol{\theta}^i, \boldsymbol{\lambda}^{i-1}, \mathbf{y}) \tag{7}$$
$$\boldsymbol{\lambda}^i \sim p(\boldsymbol{\lambda}|\boldsymbol{\theta}^i, \sigma^i, \mathbf{y}) \tag{8}$$

Under mild conditions of irreducibility and aperiodicity [12], the samples drawn converge to the joint posterior distribution as $i \to \infty$, and hence Monte Carlo posterior estimation may be performed from the output once the chain has converged.

The first two steps (7) are of standard form in the linear model (see (5)) and straightforward to sample from, involving exactly the equivalent steps to a Gibbs sampler in a linear heteroscedastic Gaussian model with fixed $\boldsymbol{\lambda}$. The third step (8) is less straightforward, although still manageable, as is now shown.

Drawing a sample from the distribution of equation (6) involves the product of a normal likelihood with the positive stable prior distribution $f_{\alpha/2,1}$, the latter of which cannot in general be evaluated exactly. However, the CMS method (section 2.3) can be used to draw samples from the prior distribution $f_{\alpha/2,1}$. This can then be used as an envelope function for rejection sampling or as the proposal to a Metropolis-Hastings (M-H) algorithm, since in both cases the acceptance probability involves only likelihood evaluations and no evaluation of the prior itself. Thus we are able to perform MCMC even in a case such as this where the conditional distributions cannot be evaluated. The basic form of the rejection and M-H methods for sampling a single λ_t are now given.

3.2.1 Rejection sampling for λ_t

It is readily seen that the likelihood forms a valid rejection function as it is bounded from above:

$$\mathrm{N}(v_t|0, \lambda_t\sigma^2) \leq \frac{1}{\sqrt{2\pi v_t^2}} \exp(-1/2)$$

Hence a suitable rejection sampler is:

1. $\lambda_t \sim f_{\alpha/2,1}$
2. $u \sim \mathrm{U}(0, \frac{1}{\sqrt{2\pi v_t^2}} \exp(-1/2))$
3. If $u > \mathrm{N}(v_t|0, \lambda_t\sigma^2)$ goto 1

3.2.2 M-H sampler for λ_t[8]

If we choose $f_{\alpha/2,1}$ as the proposal to a M-H step with $p(\lambda_t|v_t, \sigma)$ as target distribution, then the acceptance probability is obtained as:

$$\alpha(\lambda_t; \lambda_t') = \min\left(1, \frac{\mathrm{N}(v_t|0, \lambda_t'\sigma^2)}{\mathrm{N}(v_t|0, \lambda_t\sigma^2)}\right)$$

3.2.3 Hybrid rejection sampler

The Metropolis scheme of the previous section has been implemented for econometric data in very recent independent work by Tsionas [8]. However, both of these schemes are likely to have problems in practice. The rejection sampler, while having the benefit of providing exact samples from the required conditional, suffers from very poor acceptance rates, especially when the mode of the prior is much different from the mode of the likelihood. The mean acceptance rate can be obtained as $f_{\alpha,0}(v_t)\sqrt{2\pi}v_t \exp(-0.5)$, a function with low values when v_t is very small or very large, as expected. Likewise, the M-H scheme suffers from very slow movement of components around the state space. We propose here an extension to the rejection sampling method which uses tail approximations to the prior and likelihood to give more effective sampling in difficult regions where the prior and likelihood are in conflict.

We can identify three possible regimes, as illustrated in figure 1, which correspond to various different locations of the likelihood (data) relative to the prior. In each of these regimes a different sampling strategy is proposed:

1. **Data much 'larger' than prior**, fig. 1(bottom). Here the upper tail of the prior and the likelihood dominate the conditional. Hence we may approximate the prior using its upper asymptotic tail approximation (see section 2.2):

$$\begin{aligned}\lim_{\lambda_t\to\infty} f(\lambda_t|v_t) &\propto \lim_{\lambda_t\to\infty}\{f_{\alpha/2,1}(\lambda_t)\} \times \mathrm{N}(v_t|0, \lambda_t\sigma^2) \\ &\propto \lambda_t^{-\alpha/2-1} \times \mathrm{N}(v_t|0, \lambda_t\sigma^2) \\ &= \mathrm{IG}(\lambda_t|(\alpha+1)/2, v_t^2/2)\end{aligned}$$

i.e. in this case the conditional may be approximated in the limit by an inverted gamma distribution, from which we may sample directly using standard methods.

2. **Data in accordance with prior**, fig. 1 (middle). When the prior and the likelihood are in approximate agreement, rejection sampling can be used since the rejection rate will be acceptably low.

3. **Data much 'smaller' than prior**, fig. 1 (top). Here the likelihood lies to the left of the prior. In this case the prior and the upper tail of the likelihood dominate the conditional probability, as shown on the dash-dotted curve. Even though the prior may be generating samples in the correct region, the acceptance probabilities can still be very low, as the prior lies in the far right tail of the likelihood. Since there is no significant probability mass far to the left of the prior we can simply truncate the likelihood at some value $\epsilon >> v_t^2$ and perform rejection sampling for $\lambda > \epsilon$ using a suitably renormalised likelihood function to achieve good acceptance rates.

Steps 1. and 3. involve approximations. However, the approximations will be good provided they are only applied when the likelihood lies well into the tails of the prior. We will see that a hybrid scheme that switches between the three possible methods performs very efficiently when compared with either the basic rejection sampler or M-H method.

3.3 MCEM implementation

The EM method maximises at each iteration the expectation of the joint log-posterior w.r.t. the nuisance parameters. $\boldsymbol{\lambda}$ is a nuisance parameter which we must integrate out in order to perform correct inference about $\boldsymbol{\theta}$ and σ. The ith iteration of the EM algorithm for this problem can be written as:

$$\boldsymbol{\theta}^i, \sigma^i = \underset{\boldsymbol{\theta},\sigma}{\text{argmax}}\{E[log(p(\mathbf{y}|\boldsymbol{\theta}, \sigma, \boldsymbol{\lambda}))] + log(p(\boldsymbol{\theta}, \sigma))\}$$

where the expectation is understood to be w.r.t. $p(\boldsymbol{\lambda}|\boldsymbol{\theta}^{i-1}, \sigma^{i-1}, \mathbf{y})$. For the stable noise model in question it is not possible to perform the expectation analytically. However, it is straightforward to make a Monte Carlo approximation using some of the ideas of the last section. The expectation can be simplified as:

$$\begin{aligned}&E[log(p(\mathbf{y}|\sigma, \boldsymbol{\lambda}, \mathbf{y}))] \\ &= C - \frac{N}{2}\log(2\pi\sigma^2) - \sum_{t=1}^{N} \frac{1}{2\sigma^2} E[1/\lambda_t](y_t - \mathbf{g}_t\boldsymbol{\theta})^2\end{aligned}$$

This expression is equal, within an additive constant C, to the log-likelihood of the conditionally Gaussian model (4), with $E[1/\lambda_t]$ replacing $1/\lambda_t$. Hence the maximisation step is identical to that of the Gaussian model and it can be seen that EM in this case is simply an iteratively reweighted least squares (IRLS) scheme, with weights at each iteration given by $E[1/\lambda_t]^{1/2}$. IRLS has previously been proposed for estimation of parameters of stable processes [2, 13] Here the required expectations can be estimated by a Monte Carlo average. In particular, if we draw samples $\boldsymbol{\lambda}^i$, $i = 1...M$ from $p(\boldsymbol{\lambda}|\boldsymbol{\theta}^{i-1}, \sigma^{i-1}, \mathbf{y})$ using the proposed hybrid rejection sampler then a Monte Carlo estimate for the expectation is obtained as $E[1/\lambda_t] \approx \frac{1}{M}\sum_{i=1}^{M} 1/\lambda_t^i$ Of course, this is a rather heavy-handed way to calculate the expectations, and we do not provide any simulations using this method, preferring the relative efficiency of MCMC in this case.

4 Results

Simulations have been carried out for autorgressive time series driven by symmetric stable innovations:

$$y_t = \sum_{p=1}^{P} a_p y_{t-p} + \sigma_e e_t, \qquad e_t \overset{i.i.d.}{\sim} f_{\alpha,0}(1, 0)$$

Experiments so far show that the Metropolis-Hastings sampler of section 3.2.2 is very slow moving for cases where the prior and likelihood for λ_t are in conflict, whereas the hybrid rejection sampler explores the state space much more freely. We will report on objective comparisons in future publications. One application of these methods is in modelling impulsive excitation noise in speech and audio. Figure 2 shows a pitched audio signal (top) to which an autoregressive model with $f_{1.1,0}$ innovations has been fitted using the hybrid rejection MCMC method. The resulting Monte Carlo estimate of e_t (bottom) shows clearly where the impulsive pitch pulses have occurred.

5 Discussion

In this paper we have described methods for performing inference in the presence of symmetric α-stable noise. The approach needs much more extensive evaluation, which will be reported on in future publications, but offers the potential for signficant improvements in short datasets over methods that rely on asymptotic results. A further advantage of this framework is the conditionally Gaussian structure that arises from the SMiN reparameterisation of the symmetric stable distribution. This means that existing methodology developed for the Gaussian case can be re-used for the symmetric stable case. In fact, it was observed that EM for these models would correspond to a particular form of the iteratively reweighted least squares approach to parameter estimation, and this relationship will be explored further.

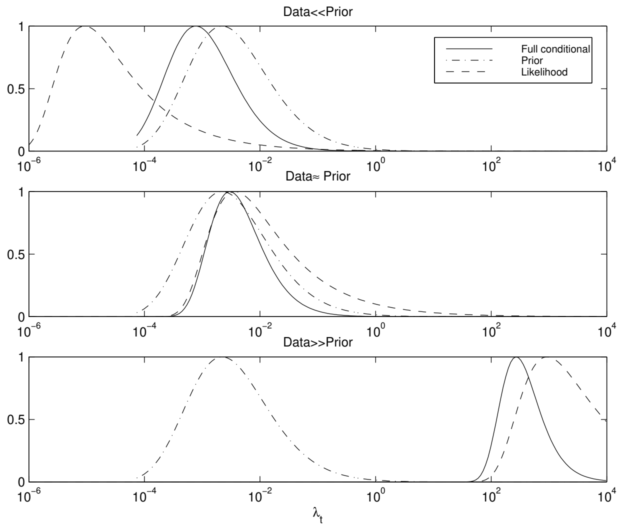

Figure 1: Illustration of the three possible regimes of prior/likelihood interaction

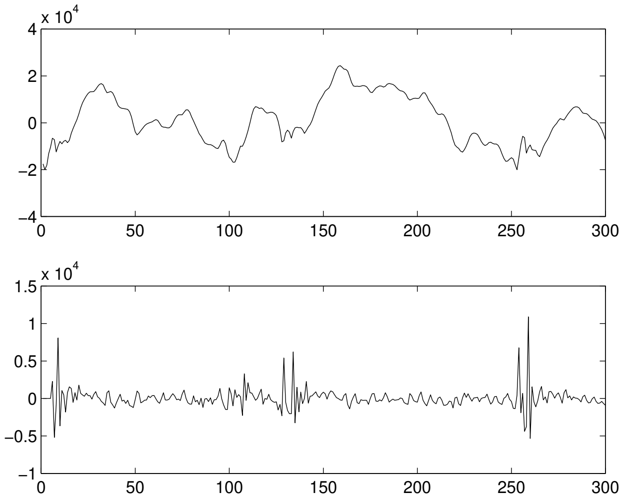

Figure 2: Pitched audio signal (top) and estimated innovations (bottom), showing clearly the impulsive elements in the innovation sequence

We have discussed the application of the methods to very simple linear parameter inference problems. However, the nature of MCMC methods allows the new methods to be incorporated readily as components of much more complex models involving non-linearity, non-stationarity, unobserved states and model selection. Finally, there has been no mention of uncertainty in the value of α, the characteristic exponent of the distribution. This can be handled within the same scheme using the methods of Buckle [6] or Tsionas [8], although we do not report any results here.

References

[1] C. L. Nikias and M. Shao. *Signal Processing with α-Stable Distributions and Applications*. John Wiley and Sons, 1995.

[2] E.C. Kuruoglu. *Signal processing in α-stable environments: a least l_p-norm approach*. PhD thesis, Cambridge University Engineering Department, 1998.

[3] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. New York:Wiley, second edition, 1971.

[4] G. Samorodnitsky and M.S. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*. New York:Chapman and Hall, 1994.

[5] V. Zolotarev. *One-dimensional stable distributions*. American Mathematical Society, Providence, RI, 1986.

[6] D.J. Buckle. Bayesian inference for stable distributions. *Journal of American Statistical Association*, 90(430):605–613, 1995.

[7] Z. Quiou and N. Ravishanker. Bayesian inference for time series with stable innovations. *Journal of Time Series Analysis*, 19(2):235–250, March 1998.

[8] E.G. Tsionas. Monte carlo inference in econometric models with symmetric stable disturbances. *J. Econometrics*, 88:365–401, 1999.

[9] S.J. Godsill. Inference and modelling for time series with symmetric stable disturbances. Joint Statistical Meetings, Special Session on Bayesian Time Series, Dallas.

[10] J.M. Chambers, C.L. Mallows, and B.W. Stuck. A method for simulating stable random variables. *Journal of American Statistical Association*, 71:340–344, 1976.

[11] R. Weron. On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. Technical report, Hugo Steinhaus Center for Stochastic Methods, Wroclaw, 1996.

[12] L. Tierney. Markov chains for exploring posterior disiributions (with discussion). *The Annals of Statistics*, (22):1701–1762, 1994.

[13] E.C. Kuruoglu, P.J.W. Rayner, and W.J. Fitzgerald. Least l_p-norm estimation of autoregressive model coefficients of symmetric α-stable processes. *IEEE Signal Processing Letters*, 4(7):201–203, July 1997.