Stochastic Volatility Models

May 2, 2019

The material in these notes draws from several sources, including Section 10 of Chapter 6 of S&S, Section 4 of Chapter 11 of Chan (2010), and a very nice paper documenting the `stochvol` package for `R` by the software's author, Gregor Kastner.

Introduction to the Stochastic Volatility Model

The stochastic volatility model is a **nonlinear state-space model**, which provides an alternative to the **ARCH** and **GARCH** models we discussed previously.

For a univariate time series of length n we assume

$$\begin{aligned} y_t &= \exp\left\{h_t/2\right\} v_t && \text{Observation Equation} \\ \left(h_t - \mu_h\right) &= \phi\left(h_{t-1} - \mu_h\right) + w_t && \text{State Equation,} \end{aligned}$$

where $|\phi| < 1$, $v_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 1\right)$, $w_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2_w\right)$ and $h_1 \sim \mathcal{N}\left(\mu_h, \sigma^2_w / \left(1 - \phi^2\right)\right)$. The state equation is still a simple **AR**(1) model, but the observations y_t are a **nonlinear** function of the latent states.

Thus far throughout the course, our process has proceeded as follows:

1. Specify a model for a time series $\boldsymbol{y}$ that involves some latent variables and a handful of parameters;

2. Find the marginal likelihood maximizing values of the parameters;

3. Compute conditional means and variances of the states and forecasts by plugging in the marginal likelihood maximizing values of the parameters.

We run into two serious computational problems when we work with this model.

($\times$) The full conditional distributions of the states given the observed data are not available in closed form, nor are the corresponding conditional means and variances

- Because this we don't have computationally efficient or speedy ways of computing the predictions $\mathbb{E}\left[h_t|y_1,\dots,y_{t-1}\right]$ and their variances $\mathbb{V}\left[h_t|y_1,\dots,y_{t-1}\right]$, filtered values $\mathbb{E}\left[h_t|y_1,\dots,y_t\right]$ and their variances $\mathbb{V}\left[h_t|y_1,\dots,y_t\right]$, smoothed values $\mathbb{E}\left[h_t|\boldsymbol{y}\right]$ and their variances $\mathbb{V}\left[h_t|\boldsymbol{y}\right]$, or forecasts $\mathbb{E}\left[y_{n+k}|\boldsymbol{y}\right]$ and their variances $\mathbb{V}\left[y_{n+k}|\boldsymbol{y}\right]$ even if we treat the parameters μ_h, σ^2_w, and ϕ as known.

($\star$) It's impractically difficult to maximize the marginal likelihood over the parameters μ_h, σ^2_w, and ϕ, which we would usually use to construct our predictions, filtered values, smoothed values, and forecasts.

Approximating the Smoothed State and Forecast Values and Their Variances

Let's consider the the first problem ($\times$). It turns out that when the full conditional distributions are not available in closed form, we can approximate the conditional means and variances via simulation. Consider smoothed values $\mathbb{E}\left[h_t|\boldsymbol{y}\right]$ and their variances $\mathbb{V}\left[h_t|\boldsymbol{y}\right]$ - they correspond to the means and variances of the conditional distributions of the states given the data, e.g. $p\left(h_1|\boldsymbol{y}\right) = \int p\left(\boldsymbol{h}|\boldsymbol{y}\right) dh_2 \dots dh_n$. If we could simulate m values of the states $\boldsymbol{h}^{(1)},\dots,\boldsymbol{h}^{(m)}$ according to the $p\left(\boldsymbol{h}|\boldsymbol{y}\right)$, then we could approximate the smoothed values and their variances with the sample mean and variance of the simulated values:

$$\begin{aligned}\mathbb{E}\left[h_t|\boldsymbol{y}\right] &\approx \frac{1}{m}\sum_{i=1}^{m} h_t^{(i)} \\ \mathbb{V}\left[h_t|\boldsymbol{y}\right] &\approx \frac{1}{m-1}\sum_{i=1}^{m}\left(h_t^{(i)} - \frac{1}{m}\sum_{i=1}^{m} h_t^{(i)}\right)^2.\end{aligned} \tag{1}$$

Fortunately, there are very powerful **Markov Chain Monte Carlo (MCMC)** methods for simulating from any distribution that we can write out if we know how the density depends on the random variable(s) we want to simulate, even if we don't know the normalizing constant. For this problem, we want to simulate values of the random variable $\boldsymbol{h}$ according to

$$\begin{aligned}p\left(\boldsymbol{h}|\boldsymbol{y}, \phi, \mu_h, \sigma_w^2\right) &= \frac{p\left(\boldsymbol{h}, \boldsymbol{y}|\phi, \mu_h, \sigma_w^2\right)}{p\left(\boldsymbol{y}|\phi, \mu_h, \sigma_w^2\right)} \\ &= \frac{p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\phi, \mu_h, \sigma_w^2\right)}{p\left(\boldsymbol{y}|\phi, \mu_h, \sigma_w^2\right)} \\ &= \frac{\text{something we can write out as a function of } \boldsymbol{h}}{\text{normalizing constant that we don't know, but which doesn't depend on } \boldsymbol{h}}.\end{aligned}$$

This means that those powerful methods can be applied to get simulated values $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(m)}$! We won't get into the details in this class, we'll just use software that does this behind the scenes. Given a specified number of simulated values the user desires, m, the software will return m simulated values $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(m)}$ to us. Then we can approximate the smoothed values and their variances using (1). There are two important things to keep in mind:

- Successive simulated values $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(m)}$ may not be independent. This means that the m simulated values $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(m)}$ may contain less information than m **independent** draws from $p\left(\boldsymbol{h}|\boldsymbol{y}, \phi, \mu_h, \sigma_w^2\right)$. One way of quantifying the amount of information in each set of simulated values $h_t^{(1)}, \dots, h_t^{(m)}$ is to compute the **effective sample size**.
  - Without getting into the details, effective sample size takes a possibly correlated set of sampled values from a distribution, e.g. $h_t^{(1)}, \dots, h_t^{(m)}$, and returns an estimate of the number m', the number of independent draws $\tilde{h}_t^{(1)}, \dots, \tilde{h}_t^{(m')}$ that

would be needed to get the same variance. This is a tricky concept, but the key idea is that the **effective sample size** is a better way of quantifying how much information is in each set of simulated values $h_t^{(1)}, \ldots, h_t^{(m)}$ than m, the actual number of simulated values.

- We need to choose m to be large enough to get a good approximation of the smoothed values and their variances. There are lots of different ways of assessing whether or not m is large enough, but no hard-and-fast rule to go by. Some ad-hoc checks include:
  - Examining the effective sample size m' for each set of simulated values $h_t^{(1)}, \ldots, h_t^{(m)}$, and asking if that's a number we'd feel comfortable with if we were doing a simple random sample to estimate a population mean.
  - Visually examining the **trace plots** for each set of simulated values $h_t^{(1)}, \ldots, h_t^{(m)}$. A **trace plot** just plots the simulated values of $h_t^{(1)}, \ldots, h_t^{(m)}$ on the y-axis in the order they were drawn. The approximate smoothed values and variances may be bad approximations if the trace plots show any slowly varying or systematic trends. If there are slowly varying or systematic trends, increasing m and/or **thinning**, i.e. not saving every simulated value, can help.
  - Repeating the procedure of drawing m simulated values $\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(m)}$ several times, and comparing the approximate smoothed values and their variances each time. If m is large enough, each repeated procedure should produce nearly identical approximate smoothed values and their variances.
  - There are also various diagnostic tests that you could implement, but we won't discuss them here.

This tells us how to approximate the smoothed values and their variances, but what about the forecasts and their variances? Given m simulated values $\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(m)}$, we can simulate a future state value h_{n+k} and a forecast y_{n+k} according to the stochastic volatility model. More specifically, we would do the following for each value of i for $l = 1, \dots, k$:

$$h_{n+l}^{(i)} \sim \mathcal{N}\left(\mu_h^{(i)} + \phi^{(i)}\left(h_{n+l-1}^{(i)} - \mu_h^{(i)}\right), \sigma_w^{2(i)}\right) \text{ and } y_{n+l}^{(i)} \sim \mathcal{N}\left(0, \exp\left\{h_{n+l}^{(i)}\right\}\right).$$

Then, we can approximate the k-step ahead forecast and its variances:

$$\mathbb{E}\left[y_{n+k}|y_1, \dots, y_n\right] \approx \frac{1}{m}\sum_{i=1}^m y_{n+k}^{(i)} \text{ and } \mathbb{V}\left[y_{n+k}|y_1, \dots, y_n\right] \approx \frac{1}{m-1}\sum_{i=1}^m \left(y_{n+k}^{(i)} - \frac{1}{m}\sum_{i=1}^m y_{n+k}^{(i)}\right)^2. \quad (2)$$

Analogously, the k-step ahead prediction of the state and its variance can be approximated:

$$\mathbb{E}\left[h_{n+k}|y_1, \dots, y_n\right] \approx \frac{1}{m}\sum_{i=1}^m h_{n+k}^{(i)} \text{ and } \mathbb{V}\left[h_{n+k}|y_1, \dots, y_n\right] \approx \frac{1}{m}\sum_{i=1}^m \left(h_{n+k}^{(i)} - \frac{1}{m-1}\sum_{i=1}^m h_{n+k}^{(i)}\right)^2. \quad (3)$$

In general, most statistics software that uses MCMC methods to simulate latent states under time series models like the stochastic volatility model will just return draws from the conditional distribution of the states $\boldsymbol{h}$ given **all** of the observed data $\boldsymbol{y}$, which means that the output is not useful for computing the within-sample predictions or filtered values of the states and their variances because they correspond to means and variances of conditional distributions of states $\boldsymbol{h}$ given **some** of the observed data.

Dealing with the Stochastic Volatility Parameters

Now let's consider the the first problem ($\star$). So far, we've assumed that the parameters σ_w^2, ϕ, and μ_h are known. When we worked with the linear state-space model, we estimated the parameters by maximizing the marginal likelihood of the data $\boldsymbol{y}$. Unfortunately, that just isn't feasible when we are working with the stochastic volatility model. Under the stochastic volatility model, the marginal likelihood of the data $\boldsymbol{y}$ is given by:

$$p\left(\boldsymbol{y}|\sigma_w^2, \phi, \mu_h\right) = \int p\left(\boldsymbol{y}|\boldsymbol{h}\right) p\left(\boldsymbol{h}|\mu_h, \phi, \sigma_w^2\right) d\boldsymbol{h}. \quad (4)$$

Neither of the two methods we discussed in the linear-state-space setting is feasible here.

- **Direct maximum marginal likelihood** is not feasible because we don't have convenient distributional facts that tell us what the marginal distribution $\boldsymbol{y}$ is.

- **Expectation-maximization (EM) maximum marginal likelihood**, which maximizes $\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\mu_h,\phi,\sigma^2_w\right)\right)|\boldsymbol{y}\right]$ is not feasible because $\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\mu_h,\phi,\sigma^2_w\right)\right)|\boldsymbol{y}\right]$ cannot be simplified to depend on a handful of conditional expectations that we can compute easily and quickly. Some more details explaining this are provided for any especially curious readers at the end of these notes.

Saved by the Bayes!

It turns out that assuming prior distributions for the parameters σ^2_w, ϕ, and μ_h can help us out, even though doing so appears to make our model more complex. We often call models that include prior distributions for the parameters **Bayesian** models, although there are other ways of thinking about a model that includes parameters for the parameters.

If we assume prior distributions for the parameters, then instead of computing the conditional expectations and variances we need for prediction, filtering, smoothing, and forecasting for fixed, marginal likelihood maximizing values of the parameters σ^2_w, ϕ, and μ_h, we will compute the conditional expectations and variances we need for prediction, filtering, smoothing, and forecasting averaging over possible values of the parameters σ^2_w, ϕ, and μ_h. To make this more explicit, consider the smoothed value $\mathbb{E}\left[h_1|\boldsymbol{y}\right]$. When we assume prior distributions for parameters σ^2_w, ϕ, and μ_h, it is defined as:

$$\mathbb{E}\left[h_1|\boldsymbol{y}\right]=\int h_1 p\left(\boldsymbol{h},\phi,\mu_h,\sigma^2_w|\boldsymbol{y}\right)dh_1\dots dh_n d\phi d\mu_h d\sigma^2_w, \tag{5}$$

which can be thought of as a weighted average of possible values of h_1, where the weights are how likely each value of h_1 is given the observed data, a set of parameter values, and a set of other state values $h_2,\dots,h_n$ and we average over all possible parameter values and other state values. When we are thinking of this as a Bayesian model, we will often refer to $p\left(\boldsymbol{h},\phi,\mu_h,\sigma^2_w|\boldsymbol{y}\right)$ as the **posterior distribution**, which is proportional to the joint likelihood of the data $\boldsymbol{y}$, the states and the parameters:

$$\begin{aligned} p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w | \boldsymbol{y}\right) &= \frac{p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w, \boldsymbol{y}\right)}{p\left(\boldsymbol{y}\right)} \\ &= \frac{p\left(\boldsymbol{y}|\boldsymbol{h}\right) p\left(\boldsymbol{h}|\mu_h, \phi, \sigma^2_w\right) p\left(\mu_h\right) p\left(\phi\right) p\left(\sigma^2_w\right)}{p\left(\boldsymbol{y}\right)}, \end{aligned}$$

where $p\left(\mu_h\right)$, $p\left(\phi\right)$, and $p\left(\sigma^2_w\right)$ are the densities corresponding to the prior distributions we assumed for μ_h, ϕ, and σ^2_w.

We can use the same MCMC methods we discussed previously to simulate values the states $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(m)}$ as well as the parameters $\phi^{(1)}, \dots, \phi^{(m)}$, $\mu_h^{(1)}, \dots, \mu_h^{(m)}$, and $\sigma_w^{2(1)}, \dots, \sigma_w^{2(m)}$ according to $p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w | \boldsymbol{y}\right)$. The MCMC methods proceed iteratively:

- Simulate $\boldsymbol{h}^{(i)}$ according to $p\left(\boldsymbol{h}|\boldsymbol{y}, \phi^{(i-1)}, \mu_h^{(i-1)}, \sigma_w^{2(i-1)}\right)$
- Simulate $\phi^{(i)}$ according to $p\left(\phi|\boldsymbol{y}, \boldsymbol{h}^{(i)}, \mu_h^{(i-1)}, \sigma_w^{2(i-1)}\right)$
- Simulate $\mu_h^{(i)}$ according to $p\left(\mu_h|\boldsymbol{y}, \boldsymbol{h}^{(i)}, \phi_h^{(i)}, \sigma_w^{2(i-1)}\right)$
- Simulate $\sigma_w^{2(i)}$ according to $p\left(\sigma^2_w|\boldsymbol{y}, \boldsymbol{h}^{(i)}, \phi^{(i)}, \mu_h^{(i)}\right)$.

Each of these steps can be handled by standard MCMC methods, because we can write out each conditional distribution up to a normalizing constant.

One we have collected m simulated values from $p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w | \boldsymbol{y}\right)$, we can approximate the smoothed means and variances of the states just by taking the sample means and variances of each set of simulated values, as described in (1). We can also use the m simulated values from $p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w | \boldsymbol{y}\right)$ to approximate the forecasts and their variances as described in (2), as well as the out-of-sample predictions of the states and their variances as described in (3). In order to understand what the data and priors combined suggest about the values of the parameters ϕ, μ_h, and σ^2_w, we might also want to examine the approximate marginal posterior distributions of the parameters, $p\left(\phi|\boldsymbol{y}\right)$, $p\left(\mu_h|\boldsymbol{y}\right)$, and $p\left(\sigma^2_w|\boldsymbol{y}\right)$. Approximations can be obtained by examining histograms and kernel densities of the draws $\phi^{(1)}, \dots, \phi^{(m)}$, $\mu_h^{(1)}, \dots, \mu_h^{(m)}$, and $\sigma_w^{2(1)}, \dots, \sigma_w^{2(m)}$, respectively. In particular, we might want to approximate the posterior means and variances of each parameter, e.g.

$$\mathbb{E}\left[\phi|\boldsymbol{y}\right] \approx \frac{1}{m}\sum_{i=1}^{m}\phi^{(i)} \text{ and } \mathbb{V}\left[\phi|\boldsymbol{y}\right] \approx \frac{1}{m}\sum_{i=1}^{m}\left(\phi^{(i)} - \frac{1}{m}\sum_{i=1}^{m}\phi^{(i)}\right)^{2}.$$

Choosing Priors

To actually take this approach, we need to choose prior distributions for the parameters. In practice, people often assume normal distributions for continuous parameters, gamma distributions for positive parameters, and beta prior distributions for parameters that constrained to an interval and select values for b_μ, B_μ, a_0, b_0, and $B_{\sigma^2_w}$ that reflect their prior beliefs about the parameters:

$$\mu_h \sim \mathcal{N}\left(b_\mu, B_\mu\right), \ \left(\phi+1\right)/2 \sim \mathcal{B}\left(a_0, b_0\right), \text{ and } \sigma^2_w \sim \mathcal{G}\left(\text{shape} = \frac{1}{2}, \text{rate} = \frac{1}{2B_{\sigma^2_w}}\right).$$

These priors have some interesting special cases and properties:

- When $a_0 = b_0 = 1$, we get a uniform prior for ϕ on $[-1, 1]$.
- This gamma prior for σ^2_w is the same as a χ^2_1 prior for $\sigma^2_w/B_{\sigma^2_w}$.

Ideally, we would have well informed beliefs about likely values of all of the parameters, however in practice we are often quite uncertain. For this reason, we will usually try to choose **uninformative priors**, which are very weakly concentrated about plausible values. If two different priors are both very uninformative, the posterior distribution $p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w|\boldsymbol{y}\right)$ will be similar under both priors because the data contributes more information to the posterior than the prior. To clarify what this means a bit, recall that we the posterior can be written as proportional to the product of some likelihood terms and some prior terms:

$$p\left(\boldsymbol{h}, \phi, \mu_h, \sigma^2_w|\boldsymbol{y}\right) = \frac{\overbrace{p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\mu_h, \phi, \sigma^2_w\right)}^{\text{information from data}}\overbrace{p\left(\mu_h\right)p\left(\phi\right)p\left(\sigma^2_w\right)}^{\text{information from prior}}}{p\left(\boldsymbol{y}\right)}.$$

One way to check how informative a prior is is to compare the posterior distribution for a parameter to the prior distribution - if the posterior distribution looks a lot like the prior even when we have a lot of data, then we might have used a prior that is more informative than we had intended. We might also try several different priors and compare results - the results for one prior look very different from the rest, then that prior might be stronger than we had intended.

In general, common uninformative priors for parameters of this model include:

- Set $b_\mu = 0$ and B_μ to be very large, e.g. $B_\mu = 1,000$, which gives a prior for μ_h that is weakly centered about 0.
- Set $a_0 = 1$ and $b_0 = 1$, which gives a prior for ϕ that is weakly centered about 0.
  - We need to be pretty careful with this prior distribution, because uninformative priors suggest that the state process is as likely to be white noise as it is to be non-stationary. This can be both implausible and computationally troublesome, which is why some software uses more informative priors for ϕ by default.
- Set $B_{\sigma^2_w} = 1$, which gives a prior for σ^2_w that is weakly centered about 1.

Whenever you use any software, you should check what the defaults are!

Appendix

Expectation-Maximization (EM) Maximum Marginal Likelihood for the Stochastic Volatility Model One other way we can compute maximum likelihood estimates of σ^2_w, ϕ, and μ_h is to use the expectation-maximization (EM) algorithm, which is an algorithm for maximizing a function that corresponds to an integral over some latent variables, which in this case are the latent states h_t. The EM algorithm allows us to maximize the marginal likelihood in (4) by maximizing $\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\mu_h, \phi, \sigma^2_w\right)\right)|\boldsymbol{y}\right]$ using an iterative procedure. It isn't immediately obvious how this helps us, but we will get a sense by working simplifying the conditional expectation of the joint log-likelihood:

$$
\begin{aligned}
&\mathbb{E}\left[\log\left(p\left(\boldsymbol{y}|\boldsymbol{h}\right)p\left(\boldsymbol{h}|\mu_h,\phi,\sigma_w^2\right)\right)|\boldsymbol{y}\right] = \\
&\mathbb{E}\left[\log\left(p\left(y_1|h_1\right)p\left(h_1|\mu_h,\phi,\sigma_w^2\right)\prod_{t=2}^{n}p\left(y_t|h_t,\right)p\left(h_t|h_{t-1},\phi,\sigma_w^2\right)\right)|\boldsymbol{y}\right] = \\
&K-\frac{1}{2}\log\left(1-\phi^2\right)-\frac{n}{2}\log\left(\sigma_w^2\right)+ \\
&\mathbb{E}\left[-\frac{1}{2}\left(\sum_{t=1}^{n}h_t+\frac{y_t^2}{\exp\left\{h_t\right\}}\right)-\frac{1-\phi^2}{2\sigma_w^2}\left(h_1-\mu_h\right)^2-\frac{1}{2\sigma_w^2}\left(\sum_{t=2}^{n}\left(\left(h_t-\mu_h\right)-\phi\left(h_{t-1}-\mu_h\right)\right)^2\right)|\boldsymbol{y}\right] = \\
&K-\frac{1}{2}\log\left(1-\phi^2\right)-\frac{n}{2}\log\left(\sigma_w^2\right)-\frac{n\left(1-\phi^2\right)\mu_h^2}{2\sigma_w^2}+ \qquad (6) \\
&-\frac{1}{2}\left(\sum_{t=1}^{n}\mathbb{E}\left[h_t|\boldsymbol{y}\right]+y_t^2\mathbb{E}\left[\exp\left\{-h_t\right\}|\boldsymbol{y}\right]\right)+ \\
&-\frac{1-\phi^2}{2\sigma_w^2}\left(\mathbb{E}\left[h_1^2|\boldsymbol{y}\right]-\mu_h\mathbb{E}\left[h_1|\boldsymbol{y}\right]\right)+ \\
&-\frac{1}{2\sigma_w^2}\left(\sum_{t=2}^{n}\left(\mathbb{E}\left[h_t^2|\boldsymbol{y}\right]+\phi^2\mathbb{E}\left[h_{t-1}^2|\boldsymbol{y}\right]\right)+ \right. \\
&-\frac{1}{2\sigma_w^2}\left(\sum_{t=2}^{n}\left(2\mu_h\left(\phi-1\right)h_t+2\phi\mu_h\left(1-\phi\right)\mathbb{E}\left[h_{t-1}|\boldsymbol{y}\right]-2\phi\mathbb{E}\left[h_th_{t-1}|\boldsymbol{y}\right]\right)\right)
\end{aligned}
$$

where $x_1 = \mu$. where K is a constant that doesn't depend on the data $\boldsymbol{y}$, or the latent states $\boldsymbol{h}$, or the parameters ϕ, σ_w^2, and μ_h. The problem we have is that (6) depends on conditional expectations that we don't have a closed form way to compute, even when the parameters are ϕ, σ_w^2, and μ_h fixed. This means that we don't have a speedy way to do the E-step of the EM-algorithm. There is a vast literature on approximating the E-step, but EM algorithms that use an approximate E-step tend to be very slow and may converge especially poorly so we won't consider them.