An Objective Prior for Hyperparameters in Normal Hierarchical Models

Jim Berger, Duke University

with Chengyuan Song and Dongchu Sun

(following up on work with Bill Strawderman and Dejun Tang)

4th Bayesian, Fiducial and Frequentist Workshop
Harvard University
May 3, 2017

History (personal) of Bayes/Frequentist interaction in shrinkage estimation of means; this is a reminder of the long history of BF

- Stein said “Shrink least squares estimates of means.”
- Bayesians said “Where should we shrink to?” and declared that the answer could be found in Bayesian hierarchical modeling.
- Efron and Morris said “We can do hierarchical modeling in an empirical Bayesian fashion, preserving a frequentist interpretation.”
- Bayesians said “There are problems in EB, especially in estimating variance components” (example to follow). “These problems can be corrected by utilizing full objective Bayesian analysis with MCMC.”
- Stein said “There is also a problem in covariance matrix estimation; eigenvalues of covariance matrices need to be shrunk together.”
- To correct the problems in EB (including covariance matrix estimation), Bayesians needed to develop good objective priors, for the HB hyperparameters, that would work for any normal hierarchical model.
- Doing this has required use of Brown’s frequentist tools of admissibility.

A prototypical normal hierarchical model:

For $i = 1, 2, \ldots, m$,

- $\boldsymbol{X}_i = \boldsymbol{\theta}_i + \epsilon_i, \qquad \epsilon_i \sim N_k(\cdot \mid \boldsymbol{0}, \boldsymbol{\Sigma}_i)$,
  the $\boldsymbol{X}_i$ and $\boldsymbol{\theta}_i$ being $k \times 1$ vectors, $k \geq 2$, with the $\boldsymbol{\Sigma}_i$ known.
  - If $\boldsymbol{X}_i = \boldsymbol{B}_i\boldsymbol{\theta}_i + \epsilon_i$ for given design matrix $\boldsymbol{B}_i$, transform to $\boldsymbol{X}_i^* = (\boldsymbol{B}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i$, which will be distributed as above.
- $\boldsymbol{\theta}_i = \boldsymbol{z}_i\boldsymbol{\beta} + \epsilon_i^*, \quad \epsilon_i^* \sim N_k(\cdot \mid 0, \boldsymbol{V})$,
  with the $\boldsymbol{z}_i$ being specified $k \times l$ covariate matrices.
  - $\boldsymbol{\beta}$ is an $l \times 1$ unknown 'hyper-mean' vector, $l \geq 2$;
  - $\boldsymbol{V}$ is an unknown $k \times k$ 'hyper-covariance matrix'.

Goal: Find good hyperpriors $\pi(\boldsymbol{\beta}, \boldsymbol{V}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{V})$ (independence assumed).

Why is a Bayesian approach to hierarchical modeling needed?

The simplest illustration: For $i = 1, \ldots, m$, suppose

$$X_i \sim \text{Normal}(\cdot \mid \theta_i, 1) \quad \text{and} \quad \theta_i \sim \text{Normal}(\cdot \mid \beta, V).$$

First – the difficulties of empirical Bayes and frequentist estimation of V:

The marginal density of X_i given (β, V) is found by integrating out the θ_i from the overall density, resulting in $X_i \sim \text{Normal}(\cdot \mid \beta, 1 + V)$, and yielding the marginal likelihood for the data $\boldsymbol{x} = (x_1, \ldots, x_m)$ and with $s^2 = \sum (x_i - \bar{x})^2$,

$$m(\boldsymbol{x} \mid \beta, V) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi(1+V)}} e^{\left[-\frac{(x_i - \beta)^2}{2(1+V)}\right]} \propto \frac{1}{(1+V)^{m/2}} \exp\left\{-\frac{n(\bar{x} - \beta)^2 + s^2}{2(1+V)}\right\}.$$

While the standard estimate $\hat{\beta} = \bar{x}$ is fine,

- if $s^2 < m$, the mle for V is $\hat{V}_{mle} = 0$;
- if $s^2 < m - 1$, the unbiased estimate of V, namely $\hat{V}_u = \frac{s^2}{m-1} - 1$, is negative.
- With numerous variance components, this is a common occurrence. Even here, for $m = 5$ and $V = 1$, $\Pr(S^2 < m) = 0.264$.

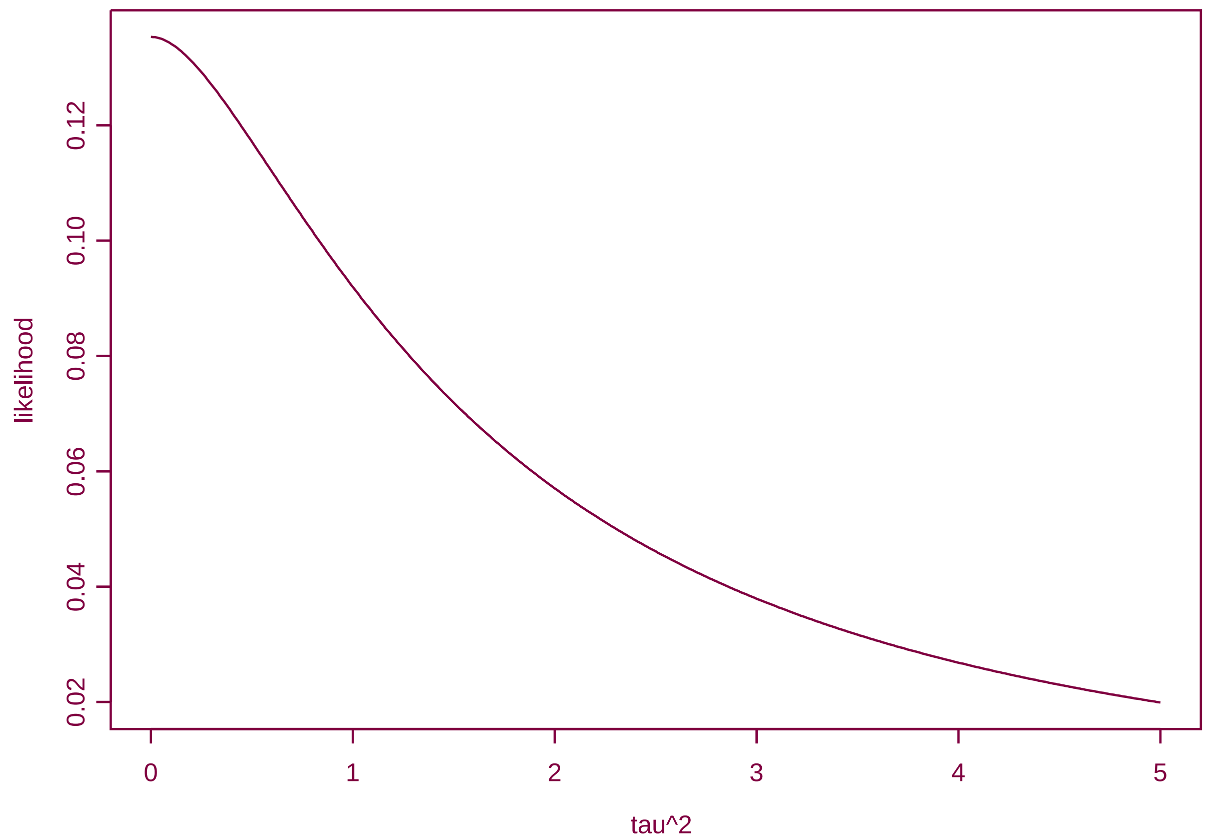

Figure 1: Marginal likelihood function of V (after integrating out β) when $m = 4$ and $s^2 = 4$ is observed. Note that it decreases slowly, indicating considerable uncertainty about V, even though the mle is 0.

Neglecting uncertainty in V affects the analysis in an incorrectly aggressive fashion.

Setting V to 0, when that is the MLE, is equivalent to setting $\theta_1 = \ldots = \theta_m$ (since the $\theta_i \sim Normal(\cdot \mid \beta, V)$), which is silly.

Frequentist methods have difficulty incorporating the uncertainty in V, because the maximum is achieved at a boundary.

Objective Bayes analysis

- leads to a posterior for V that reflects the uncertainty in the likelihood;
- can be easily implemented computationally for very complex hierarchical models using MCMC, more easily than likelihood methods.

But choice of 'hyperpriors' in hierarchical Bayesian analysis requires care.

- In the previous example, the *Jeffreys prior* for a mean and variance, $\pi(\beta, V) = 1/V$, results in an improper posterior. Commonly used *vague proper conjugate priors*, $\pi(\beta, V) \propto \boldsymbol{V}^{-(1+\epsilon)} e^{-\epsilon/V}$, will cause the posterior to concentrate near 0, having the same bad practical effect.

- Objective priors can also be too diffuse:
  - The constant prior for $\boldsymbol{\beta}$ is too diffuse for $k > 2$ (Stein, 1956, in the non-hierarchical setting; initiating the field of shrinkage estimation).
  - The constant prior for $\boldsymbol{V}$ yields a proper posterior only when $m > 2k$; this is much too large, since roughly k observations should make $\boldsymbol{V}$ identifiable.
    - Thus roughly k observations are needed just to correct for the over-diffuseness of the prior.
  - The same problems (or worse) occur for diffuse proper conjugate priors.

Addressing overdiffuseness through Admissibility and Inadmissibility

Consider estimating $\boldsymbol{\theta}$ by its posterior mean $\boldsymbol{\delta}^{\pi}(\boldsymbol{x})$, under mean squared error frequentist risk $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}) = E_{\boldsymbol{\theta}}^{\boldsymbol{X}}[(\boldsymbol{\theta} - \boldsymbol{\delta}^{\pi}(\boldsymbol{X}))^t(\boldsymbol{\theta} - \boldsymbol{\delta}^{\pi}(\boldsymbol{X}))]$.

Definition: $\boldsymbol{\delta}^{\pi}$ is admissible [inadmissible] if it cannot [can] be improved in risk (improvement meaning there is a $\delta^*(\boldsymbol{x})$ such that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^*) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi})$ for all $\boldsymbol{\theta}$ with strict inequality for some $\boldsymbol{\theta}$).

- Proper priors yield admissible estimators.
- Too diffuse improper priors yield inadmissible estimators.
- Priors ‘on the boundary of admissibility’ are typically exactly balanced between being too vague and too concentrated.

Proving Admissibility and Inadmissibility

Proofs are based on the results in Brown (1971): suppose that $m(\boldsymbol{x}) = \int\int\int f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \boldsymbol{V})\pi(\boldsymbol{\beta})\pi(\boldsymbol{V})d\boldsymbol{V}d\boldsymbol{\beta}d\boldsymbol{\theta}$ is the marginal density function, and define

$$\overline{m}(r) = \int m(\boldsymbol{x})d\phi(\boldsymbol{x}), \quad \underline{m}(r) = \int \frac{1}{m(\boldsymbol{x})}d\phi(\boldsymbol{x}),$$

where $\phi(\cdot)$ is the uniform probability measure on the surface of the sphere of radius $r = \|\boldsymbol{x}\|$.

Fact 1 *$\delta^\pi(\boldsymbol{x})$ is admissible if $\delta^\pi(\boldsymbol{x}) - \boldsymbol{x}$ is uniformly bounded and*

$$\int_c^\infty [r^{mk-1}\overline{m}(r)]^{-1}dr = \infty\,.$$

Fact 2 *$\delta^\pi(\boldsymbol{x})$ is inadmissible if*

$$\int_c^\infty r^{1-mk}\underline{m}(r)dr < \infty\,.$$

Results for $\pi(\boldsymbol{\beta})$:

- The constant prior $\pi(\boldsymbol{\beta}) = 1$ results in inadmissibility, except when $l = 2$.
- We recommend the prior $\pi(\boldsymbol{\beta}) \propto [1 + ||\boldsymbol{\beta}||^2]^{-(l-1)/2}$; it is excellent from the perspective of admissibility for all l. (It is not quite on the boundary of admissibility, but is close; the exponent $-(l-2)/2$ is the boundary.) To compute with this prior, use the equivalent representation

$$\boldsymbol{\beta} \mid \lambda \sim N_l(\cdot \mid \mathbf{0}, \lambda \boldsymbol{I}), \;\; \lambda \sim \lambda^{-1/2} e^{-1/2\lambda},$$

  - sample λ from its full conditional, the Inverse Gamma$(\cdot \mid (l-1)/2, 2/[1 + ||\boldsymbol{\beta}||^2])$ density;
  - given λ (and $\boldsymbol{V}$ and the $\boldsymbol{\theta}_i$), Gibbs sampling of $\boldsymbol{\beta}$ can be done from its full conditional, which is

$$N_l\left(\left(\frac{1}{\lambda}\boldsymbol{I} + \sum_{i=1}^{m} \boldsymbol{z}_i' \boldsymbol{V}^{-1} \boldsymbol{z}_i\right)^{-1} \sum_{i=1}^{m} \boldsymbol{z}_i' \boldsymbol{V}^{-1} \boldsymbol{\theta}_i, \left(\frac{1}{\lambda}\boldsymbol{I} + \sum_{i=1}^{m} \boldsymbol{z}_i' \boldsymbol{V}^{-1} \boldsymbol{z}_i\right)^{-1}\right).$$

Background on Covariance Matrix Priors

Consider i.i.d. multivariate normal data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where each column vector $\boldsymbol{x}_i$ arises from the $\mathrm{N}_k(\boldsymbol{x} \mid \mathbf{0}, \boldsymbol{\Sigma})$ density.

The sufficient statistic for Σ is easily seen to be $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i\, \boldsymbol{x}_i'$.

A commonly used prior for $\boldsymbol{\Sigma}$ is the inverse Wishart prior with mean proportional to the identity, for some specified a and b:

$$\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-a/2} \exp\{-\tfrac{1}{2}\mathrm{tr}[b\, \boldsymbol{\Sigma}^{-1}]\}.$$

A frequently used objective version of this prior (choosing $a = k + 1$ and $b = 0$) is the Jeffreys-rule prior

$$\pi^J(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(k+1)/2}.$$

Stein (1975, 1977) had shown that $\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{S}}{n}$ is seriously inadmissible, and can be improved by shrinking the eigenvalues of $\frac{\boldsymbol{S}}{n}$ together. $\hat{\boldsymbol{\Sigma}}$ happens to be

- the frequentist unbiased estimate,
- the maximum likelihood estimate,
- the Bayes rule using the Jeffreys-rule prior.

Thus, there is something seriously wrong with the Jeffreys-rule prior for a covariance matrix.

An interesting transformation: Write $\boldsymbol{\Sigma} = \boldsymbol{H}^t \boldsymbol{D} \boldsymbol{H}$, where $\boldsymbol{H}$ is an orthonormal matrix and $\boldsymbol{D}$ is a diagonal matrix with diagonal entries $d_1 > d_2 > \cdots > d_k$. Change of variables yields for the inverse Wishart prior

$$\pi(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma} \propto \left(\prod_{j=1}^{k} d_j{}^{-a/2} e^{-b/(2d_j)} \right) I_{[d_1 > \cdots > d_k]} \prod_{i<j} (d_i - d_j)\, d\boldsymbol{D}\, d\boldsymbol{H}\,;$$

for the Jeffreys-rule prior, $a = k+1$ and $b = 0$.

- Being uniform over (the rotation) $\boldsymbol{H}$ is natural.
- The term involving a product of constrained inverse gamma distributions for the d_j is natural.
- What about the term $\prod_{i<j}(d_i - d_j)$?
  - This assigns near zero density when any eigenvalues are close to each other, so that the prior pushes the eigenvalues away from each other.
  - This is why Stein got much better answers when he shrunk the eigenvalues of $\frac{\boldsymbol{S}}{n}$ together (the Jeffreys prior had forced them apart).
  - Inverse Wishart priors are also all likely bad.

A Modified Reference Prior: Berger, Strawderman and Tang (2005) proposed using the modified reference prior

$$\pi^{HR}(\boldsymbol{V}) = \frac{1}{|\boldsymbol{V}|^{(1-\frac{1}{2k})}\prod_{i<j}(d_i - d_j)}\, d\boldsymbol{V} = \frac{1}{|\boldsymbol{D}|^{(1-\frac{1}{2k})}}\, d\boldsymbol{D}\, d\boldsymbol{H}\,.$$

(defined as $\frac{1}{\sqrt{V}}$ if $k = 1$). This

- does not force the eigenvalues apart;
- results in a proper posterior when $m \geq 2$;
- is on the "boundary of admissibility."

New result: This prior results in admissible estimates.

Four Methods of Sampling From the Full Conditional of $\boldsymbol{V}$

Method 1. Yang and Berger (1994) used the Metropolized hit-and-run sampler for the log transformation of a covariance matrix.

Method 2. Direct Metropolis sampling of $\boldsymbol{V}$:

Step 0. Start with $\boldsymbol{V}^0 = \boldsymbol{I}$ or the marginal maximum likelihood estimate.

Step 1. At iteration r, generate $\boldsymbol{V}^* \sim$ Inverse Wishart$(\boldsymbol{W}(\boldsymbol{\theta},\boldsymbol{\beta}), m)$, where $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\theta},\boldsymbol{\beta}) = \sum_{i=1}^m (\boldsymbol{\theta}_i - \boldsymbol{Z}_i'\boldsymbol{\beta})(\boldsymbol{\theta}_i - \boldsymbol{Z}_i'\boldsymbol{\beta})^t$.

Step 2. Set $\boldsymbol{V}^{r+1} = \begin{cases} \boldsymbol{V}^* & \text{with probability } \alpha, \\ \boldsymbol{V}^r & \text{otherwise,} \end{cases}$

where

$$\alpha = \min\left\{1, \frac{\prod_{i<j}(d_i^* - d_j^*)}{\prod_{i<j}(d_i^r - d_j^r)} \cdot \frac{|\boldsymbol{V}^r|^{(k-1+k^{-1})/2}}{|\boldsymbol{V}^*|^{(k-1+k^{-1})/2}}\right\},$$

the d_i^* and d_i^r being the eigenvalues of $\boldsymbol{V}^*$ and $\boldsymbol{V}^r$, respectively.

Step 3. Iterate Steps 1 and 2 as needed.

Two newer methods are based on eigendecomposition of $\boldsymbol{V}$.
Defining $r = \frac{m}{2} + 1 - \frac{1}{2k}$, the full conditional for $\boldsymbol{V}$ can be written

$$\pi(\boldsymbol{V} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) \propto \frac{1}{|\boldsymbol{V}|^r \prod_{i<j}(d_i - d_j)} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{W})\right).$$

Writing $\boldsymbol{V} = \boldsymbol{O}'\boldsymbol{D}\boldsymbol{O}$, where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{D}$ is the diagonal matrix of ordered eigenvalues, it is shown in Yang and Berger (1994) that the full conditional can be transformed to

$$\begin{aligned}\pi(\boldsymbol{D}, \boldsymbol{O} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) &\propto \frac{1}{|\boldsymbol{D}|^r} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{O}\boldsymbol{D}^{-1}\boldsymbol{O}'\boldsymbol{W})\right) 1_{\{d_1 > d_2, \ldots, > d_k\}} d\boldsymbol{D} d\boldsymbol{O} \\ &= \frac{1}{|\boldsymbol{D}|^r} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{O}'\boldsymbol{W}\boldsymbol{O})\right) 1_{\{d_1 > d_2, \ldots, > d_k\}} d\boldsymbol{D} d\boldsymbol{O}.\end{aligned}$$

Method 3: Hoff (2009) developed a reasonable method for sampling from $\boldsymbol{O}$.

Method 4: A new Gibbs sampling method that produces *exact draws* from the full conditionals of the variables in $\boldsymbol{D}$ and $\boldsymbol{O}$ and mixes very well.

To sample $\boldsymbol{D}$ from the full conditional given $\boldsymbol{O}$ and $\boldsymbol{W}$, note that

$$\pi(\boldsymbol{D} \mid \boldsymbol{O}, \boldsymbol{W}) \propto \left[\prod_{i=1}^{k} \frac{1}{d_i^r} e^{-c_i/d_i}\right] 1_{\{d_1 > d_2, \ldots, > d_k\}} d\boldsymbol{D}\,,$$

where c_i is the (i, i) element of $\boldsymbol{O}'\boldsymbol{W}\boldsymbol{O}/2$. To remove the constraints, first transform to $v_i = 1/d_i$ (so that $v_1 < v_2, \ldots, < v_k$), then write $v_i = \sum_{j=1}^{i} \delta_j$; the δ_j are now unconstrained positive numbers. The full conditional of δ_j is (where $k_i = \sum_{j=i}^{k} c_j$)

$$\pi(\delta_j \mid \boldsymbol{O}, \boldsymbol{W}, \boldsymbol{\delta}_{(-j)}) \propto \left[\prod_{i=1}^{k} (\sum_{j=1}^{i} \delta_j)^{[r-2]}\right] e^{-k_j \delta_j}\,.$$

This is log-concave and hence easy to exactly sample by rejection sampling.

The full conditional of o_{ij} can be shown to be

$$[o_{ij} \mid others] \propto \exp\{c_{ij}\cos^2 o_{ij} + d_{ij}\cos\sin o_{ij} + e_{ij}\cos o_{ij} + f_{ij}\sin o_{ij}\}\,,$$

where c_{ij}, d_{ij}, e_{ij}, and f_{ij} are easily computable constants.

A simple rejection sampler to draw from this is as follows.

- Find the mle $\hat{o}_{ij}$. This requires solving a quartic equation.
- Compute the observed Fisher information $\hat{I}_{ij}$.
- Use, as a proposal $p(o_{ij})$, the t-distribution with 4 degrees of freedom and mean and variance $\hat{o}_{ij}$ and $\hat{I}_{ij}^{-1}$, constrained to the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$.
- Compute

$$K = \sup_{\{-\pi/2 < 0_{ij} < \pi/2\}} \frac{\pi(o_{ij})}{p(o_{ij})}\,.$$

- Do rejection sampling with probability $\pi(o_{ij})/[Kp(o_{ij})]$.

Table 1: The computational performance of the four methods (k=5)

Dimension = 5	Time/1000 iterations	#iterations to convergence	Convergence time
Hit and Run with $\log \boldsymbol{V}$	3.412(s)	1.3×10^7	4.4356×10^4(s)
Metropolis	2.268(s)	1.8×10^7	4.0824×10^4(s)
Hoff (+ new method for $\boldsymbol{D}$)	8.947(s)	8×10^5	7.158×10^3(s)
New method	10.091(s)	1.6×10^5	1.614×10^3(s)

Table 2: The computational performance of the four methods (k=10)

Dimension =10	Time/1000 iterations	#iterations to convergence	Convergence time
Hit and Run with $\log \boldsymbol{V}$	5.053(s)	2.8×10^7	1.415×10^5(s)
Metropolis	3.272(s)	3.4×10^7	1.112×10^5(s)
Hoff (+ new method for $\boldsymbol{D}$)	20.495(s)	4.3×10^6	8.813×10^4(s)
New method	34.091(s)	4×10^5	1.363×10^4(s)

Table 3. The mean square error(MSE) of method M_{ij}

$i = 1$: constant prior for $\boldsymbol{\beta}$; $i = 2$: $N(0, I)$ prior for $\boldsymbol{\beta}$; $i = 3$: suggested prior for $\boldsymbol{\beta}$

$j = 1$: constant for $\boldsymbol{V}$; $j = 2, 3$:Jeffreys and reference for $\boldsymbol{V}$; $j = 4$: suggested for $\boldsymbol{V}$

$k_1 = 4, m_1 = 10$, $k_2 = 5, m_2 = 15$; $\boldsymbol{\beta}_1 = \mathbf{1}_k$, $\boldsymbol{\beta}_2 = 50\mathbf{1}_k$; $\boldsymbol{V}_1 = \boldsymbol{I}_k$, $\boldsymbol{V}_2 = diag\{8k-7, \ldots, 9, 1\}$

	$k_1 \beta_1 V_1$	$k_1 \beta_2 V_1$	$k_1 \beta_1 V_2$	$k_1 \beta_2 V_2$	$k_2 \beta_1 V_1$	$k_2 \beta_2 V_1$	$k_2 \beta_1 V_2$	$k_2 \beta_2 V_2$
M_{11}	68.481	71.552	76.541	84.039	111.507	128.434	134.340	145.854
M_{12}	53.346	58.592	73.334	87.801	91.378	120.832	131.415	147.922
M_{13}	50.649	56.135	72.512	82.752	85.465	115.641	130.210	142.433
M_{14}	44.816	51.743	65.373	79.331	79.972	108.060	127.290	138.479
M_{21}	64.535	70.355	73.514	83.743	113.912	131.937	131.491	148.484
M_{22}	50.565	61.872	74.752	90.341	99.409	125.294	135.927	156.847
M_{23}	48.989	62.930	71.839	86.867	96.968	124.577	128.472	149.775
M_{24}	44.669	56.133	65.563	80.857	100.694	112.786	126.933	141.599
M_{31}	64.897	69.076	73.717	81.565	108.946	125.860	131.537	143.387
M_{32}	48.725	54.799	70.865	82.310	88.961	114.740	128.969	143.568
M_{33}	47.030	53.319	70.706	80.075	83.801	112.450	127.643	140.915
M_{34}	42.735	44.746	63.311	76.338	77.129	107.277	123.973	134.529

Higher levels of a hierarchical model: *The same priors for $\boldsymbol{\beta}$ and $\boldsymbol{V}$ should work for higher levels of a hierarchical model.*

Consider the following hierarchical model, where $m \geq 2$, $p \geq 1$, and $s \geq 2$; note that $k = ps$.

$$\begin{cases} \textit{Level } 1: & \boldsymbol{x}_i = \boldsymbol{\theta}_i + N_k(\mathbf{0}, \boldsymbol{I}_k), \quad i = 1, 2, \ldots, m; \\ \textit{Level } 2: & \boldsymbol{\theta}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + N_k(\mathbf{0}, \boldsymbol{V}), \quad \boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_s^t); \\ \textit{Level } 3: & \boldsymbol{\beta}_j = \boldsymbol{\eta} + N_p(\mathbf{0}, \boldsymbol{W}), \quad j = 1, 2, \ldots, s, \end{cases} \tag{1}$$

Here $\boldsymbol{Z}_i$ is an $k \times sp$ known matrix, and $(\boldsymbol{\eta}, \boldsymbol{V}, \boldsymbol{W})$ are unknown parameters.

For the unknown parameters $(\boldsymbol{\eta}, \boldsymbol{V}, \boldsymbol{W})$, utilize the independent priors,

$$\begin{aligned} \pi(\boldsymbol{\eta}) &\propto \frac{1}{(1 + ||\boldsymbol{\eta}||^2)^{(p-1)/2}}, \ \boldsymbol{\eta} \in \mathbb{R}^p, \\ \pi(\boldsymbol{V}) &\propto \frac{1}{|\boldsymbol{V}|^{1-1/(2k)} \prod_{1 \leq i < j \leq k} (v_i - v_j)}, \ \boldsymbol{V} > 0, \\ \pi(\boldsymbol{W}) &\propto \frac{1}{|\boldsymbol{W}|^{1-1/(2p)} \prod_{1 \leq i < j \leq p} (w_i - w_j)}, \ \boldsymbol{W} > 0\,. \end{aligned}$$

Theorem 1 *Assume that* $\mathbf{Z}$ *has rank* ps*. Then the posterior distribution is always proper if* $p \geq 2$*, and is proper when* $p = 1$ *if* $s = 3$ *and* $m \geq 5$*; if* $s = 4$ *and* $m \geq 3$*; and always for larger* s*.*

- This theorem likely generalizes to hierarchical models having any number of hierarchies. (We almost have a proof)
- It also is likely that the resulting Bayes estimator is admissible.
- The Gibbs sampling algorithm is essentially the same.

Summary

- Starting with the key insights of Stein into shrinkage estimation (of both means and covariance matrices);
- utilizing the hierarchical Bayesian framework to implement modeled shrinkage;
- employing objective Bayesian reference prior theory to understand the key needed property of covariance matrix priors (do not allow them to force eigenvalues apart);
- utilizing the theory of Brown (1971) to find the optimal versions of these priors (on the "boundary of admissibility");
- and finding efficient MCMC implementation schemes for these priors, that work at any level of a hierarchical model;

has produced a plausible answer to the 40+ year-old question of objective prior choice for any normal hierarchical model.

THANKS