Brittleness of Hierarchical Bayesian Modeling with Real Data:
Challenges, Failures, and Theoretical Gaps

Aaron Luttman
Mathematics and Data Analysis Group
Defense Experimentation and Stockpile Stewardship – Nevada Operations
National Security Technologies, LLC

With contributions from: Marylesa Howard, Michael Fowler, Kevin Joyce,
Johnathan Bardsley, Jesse Adams, Maggie Hock

This work was done by National Security Technologies, LLC, under Contract
No. DE-AC52-06NA25946 with the U.S. Department of Energy and supported
by the Site-Directed Research and Development Program.
Bayesian Models for Linear Inverse Problems

Given the deterministic linear model

\[ Au = b, \]

we use the following 3 likelihood functions:

1) Gaussian Likelihood – Least Squares
2) Gaussian Likelihood – Weighted Least Squares
3) Poisson Likelihood

corresponding to the stochastic models

\[
\begin{align*}
    b &= Au + \varepsilon & \varepsilon &\sim \mathcal{N}(0, \sigma^2), \\
    b &= \text{Poiss}(Au).
\end{align*}
\]
Likelihoods

Thus, per the choice of stochastic model, the discrete likelihoods are

\[ p(b|u, \lambda) \propto \lambda^{n/2} \exp \left( -\frac{1}{2} \| A u - b \|^2 \right) \quad \text{LS} \]

\[ p(b|u) \propto \exp \left( -\frac{1}{2} \| C^{-1/2} (A u - b) \|^2 \right) \quad \text{WLS} \]

\[ p(b|u) \propto \exp \left( -(A u)^T 1 + b^T \log (A u) \right) \quad \text{Poisson}, \]

where \( C = \text{diag}(b) \), \( 1 \) is a vector of all 1’s, and \( \lambda \) is a scale parameter.

Here \( b \in \mathbb{R}^m \), \( u \in \mathbb{R}^n \), and \( A \in \mathbb{R}^{m \times n} \).
In this presentation we will discuss only priors of the type
\( u \sim \mathcal{N}(0, (\delta L)^{-1}) \), which take the form

\[
p(u) \propto |L|^{1/2} \delta^{n/2} \exp \left( -\frac{\delta}{2} u^T L u \right).
\]

Commonly, \( L = I \) (classical Tychonov regularization) or \( L \) is of the form

\[L = D^T W D,\]

where typically \( W = I \) (Laplacian regularization) or \( W \) is a linearization of the total variation seminorm.

Here \( \delta \) is a prior scale parameter (regularization parameter) that must be selected.
Hierarchical Modeling – Basic Formulation

Given the desire to not select a regularization parameter explicitly, we can also put a prior distribution on the “prior scale” parameter. This gives

\[ p(u, \lambda, \delta | b) \propto p(b | u, \lambda) p(u | \delta) p(\delta) p(\lambda). \]

We often like to choose \( p(\delta) \) and \( p(\lambda) \) to be uninformative, while also being conjugate priors, which make

\[ p(\delta) \propto \delta^{\theta - 1} \exp(-\phi \delta) \quad \text{and} \quad p(\lambda) \propto \lambda^{\alpha - 1} \exp(-\beta \lambda) \]

convenient choices. This adds four new hyperparameters that must be selected, but \((\alpha, \beta)\) and \((\theta, \phi)\) can be chosen to make these priors pretty flat.
Hierarchical Modeling – Basic Formulation

Given the desire to not select a regularization parameter explicitly, we can also put a prior distribution on the “prior scale” parameter. This gives

\[
p(u, \lambda, \delta | b) \propto p(b | u, \lambda) p(u | \delta) p(\delta) p(\lambda).
\]

We often like to choose \( p(\delta) \) and \( p(\lambda) \) to be uninformative, while also being conjugate priors, which make

\[
p(\delta) \propto \delta^{\theta-1} \exp(-\phi \delta) \quad \text{and} \quad p(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta \lambda)
\]

convenient choices. This adds four new hyperparameters that must be selected, but \((\alpha, \beta)\) and \((\theta, \phi)\) can be chosen to make these priors pretty flat.
Rather than imposing such stringent requirements on $L$, we can also place a distribution on $L$ and sample from that distribution.

Natural choices for the distribution on $L$ are the Wishart or inverse Wishart.

Taking $p(L) \sim \text{Wishart} (\Delta, \nu)$ gives

$$p(u, \lambda, \delta, L | b) \propto p(b | u, \lambda) p(u | L, \delta) p(L) p(\delta) p(\lambda)$$

$$= |L|^\frac{\nu - n}{2} \lambda^\frac{n}{2} + \alpha - 1 \delta^\frac{n}{2} + \theta - 1 \exp(-\beta \lambda - \phi \delta)$$

$$\times \exp\left(-\frac{\lambda}{2} \|Au - b\|^2 - \text{Tr}\left(\Delta^{-1}L\right) - \frac{\delta}{2} u^T Lu\right)^*$$

with hyperprior parameters $\{\alpha, \beta, \theta, \phi, \Delta, \nu\}$.  

*Corresponding formulas hold for the weighted least squares and Poisson likelihoods.
Sampling Schemes

In order compute samples from the posteriors, there are 2 scenarios:

1. For the Least Squares and Weighted Least Squares likelihoods
   - All the priors are appropriately conjugate to the likelihood,
   - we can compute the full conditionals, and
   - use straight hierarchical Gibbs sampling.

2. For the Poisson likelihood
   - The priors are not conjugate,
   - we use Metropolis-Hastings with a Gaussian acceptance proposal
     and
   - Gibbs sampling on the hyperprior (i.e. M-H with trivial
     acceptance).
Our Applications – Source Shape Reconstruction*

Problem: Reconstruct the shape of a radiation source from the image of an “L” Rolled Edge.

\[ b(s, t) = (Au)(s, t) = \int_{s'}^{s''} \int_{t'}^{t''} u(\eta, \mu) d\eta d\mu \]

- Any of the 3 likelihoods (Least squares best)
- Non-negativity constraints included
- Laplacian prior
- Not too ill-posed

Our Applications – Deconvolution*

**Problem:** Deconvolve the system impulse response (blur) out of a measured signal/image.

\[ b(s) = (Au)(s) = \int_{\Omega} a(s - t)u(t) \, dt \]

- Severely ill-posed
- Non-negativity can be included
- Any of the 3 likelihoods
- Any of the above priors, depending on what features in the signal are of greatest interest.


---

Managed and Operated by National Security Technologies, LLC
Nevada National Security Site
Our Applications – Abel Inversion*

**Problem:** Reconstruct volumetric density of cylindrically symmetric objects from single image projections.

\[ b(x, z) = (Au)(x, z) \]
\[ = 2 \int_{\sqrt{x^2 - r^2}}^{R} \frac{ru(r, z)}{\sqrt{x^2 - r^2}} \, dr \]

- Least square likelihood only
- Non-negativity can be included (but we don’t currently)
- Feature-preserving or -seeking priors are essential
- Not too ill-posed

Our Problems and Challenges

1. Selection of and sensitivity to hyperprior parameters,

2. Extracting meaning from computed samples of $L$ in the feature-seeking paradigm,

3. Determining MCMC chain convergence,

4. ...and many others that we don’t have the time to present here...
Theoretical Gaps – Hyperprior Parameters

What are the uses and benefits of hierarchical modeling?

For us: regularization parameter selection is something we want to avoid...better to let the data drive the selection.
“The argument is thus that two insensitive parameters are better than one sensitive parameter.”

–Fowler, Howard, Luttman, Mitchell, Webb

I certainly believe that this statement is TRUE, but is it RELEVANT?
Theoretical Gaps – Hyperprior Parameters

“The argument is thus that two insensitive parameters are better than one sensitive parameter.”

–Fowler, Howard, Luttman, Mitchell, Webb

I certainly believe that this statement is TRUE, but is it RELEVANT?
These are 8 Abel reconstructions with varying hyperprior parameters. Notice that the mean of the reconstruction samples is pretty consistent – and accurate – but the credible intervals vary wildly.
Theoretical Gaps – Hyperprior Parameters

As noted above, in our least-squares, “feature-seeking” model – modeling $\mathbf{L}$ with a Wishart – the full set of hyperprior parameters is

$$\{ \alpha, \beta, \theta, \phi, \Delta, \nu \}$$

$\lambda \Gamma$  $\delta \Gamma$  $\mathbf{L} - \text{Wishart}$

The Wishart parameters...

$\nu = n - 1$

$\Delta$ matters, but is not game-changingly sensitive (see right)
Theoretical Gaps – Hyperprior Parameters

As noted above, in our least-squares, “feature-seeking” model – modeling \( \mathbf{L} \) with a Wishart – the full set of hyperprior parameters is

\[
\{ \alpha, \beta, \theta, \phi, \Delta, \nu \}
\]

\( \lambda - \Gamma \) \quad \delta - \Gamma \quad \mathbf{L} - \text{Wishart} \)

The Wishart parameters...

\( \nu = n - 1 \)

\( \Delta \) matters, but is not game-changingly sensitive (see right)
Recall

\[ p(u, \lambda, \delta, L|b) \propto p(b|u, \lambda) p(u|L, \delta) p(L) p(\delta) p(\lambda) \]

It would be nice to be able to estimate the uncertainty in the “edge” locations in our reconstructions, based on the variation in \( L \).

**But how?**

Final \( L \) as image (above) and the main diagonal (left).
Our Challenges – Extracting Meaning from \( \text{L} \) Samples

1. Look at the sample progression of the main diagonal.
2. Characterize how the main diagonal captures the discontinuities in the reconstruction.
3. Other Ideas?
Our Challenges – Extracting Meaning from L Samples

1. Look at the sample progression of the main diagonal.
2. Characterize how the main diagonal captures the discontinuities in the reconstruction.
3. Other Ideas?
Our Problems – MCMC Chain Convergence

“Time” series of $\delta$ samples

Ostensibly, the goal is for the hyperparameter time series to look like white noise.

In practice, they don’t.
Our Problems – MCMC Chain Convergence

“Time” series of $\delta$ samples

Ostensibly, the goal is for the hyperparameter time series to look like white noise.

In practice, they don’t.

Do we care?