1 Introduction

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

2 Efficient Bayesian Parameter Estimation ?

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:

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

? 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 $W_t = 1 - D_t$, **?** show that iid samples can be obtained. Unclear how to select W_t for a time series model.

When the variances are unknown, **?** 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, **?** 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:

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

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 ?

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, ?

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, ?

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 Demptster 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,

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

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, ?

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" (**?**) $I_{obs} I_{aug}^{-1}$ where I_{aug} is the expected Fisher information and I_{obs} is the observed Fisher information matrix (see **?**).

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 $T_i^* = E[b_i^2 | Y_{obs}, \theta^*]$, $\theta = (\beta, \sigma^2, T)$

7 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
2. Dempster, Laird and Rubin 1977:Maximum Likelihood from incomplete data via the EM algorithm