1 Introduction

This document is compilation of notes on various papers and books from the relevant literature.

2 Efficient Bayesian Parameter Estimation Frühwirth-Schnatter [2004]

Consider a common statespace model:

$$\begin{aligned} \beta_t &= \phi\beta_{t-1} + (1-\phi)\mu + w_t & w_t &\sim N(0, \sigma_w^2) \\ y_t &= Z_t\beta_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \end{aligned}$$

Usual MCMC algorithm: let $\theta = (\mu, \phi, \sigma_w^2, \sigma_\epsilon^2)$. Then DA algorithm with two steps: $p(\theta|\beta, y)$ and $p(\beta|\theta, y)$.

Suppose $\phi = 0$, i.e. a random effects model. Centered parameterization: $\tilde{\beta}_t = \beta_t - \mu$. Let $D = 1 - V(y_t|\beta_t)/V(y_t)$ (depends on STN ratio, obv). $D > 1/2$ means centered is better, $D < 1/2$ means noncentered is better. (D is roughly the STN ratio)

Now suppose $\phi \neq 0$. Same $\tilde{\beta}_t$. Implied model:

$$\begin{aligned} \tilde{\beta}_t &= \phi\tilde{\beta}_{t-1} + w_t & w_t &\sim N(0, \sigma_w^2) \\ y_t &= Z_t\mu + Z_t\tilde{\beta}_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \end{aligned}$$

Pitt and Shephard [1999] prove that for $Z_t = 1$ with known variances: $\phi \to 1 \implies$ the convergence rate of the centered parameterisation goes to 0, whereas the convergence rate of the noncentered paramerization goes to 1. So for the limiting random walk model, the noncentered parameterization does not converge geometrically regardless of the STN. But when $\phi < 1$ the variances matter, CP better than NCP when $\sigma_w^2/(1-\phi)^2 > \sigma_\epsilon^2$. (NOTE: only centered in location, NOT scale)

Also a section on partial noncentering (not as relevant):

$$\beta_t^w = W_t\tilde{\beta}_t + (1 - W_t)\beta_t.$$

With $W_t = 1 - D_t$, Bernardo et al. [2003] show that iid samples can be obtained. Unclear how to select W_t for a time series model.

When the variances are unknown, Meng and Van Dyk [1998] showed that for a random effects model NC in location, when D is small (i.e. low STN) we have a poor sampler. Solution: rescale the state vector (noncentered in scale):

$$\beta_t^* = \frac{\tilde{\beta}_t}{\sigma_w}$$

Can also do partial noncentering:

$$\beta_t^a = \frac{\tilde{\beta}_t}{\sigma_w^A}$$

For random effects model, Meng and Van Dyk [1998] suggest $A = 2(1-D)/(2-D)$.

No one knows what happens when you NC a *time series* in the scale parameter. Simulations: known $\phi = 0.1, 0.95$, unknown variances. Data has drawn from $\sigma_w^2 = 1, 0.05, 0.001$ and $\sigma_\epsilon^2 = .1$, also Z_t is randomly $-1, 0, 1$. For $\phi = 0.1$ NC in location and scale improves the "preferred" sampler (based on D e.g.) in all cases except for μ when $\sigma_w^2 = 1$. For $\phi = 0.95$, on the other hand, when σ_w^2 is smaller the NCP is worse for σ_w^2 and μ. However when σ_w^2 is larger the NCP is better for σ_w^2 and just as good for the other parameters.

What if ϕ is unknown... (> 0). Basically nothing changes if σ_w^2 is not too small, but whe it's close to 0, the model is "nearly oversized" - main problem is that ϕ is still in the system equation while everything else (in the NCP) is in the observation equation. So new parameterization:

$$\begin{aligned} w_t &\sim N(0,1) \\ y_t = Z_t\mu + Z_t\sigma_w\beta_t^* + \epsilon_t &\qquad \epsilon_t \sim N(0,\sigma_\epsilon^2) \end{aligned}$$

where $\beta_t^* = \phi\beta_{t-1}^* + w_t$. Missing data are defined as $\tilde{X} = (\beta_0^*, w_{1:T})$. Removes all model paramters from system equation. Full Gibbs no longer possible - use a random walk metropolis hastings algorithm. The result is that if σ_w^2 is very small, results improve (specifically for ϕ which typically has the worst problems), but mostly when ϕ is small. **They try some other parameterizations, but ultimately find that nothing seems to do better than one of 1) standard CP 2) NCP for disturbances (my scaled disturbances).**

3 Efficient Parameterisations for Normal Linear Mixed Models Gelfand et al. [1995]

Start with a basic model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \sim N(0,\sigma_e^2)$, $\beta_{ij} \sim N(0,\sigma_\beta^2)$, $\alpha_i \sim N(0,\sigma_\alpha^2)$ and $\mu \sim N(\mu_0,\sigma_\mu^2)$. Assume that all variance components are known for now. An alternative "centered parameterization" (CP) is $\eta_i = \mu+\alpha_i$ and $\rho_{ij} = \mu+\alpha_i+\beta_{ij}$ which gives $Y_{ijk} = \rho_{ij} + \epsilon_{ijk}$ where $\rho_{ij} \sim N(\eta_i,\sigma_\beta^2)$ and $\eta_i \sim N(\mu,\sigma_\alpha^2)$. Usually reparameterizations require the square root of an approximation to the joint covariance matrix, which is hard to compute in large models (requires a big martrix inverse).

Consider a different model: $Y_i : n_i \times 1$,

$$\begin{aligned} Y_i|\eta_i &\sim N(X_i\eta_i, \sigma_i^2 I_{n_i}) \\ \eta_i|\mu &\sim N(\mu, D) \end{aligned}$$

where σ_i^2 and D are known (for now). Take a flat prior on μ. (μ,η) is the CP while (μ,α) where $\alpha = \eta - \mu$ is the NCP. Posterior is multivariate normal in either case.

Conditional on μ, the Y_i are independent with $Y_i|\mu \sim N(X_i\mu, \Sigma_i)$ where $\Sigma_i = \sigma_i^2 I_{n_i} + X_i D X_i'$. Thus $\mu|Y \sim N[\hat{\mu}, (X'\Sigma^{-1}X)^{-1}]$ where

$$\begin{aligned} X' &= (X_1', ..., X_m') \\ \Sigma &= diag(\Sigma_1, ..., \Sigma_m) \\ Y' &= (Y_1', ..., Y_m') \\ \hat{\mu} &= (X'\Sigma^{-1}X^{-1}X'\Sigma^{-1}Y \end{aligned}$$

Let $A_i = X_i'\Sigma_i^{-1}$ and $A = \sum_i A_i = X'\Sigma^{-1}X$. Then we have $\eta_i|\mu, Y \sim N(B_i b_i, B_i)$ where

$$\begin{aligned} B_i &= (\sigma_i^{-2}X_i'X_i + D^{-1})^{-1} \\ b_i &= \sigma_i^{-2}X_i'Y_i + D^{-1}\mu \end{aligned}$$

which implies that $\eta|Y$ is normal with

$$\begin{aligned} E[\eta_i|Y] &= B_i\hat{b}_i \\ \hat{b}_i &= \sigma_i^{-2}X_i'Y_i + D^{-1}\hat{\mu} \\ V(\eta_i|Y) &= B_i + B_iD^{-1}A^{-1}D^{-1}B_i \\ cov(\eta_i,\mu|Y) &= B_iD^{-1}A^{-1} \\ cov(\eta_i,\eta_j|Y) &= B_iD^{-1}A^{-1}D^{-1}B_j \end{aligned}$$

whereas in $\alpha - \mu$ space we have $\alpha|Y$ normal with

$$\begin{aligned} E[\alpha_i|Y] &= B_i\hat{b}_i - \hat{\mu} \\ V(\alpha_i|Y) &= B_i + B_iD^{-1}A^{-1}D^{-1}B_i + A^{-1} - 2B_iD^{-1}A^{-1} \\ cov(\alpha_i, \mu|Y) &= B_iD^{-1}A^{-1} - A^{-1} \\ cov(\alpha_i, \alpha_j|Y) &= B_iD^{-1}A^{-1}D^{-1}B_j - (B_i + B_j)D^{-1}A^{-1} + A^{-1}. \end{aligned}$$

A matrix identity gives $B_iD^{-1} + DA_i = I_{n_i}$ Now B_iD^{-1} is PD and DA_i is PSD so B_iD^{-1} measures the relative contribution of the error variance and DA_i measures the relative contribution of the random effect variance.

When $|B_iD^{-1}|$ is near zero the CP is efficient while when it's near one the NCP is efficient. (Pf shows correlations between η's and μ's go to zero in one case, and α's and μ's in the other.

They do something similar for a more complicated model and run some simulations in order to confirm their findings.

Note: this doesn't help us at all - we're drawing the theta's jointly conditional on the other stuff. the problem is the variances!!

4 Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler, Roberts and Sahu [1997]

Let θ^t be a markov chain with stationary density $h(\theta)$. Let f be a square h-integrable function of θ and $h(f)$ denote the expectation of f under density h. Then we look at the rate at which $P^t f(\theta^0) \equiv E_h[f(\theta^t)|\theta^0]$ approaches $h(f)$ in L^2. Define ρ to be the minimum number such that for all square h-integrable function f and for all $r > \rho$

$$\lim_{t\to\infty} \left(E_h[(P^t f(\theta^0) - h(f))^2]r^{-t}\right) = 0$$

Sometimes it's impossible to compute ρ, but we can compute ρ_L which restricts the functions f to be linear. Often $\rho_L = \rho$ but generally $\rho_L \leq \rho$.

The survey the literature which says that usually random updating schemes are better, but they'll show that in two cases a deterministic scheme is better: hierarchical models in a certain class and density with non-negative partial correlations. It's also know that blocking often improves convergence, but they emphasize that it can make an algorithm converge more slowly. They also mention that "It is well known that high correlations between the coordiantes dimish the speed of convergence of the Gibbs sampler; see, for example, Hills and Smith 1992." They ultimately compare the CP to alternative parameterizations (note, only centered in the mean, not variance).

Didn't read sections 2 and 3 closely - only seems to talk about blocking and such.

4.1 Optimal parameterizations for Gaussian linear models

Theoretical result on when the CP and NCP are better for basic model: $y_i = \mu + \alpha_i + \epsilon_i$ where y_i and ϵ_i ahve been reduced by sufficiency - basically CP is better when variance of ϵ is lower than that of α and otherwise NCP better (CP better when STN ratio high). If we add another level to the model, β_{ij}, it's more complicated and it's no longer obvious that the deterministic updating scheme is better either (depends on the parameterization!).

Lots of proofs in appendicies.

5 The EM Algorithm – An Old Fok-Song Sung to a Fast New Tune, Meng and Van Dyk [1997]

starts with a history less on the EM algorithm

5.1 Augmentating data efficiently to speed up EM algorithm

It's known that the rate of convergence is determined by the fraction of missing information.

Some details on working with a multivariate t model - treat it as a chi-square mixture of normals, and treat the chi-square rv is the missing data. (One chi-sq for each data point)

From Dempster et al 1977 we know that the matrix rate of the EM algorithm is (assuming limit is an interior point)

$$DM = I - I_{obs} I_{aug}^{-1}$$

where I is the identity matrix,

$$I_{aug} = E\left[-\frac{\partial^2 \log f(Y_{aug}|\theta)}{\partial\theta\partial\theta'} \Big| Y_{obs}, \theta \right] \Big|_{\theta=\theta^*}$$
$$I_{obs} = -\left. \frac{\partial^2 L(\theta|Y_{obs})}{\partial\theta\partial\theta'} \right|_{\theta=\theta^*}$$

i.e. the expected and observed information matrices, where θ^* is a (local) MLE. The largest eigenvalue of DM, denoted r, is known as the (global) rate of convergence of the EM algorithm. $s = 1 - r$ is known as the global speed of the algorithm. s is the smallest eigenvalue of the speed matrix $S = I_{obs} I_{aug}^{-1}$.

They allow the *aug* quantities to depend on some parameter a and look for the a which maximizes s. This looks like the genesis of "parameter expanded data augmention" and they talk about the similarities to stochastic algorithms and other issues. Didn't read the rest of the details too closely, but they mention a paper by Orchard which talks about the "missing information principle."

6 Fast EM-type implementations for mixed effects models, Meng and Van Dyk [1998]

Consider a mixed effects model. The EM algorithm, treating random effects as missing data, is a popular method to fit these models (to obtain MLEs). However, it has slow convergence especially when the variances of the random effects are relatively small. There are lots of alternatives to the EM algorithm (e.g. Newton-Raphson) but they require lots of human effort.

To set up, the EM algorithm requires defining a data augmentation Y_{aug} such that $Y_{obs} = M(Y_{aug})$. for M some many-to-one mapping. The theoretical speed of convergence of the algorithm is then determined by the smallest eigenvalue of the "fraction of observed information" (Dempster et al. [1977]) $I_{obs} I_{aug}^{-1}$ where I_{aug} is the expected Fisher information and I_{obs} is the observed Fisher information matrix (see Meng and Van Dyk [1997]).

This paper's schtick is to consider parameter expanded data augmentation, again. Define $Y_{aug}(a)$ and minimize $I_{aug}(a)$ in a.

6.1 Standard and alternative implementations

Consider the mixed effects model

$$y_i = X_i'\beta + Z_i' b_i + e_i$$

with $b_i \overset{iid}{\sim} N_q(0, T)$ independent of $e_i \overset{iid}{\sim} N(0, \sigma^2 R_i)$ where R_i's are known, the Z_i's are known and are such that T is identifiable, and the X_i's are known. The standard EM implementation is to treat the b_i as missing data.

An alternative implementation scales the b_i's by $T^{-a/2}$ where a is the working parameter. $a = 1$ is natural for some versions of this model, but for others the form of T may make it too complicated to be easy and efficient. Of course the solution is a cholesky decomposition - helps make it numerically stable. Specifically, let $T = \Delta U \Delta'$, then $c_i = \Delta^{-1} b_i$ so that $c_i \sim N(0, U)$. Now we can rescale each c_i by a power of it's own

standard deviation, $u_i^{a_i}$. **Note that the ordering of the random effects changes the definition of c_i, so there are $q!$ possible data augmentations**. (This applies for us as well!!! - at least when F or G are matrices). There are plenty of variations of this that depend on whatever structure is on T.

6.2 Simulation Studies

Define

$$D^* = \frac{\sum_{i=1}^m \text{tr}(Z_i' T^* Z_i)/m}{\sigma^{2*} + \sum_{i=1}^m \text{tr}(Z_i' T^* Z_i)/m}$$

When D^* is close to 0, the standard algorithm is very slow. When D^* isn't very small or very large, the alternative algorithm does well. When D^* is very large, the standard algorithm is way better than the alternative. The cutoff is $D^* = 2/3$ - smaller than that, alternative is better. Larger than that, the standard is better. Very close to 0, the alternative is way better. This is true over a couple of different models. For some models value of D^* cutoff changes, and the difference may be small. They also set up an adaptive algorithm that picks what seems to be the best.

6.3 Theory

The best a is

$$a_0 = \frac{2(1 - \tilde{D}^*)}{2 - \tilde{D}^*}$$

where

$$\tilde{D}^* = \frac{\sum_{i=1}^m \text{tr}(Z_i' T_i^* Z_i)/m}{\sigma^{2*} + \sum_{i=1}^m \text{tr}(Z_i' T_i^* Z_i)/m}$$

Note: the tilde D^* depends on T_i and not T. When Z_i doesn't depend on i, then $\tilde{D}^* = D^*$. (Note $T_i^* = E[b_i^2 | Y_{obs}, \theta^*]$, $\theta = (\beta, \sigma^2, T)$, i.e. it depends on what we observed: Z_i, it's the posterior squared expectation). THis makes the M step really complicated though, s limiting the a's to be either 0 or 1. In that case, $a = 0$ is the minimizer if $\tilde{D}^* \geq 2/3$, while otherwise $a = 1$ is. Ultimately they suggest using $a = (1, 1, ..., 1)$ in at least a burn in period. They also note that ist suggests Gibbs methods.

7 Analytic Convergence Rates and Parameterization Issues for the Gibbs Sampler Applied to State Space Models, Pitt and Shephard [1999]

Main purpose of the paper: obtain the analytical convergence rate for the single-move Gibbs sampler applied to the states of the AR(1) states plus gaussian noise model. Then consider alternative parameterizations in the same model for a gibbs sampler with two blocks - parameters θ and states α. Similar issues in the stochastic volatility model are considered.

7.1 AR(1)plus noise model

The model is

$$\begin{aligned} y_t &= \mu + \alpha_t + \epsilon_t & \epsilon_t &\sim NID(0, \sigma_\epsilon^2) \\ \alpha_t &= \phi \alpha_{t-1} + \eta_t & \eta_t &\sim NID(0, \sigma_\eta^2) \\ & & \alpha_1 &\sim NID(0, \sigma_\eta^2/(1 - \phi^2)) \end{aligned}$$

with $|\phi| < 1$, though $|\phi| = 1$ doesn't change their results (they claim). Start with $\theta = (\sigma_\epsilon^2, \sigma_\eta^2, \phi, \mu)'$ known. We focus on drawing each of the α's from their full conditional distribution. Attempt to find the geometric

rate of convergence, i.e. the ρ such that for all $r < \rho$ and all square integrable functions f and some function $V(.) \geq 1$ such that the expectation under the target density $\pi(V) < \infty$ and

$$|E[f(\alpha^{(i)}) - \pi(f)| \leq V(\alpha^{(0)} r^i.$$

If such a $\rho < 1$ exists, that is its geometric rate of convergence. They show that in the limit of an infinite time series.

$$\rho = 4\frac{\phi^2}{(1 + \phi^2 + \sigma_\eta^2 \sigma_\epsilon^{-2})^2}$$

So as $n \to infty$, $|\phi| \to 1$,, $\sigma_\eta^2 \sigma_\epsilon^{-2} \to 0$ then $\rho \to 1$, so for persistent parameterizations the convergence rate will be close to 1, i.e. very slow. (This is duh - the more correlated the α's are, worse off you are from not blocking the α's together.)

7.2 Reparameterizations

Now consider the two models

$$y|\alpha \sim N(\mu\mathbf{1} + \alpha, \sigma_\epsilon^2 I_n) \qquad \alpha \sim N(0, D)$$

and

$$y|\omega \sim N(\omega, \sigma_\epsilon^2 I_n) \qquad \omega \sim N(\mu\mathbf{1}, D)$$

where we treat D as known and μ as unknown. The first case is the usual parameterization of the AR(1) plus noise model, the second has $\omega_t = \mu + \phi(\omega_{t-1} - \mu) + \eta_t$. A flat prior is assumed on μ and $|\phi| < 1$ to ensure that μ is identifiable. Assume that σ_x^2 and ϕ are known ($x = \eta, \epsilon$). Two block sampler - μ then α or ω. Let $\rho_\mu(1, \alpha)$ denote the lag 1 autocorrelation for α (and ω of course). Then

$$V = (\sigma_\epsilon^2 I + D^{-1})^{-1}$$
$$\rho_\mu(1, \alpha) = 1 - \frac{1}{n}\mathbf{1}'VD^{-1}\mathbf{1}$$
$$\rho_\mu(1, \omega) = 1 - \sigma_\epsilon^{-2}\frac{\mathbf{1}'VD^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}}$$

For persistent models $\rho_\mu(1, \alpha)$ is close to 1. $\rho_\mu(1, \omega) < \rho_\mu(1, \alpha)$ when $\mathbf{1}'D^{-1}\mathbf{1}\sigma_\epsilon^2 < n$. For the AR(1) plus noise model (structuring D) then becomes

$$\frac{\sigma_\epsilon^2 \sigma_\eta^{-2}}{n}\left((n-2)(1+\phi^2) + 2 - 2(n-1)\phi\right) < 1$$

in the limit

$$\sigma_\epsilon^2 \sigma_\eta^{-2}(1-\phi)^2 < 1.$$

In simulations, when atocorrelation is high and signal to noise ratio is low, awful mixing using the α's and good using the ω's. Similar results for the autocorrelation of the middle state (α or ω).

8 The Art of Data Augmentation, Van Dyk and Meng [2001]

Data is Y_{obs} and we want to sample from $p(\theta|Y_{obs}) \propto p(Y_{obs}|\theta)p(\theta)$. DA algorithm starts with $M : Y_{aug} \to Y_{obs}$, $Y_{obs} = M(Y_{aug})$ such that

$$\int_{M(Y_{aug})=Y_{obs}} p(Y_{aug}|\theta)\mu(dY_{aug}) = p(Y_{obs}|\theta)$$

(a.s. wrt some appropriate dominating measure). Then the standard DA algorithm based on this samples from $p(\theta|Y_{aug}, Y_{obs})$ and $p(Y_{aug}|\theta, Y_{obs})$. (Slice sampling can be seen as a version of this)

Based on the EM literature, there's a tradeoff between simplicity and speed - speed depends on the "fraction of missing information" so that the "larger" (in fisher info terms) the DA, the faster the convergence, but larger DAs are harder to construct.

The geometric rate of convergence of the DA algorithm is

$$\lambda^{DA}(\alpha) = 1 - \inf_{h:var[h(\theta)|Y_{obs}]=1} E[var(h(\theta)|Y_{aug}, \alpha)|Y_{obs}, \alpha]$$

where the expection is wrt the stationary density $p(\theta, Y_{aug}|Y_{obs}, \alpha)$. The maximum autocorrelation over linear compinations is

$$\sup_{x \neq 0} \text{corr}(x'\theta^{(t)}, x'\theta^{(t+1)}) = \sup_{x \neq 0} \frac{x' var[E(\theta|Y_{aug}, \alpha)|Y_{obs}, \alpha]x}{x' var(\theta|Y_{obs})x} = \rho(\mathcal{F}_B(\alpha))$$

Minimize one of these two in α is the basic strategy here, where $\mathcal{F}_B(\alpha)$ is the Bayesian fraction of missing information:

$$\mathcal{F}_B(\alpha) = I - [var(\theta|Y_{obs})]^{-1} E[var(\theta|Y_{aug}, , \alpha)|Y_{obs}, \alpha]$$

and $\rho(A)$ is the spectral radius of A. (i.e. maximum absolute value of eigenvalues of A). Note that this is often hard, but the EM criterion is easier to minimize in α:

$$\mathcal{F}_{EM}(\alpha) = I - I_{obs} I_{aug}^{-1}(\alpha)$$

where

$$I_{aug}(\alpha) = \mathrm{E}\left[- \left. \frac{\partial^2 \log p(\theta|Y_{aug}, \alpha)}{\partial\theta\dot{\partial}\theta} \right| Y_{obs}, \theta, \alpha \right] \Bigg|_{\theta=\theta^*}$$

$$I_{obs} = - \left. \frac{\partial^2 \log p(\theta|Y_{obs})}{\partial\theta\dot{\partial}\theta} \right|_{\theta=\theta^*}$$

where θ^* is the posterior mode. Authors suggest using the EM criterion because it's easier (it's the same if the augmented data posterior is normal), and thus only worrying about minimizing $I_{aug}^{-1}(\alpha)$ in alpha. The essence of the idea: when it's too difficult to compare two stochastic algorithms (DA), instead compare their deterministic counterparts (EM) to decide which stochastic algorithm to use - often leads to good-if-not-best stochastic algorithms. Often we don't even need to know θ^*.

Basic idea of the paper:

1. Minimize I_{aug} in α.
2. Put a prior on α and minimize I_{aug} in the prior on α. Sample (α, θ) jointly or marginalize out α.

They have a bunch of examples and theoretical results related to these ideas, including difficulties with an improper prior on α.

They have a fantastic note on page 40 near the top about the "art" of this process because yes, we can find the "optimal" sort of DA algorithm, but it might require us to sample from the posterior with iid draws, which is what we're trying to do in the first place!

9 Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes, Roberts et al. [2004]

consider the follow model for log stock prices

$$dx(t) = v(t)^{1/2} dB(t) \qquad\qquad t \in [0, T]$$

with $x(0) = 0$. The the variance $v(t)$ is modeled as a stationary non-Gaussian OU process with decay rate $\mu > 0$ which is driven by a homogeneous Levy process $z(.)$ with positive increments and $z(0) = 0$, e.g.:

$$dv(t) = -\mu v(t)dt + dz(t)$$

(Levy processes are precesses with independent and stationary increments). This model captures many stylized facts about asset prices, including serial dependence but not autocorrelation, volatility clustering, skewness, fat tails, etc. They assume that $v(t)$ is a sort of poisson process that jumps up suddenly then trails off exponentially (e.g. new information arriving in packets). In particular

$$v(t) = \exp(-\mu t)v(0) + \sum_{j=1}^{\infty} \exp(-\mu(t - c_j))\epsilon_j \mathbf{1}(c_j < t)$$

where $0 < c_1 < c_2 < ...$ are the arrivals of a poisson process with finite rate λ and the ϵ_j's are iid $exp(\theta)$. Here $z(t) = \sum_{j=1}^{\infty} \epsilon_j \mathbf{1}(c_j < t)$.

This model fails to capture some stylized facts about the dependence of stock returns, but if instead we let

$$v(t) = \sum_{i=1}^{m} v_i(t)$$

where each v_i has an independent levy process. Other work shows that $m = 2$ is adequate for daily financial data. Some more details here that I'm skipping.

9.1 Data augmentation

Which parameterization to use? CP:

$$\Psi = \{(c_j, \epsilon_j)\}$$

Note: $\Psi : [0, T] \times (0, \infty) \to \Re$. The NCP: $\tilde{\Psi}$ so that a prior, $\tilde{\Psi}$ is independent of the parameters. There are many ways to do this, but they all correspond to the same graphical model just with different priors on $\tilde{\Psi}$, so a different transfornation to Ψ. Over a wide range of the parameterspace in a simulation study, the NCP is just as good and sometimes significantly better than the CP. They also do a prior sensitivity analysis and a real data example.

10 A General Framework for the Parameterization of Hierarchical Models, Papaspiliopoulos et al. [2007]

Suppose we have observed data Y unknown parameters Θ with prior dentiy $p(\Theta)$ and data model $p(Y\Theta)$ which cn be conveniently expressed using a hidden *layer* X as

$$p(Y|\Theta) = \int p(Y|X, \Theta)p(X|\Theta)d\mu(X)$$

where μ is the measure wrt $p(X|\Theta)$ is defined. A reparameterization is (X^*, Θ) such that

$$X = h(X^*, \Theta, Y)$$

It's practical if $p(\Theta|X^*)$ and hence $p(\Theta|X^*, Y)$ is known up to a normalizing constant. The centered parameterization is (in a graphical model):

$$\Theta \to X \to Y.$$

The Gibbs sampler considered alternations between X and Θ, of course. The NCP is

$$\Theta, \tilde{X} \to X \to Y.$$

where $X = h(\tilde{X}, \Theta)$.

They go through a wide clases of models and show that what the order (in n) of τ_c and τ_{nc} (CP and NCP respectively) where $\tau = 1/log\gamma \approx 1/(1-\gamma)$ for $\gamma \approx 1$, where $\gamma = \sup_f \gamma_f$ and $\gamma_f = 1 - \frac{E[var(f(\Theta)|X^*,Y)|Y]}{var[f(\Theta)|Y]}$, the bayesian fraction of missing information for f maximized over all f. Sometimes the CP improves while the NCP worsens as sample size increases, e.g. in a simple hierarcichal model, and sometimes vice versa. Sometimes γ is not very useful because it measures global dependence, and depending on your sample, dependence for your sampler might be drastically different.

(They have a couple stochastic volatility examples)

In order to construct an NCP they have a couple ideas that can be combined:

1. Location: $h(\tilde{X}, \Theta) = \tilde{X} + \Theta$. $X \sim F(\dot{-}\Theta) \implies \tilde{X} \sim F(.)$.
2. Scale: $h(\tilde{X}, \Theta) = \Theta\tilde{X}$. $X \sim F(\dot{/}\Theta) \implies \tilde{X} \sim F(.)$.
3. Inverse CDF: $h(\tilde{X}, \Theta) = F_\Theta^{-1}(X)$ where $X \sim F_\Theta$ implies $\tilde{X} \sim U(0,1)$.

They acknowledge that alternating between a CP and an NCP works very well (page 12).

Correcting a NCP - partially noncentered NCP - basically making an NCP take into account the data by using the posterior mean and SD instead of the prior mean and SD to construct the parameterization. (Maybe we should try this!!!)

11 Stability of the Gibbs Sampler for Bayesian Hierarchical Models, Papaspiliopoulos and Roberts [2008]

Consider

$$Y \sim L(Y|X)$$
$$X \sim L(X|\Theta)$$

where Y is the data, X is the missing data, and Θ is the parameter. They explicitly how the relative tail behavior of $L(Y|X)$ and $L(X|\Theta)$ determines the stability of the Gibbs sampler (uniform, geometric or sub geometric convergence). And tail behavior thus determines the type of parameterization that should be adopted. Restricted to linear hierarchical models with general error distributions. So the model is

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\boldsymbol{X}_i + \boldsymbol{Z}_{1i}$$
$$\boldsymbol{X}_i = \boldsymbol{D}\boldsymbol{\Theta} + \boldsymbol{Z}_{2i}$$

where $\boldsymbol{Z}_{1i} \overset{iid}{\sim} L(\boldsymbol{Z}_1)$, $\boldsymbol{Z}_{2i} \overset{iid}{\sim} L(\boldsymbol{Z}_2)$, and both are symmetric around $\mathbf{0}$, independent of each other.

It's known that if the tails of $L(\boldsymbol{Z}_1)$ are heavier than the tails of $L(\boldsymbol{Z}_2)$ then inference for $\boldsymbol{X}$ is robust to outlying observations, whereas in the opposite case inference for $\boldsymbol{X}$ is less influenced by the prior in case of data-prior conflict.

Place an improper flat prior on Θ. The parameterization (X, Θ) is the CP, used to refer to a parameterization where the parameters and the data are conditionally independent given the missing data. The NCP is $(\tilde{X}, \Theta)$ where $\tilde{X}_i = h(X_i, \Theta)$, $h(x, \theta) = x - D\theta$.

Their basic strategy is to consider 4 different types of distribution for $L(\boldsymbol{Z}_1)$ and $L(\boldsymbol{Z}_2)$. In order of fattest tails: Cauchy (C), Double Exponential (D), Gaussian (G), and Exponential power with $\beta > 2$ (lighter than gaussian tails) (L). (C,E) denotes cauchy for Z_1 and double exponential for Z_2. We also have 3 types of stability results in order of most stable: Uniform ergodic (U), geometric ergodic (G) and non/sub-geometric (N). Basically, they find that Fatter tails for Z_1 than Z_2 leads to worse stability of the CP and better stability of the NCP. Thinner tails for Z_1 than Z_2 leads to better stability of the CP and worse of the NCP. A proper prior on Θ improves convergence properties all around. In the (E,E) model, the ratio of scale parameters matters. Heuristically though: the CP is better when Z_1 has lighter tails than Z_2 and worse when Z_1 has

fatter tails. The NCP is the reverse, and both algorithms become more stable as the tails of both Z_1 and Z_2 become lighter.

They also note that linear reparameterizations may only work when the tails of Z_1 and Z_2 are the same!

12 Cross-Fertilizing strategies for better EM mountain climbing and DA field exploration, Van Dyk et al. [2010]

Exploration of lotss of different Gibbs samplers and EM algorithms and how they inform each other. Have not read.

13 Other Papers can't find a copy

These are in the back of the state space book:

1. Papaspilipoulos, Robers and Skold 2003: whole continuum of location parameterizations
2. Shepard 1996: reparameterization in stochastic volatility models
3. Fruhwirth-Schnatter and Sogner: reparameterization in stochastic volatility models

Papers to get:

1. Orchard and Woodbury 1972: A missing information principle

References

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, page 307. Oxford University Press, USA, 2003.

Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

Sylvia Frühwirth-Schnatter. Efficient Bayesian parameter estimation for state space models based on reparameterizations. *State Space and Unobserved Component Models: Theory and Applications*, pages 123–151, 2004.

Alan E Gelfand, Sujit K Sahu, and Bradley P Carlin. Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488, 1995.

X-L Meng and David Van Dyk. Fast em-type implementations for mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):559–578, 1998.

Xiao-Li Meng and David Van Dyk. The em algorithman old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.

Omiros Papaspiliopoulos and Gareth Roberts. Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics*, pages 95–117, 2008.

Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

Michael K Pitt and Neil Shephard. Analytic convergence rates and parameterization issues for the gibbs sampler applied to state space models. *Journal of Time Series Analysis*, 20(1):63–85, 1999.

Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2): 291–317, 1997.

Gareth O Roberts, Omiros Papaspiliopoulos, and Petros Dellaportas. Bayesian inference for non-gaussian ornstein–uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):369–393, 2004.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001.

David A Van Dyk, Xiao-Li Meng, et al. Cross-fertilizing strategies for better em mountain climbing and da field exploration: A graphical guide book. *Statistical Science*, 25(4):429–449, 2010.