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

- $X_i = \theta_i + \epsilon_i$, $\epsilon_i \sim N_k(\cdot | 0, \Sigma_i)$, the $X_i$ and $\theta_i$ being $k \times 1$ vectors, $k \geq 2$, with the $\Sigma_i$ known.
  - If $X_i = B_i \theta_i + \epsilon_i$ for given design matrix $B_i$, transform to $X_i^* = (B_i^t \Sigma_i^{-1} B_i)^{-1} B_i^t \Sigma_i^{-1} X_i$, which will be distributed as above.

- $\theta_i = z_i \beta + \epsilon_i^*$, $\epsilon_i^* \sim N_k(\cdot | 0, V)$, with the $z_i$ being specified $k \times l$ covariate matrices.
  - $\beta$ is an $l \times 1$ unknown ‘hyper-mean’ vector, $l \geq 2$;
  - $V$ is an unknown $k \times k$ ‘hyper-covariance matrix’.

**Goal:** Find good hyperpriors $\pi(\beta, V) = \pi(\beta)\pi(V)$ (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 \((\beta, V)\) is found by integrating out the \( \theta_i \) from the overall density, resulting in \( X_i \sim \text{Normal}(\cdot \mid \beta, 1 + V) \), and yielding the marginal likelihood for the data \( x = (x_1, \ldots, x_m) \) and with \( s^2 = \sum(x_i - \bar{x})^2 \),

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

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

- if \( s^2 < m \), the mle for \( V \) is \( \hat{V}_{\text{mle}} = 0 \);
- 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 $\beta$) 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 $\theta_i \sim Normal(\cdot | \beta, V)$), 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, \( \pi(\beta, V) \propto V^{-(1+\epsilon)} e^{-\epsilon/V} \), 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 \( \beta \) is too diffuse for \( k > 2 \) (Stein, 1956, in the non-hierarchical setting; initiating the field of shrinkage estimation).
  - The constant prior for \( V \) yields a proper posterior only when \( m > 2k \); this is much too large, since roughly \( k \) observations should make \( V \) 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 $\theta$ by its posterior mean $\delta^\pi(x)$, under mean squared error frequentist risk $R(\theta, \delta^\pi) = E^X_\theta [(\theta - \delta^\pi(X))^t(\theta - \delta^\pi(X))]$.

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

- 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
\[ m(x) = \int \int \int f(x \mid \theta) \pi(\theta \mid \beta, V) \pi(\beta) \pi(V) dV d\beta d\theta \]
is the marginal density function, and define
\[ \overline{m}(r) = \int m(x) d\phi(x), \quad \underline{m}(r) = \int \frac{1}{m(x)} d\phi(x), \]
where \( \phi(\cdot) \) is the uniform probability measure on the surface of the sphere of radius \( r = \|x\| \).

**Fact 1** \( \delta^\pi(x) \) is admissible if \( \delta^\pi(x) - x \) is uniformly bounded and
\[ \int_c^\infty \left[ r^{mk-1} \overline{m}(r) \right]^{-1} dr = \infty. \]

**Fact 2** \( \delta^\pi(x) \) is inadmissible if
\[ \int_c^\infty r^{1-mk} \underline{m}(r) dr < \infty. \]
Results for $\pi(\beta)$:

- The constant prior $\pi(\beta) = 1$ results in inadmissibility, except when $l = 2$.

- We recommend the prior $\pi(\beta) \propto [1 + ||\beta||^2]^{-(l-1)/2}$; 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

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

- sample $\lambda$ from its full conditional, the Inverse Gamma($\cdot \mid (l - 1)/2, 2/[1 + ||\beta||^2]$) density;

- given $\lambda$ (and $V$ and the $\theta_i$), Gibbs sampling of $\beta$ can be done from its full conditional, which is

$$N_l \left( \left( \frac{1}{\lambda} I + \sum_{i=1}^m z'_i V^{-1} z_i \right)^{-1} \sum_{i=1}^m z'_i V^{-1} \theta_i, \left( \frac{1}{\lambda} I + \sum_{i=1}^m z'_i V^{-1} z_i \right)^{-1} \right).$$
Background on Covariance Matrix Priors

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

The sufficient statistic for \(\Sigma\) is easily seen to be \(S = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'\).

A commonly used prior for \(\Sigma\) is the inverse Wishart prior with mean proportional to the identity, for some specified \(a\) and \(b\):

\[
\pi(\Sigma) \propto |\Sigma|^{-a/2} \exp\left\{ -\frac{1}{2} \text{tr}[b \Sigma^{-1}] \right\}.
\]

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

\[
\pi^J(\Sigma) \propto |\Sigma|^{-(k+1)/2}.
\]
Stein (1975, 1977) had shown that $\hat{\Sigma} = \frac{S}{n}$ is seriously inadmissible, and can be improved by shrinking the eigenvalues of $\frac{S}{n}$ together. $\hat{\Sigma}$ 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 $\Sigma = H^t D H$, where $H$ is an orthonormal matrix and $D$ is a diagonal matrix with diagonal entries $d_1 > d_2 > \cdots > d_k$. Change of variables yields for the inverse Wishart prior

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

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

- Being uniform over (the rotation) $H$ 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 $\frac{S}{n}$ 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

\[
\pi^{HR}(V) = \frac{1}{|V|^{(1-\frac{1}{2k})}} \prod_{i<j} (d_i - d_j) dV
\]

\[
= \frac{1}{|D|^{(1-\frac{1}{2k})}} dD dH .
\]

(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 $V$

**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 $V$:

- **Step 0.** Start with $V^0 = I$ or the marginal maximum likelihood estimate.

- **Step 1.** At iteration $r$, generate $V^\star \sim \text{Inverse Wishart}(W(\theta, \beta), m)$, where $W = W(\theta, \beta) = \sum_{i=1}^{m} (\theta_i - Z_i^l \beta)(\theta_i - Z_i^l \beta)^t$.

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

  where

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

  the $d_i^\star$ and $d_i^r$ being the eigenvalues of $V^\star$ and $V^r$, respectively.

- **Step 3.** Iterate Steps 1 and 2 as needed.
Two newer methods are based on eigendecomposition of \( V \). Defining \( r = \frac{m}{2} + 1 - \frac{1}{2k} \), the full conditional for \( V \) can be written

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

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

\[
\pi(D, O \mid \theta, \beta) \propto \frac{1}{|D|^r} \exp \left( -\frac{1}{2} \text{tr}(OD^{-1}O'W) \right) 1_{\{d_1 > d_2, \ldots, > d_k\}} dDdO
\]

\[
= \frac{1}{|D|^r} \exp \left( -\frac{1}{2} \text{tr}(D^{-1}O'WO) \right) 1_{\{d_1 > d_2, \ldots, > d_k\}} dDdO.
\]

Method 3: Hoff (2009) developed a reasonable method for sampling from \( O \).
Method 4: A new Gibbs sampling method that produces *exact draws* from the full conditionals of the variables in $D$ and $O$ and mixes very well.

To sample $D$ from the full conditional given $O$ and $W$, note that

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

where $c_i$ is the $(i, i)$ element of $O'WO/2$. 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 $\delta_j$ are now unconstrained positive numbers. The full conditional of $\delta_j$ is (where $k_i = \sum_{j=i}^{k} c_j$)

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

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 $(-\pi/2, \pi/2)$.
- Compute

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

- Do rejection sampling with probability $\pi(o_{ij})/[Kp(o_{ij})]$. 
Table 1: The computational performance of the four methods (k=5)

<table>
<thead>
<tr>
<th>Dimension = 5</th>
<th>Time/1000 iterations</th>
<th>#iterations to convergence</th>
<th>Convergence time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit and Run with log $V$</td>
<td>3.412(s)</td>
<td>$1.3 \times 10^7$</td>
<td>$4.4356 \times 10^4$(s)</td>
</tr>
<tr>
<td>Metropolis</td>
<td>2.268(s)</td>
<td>$1.8 \times 10^7$</td>
<td>$4.0824 \times 10^4$(s)</td>
</tr>
<tr>
<td>Hoff (+ new method for $D$)</td>
<td>8.947(s)</td>
<td>$8 \times 10^5$</td>
<td>$7.158 \times 10^3$(s)</td>
</tr>
<tr>
<td>New method</td>
<td>10.091(s)</td>
<td>$1.6 \times 10^5$</td>
<td>$1.614 \times 10^3$(s)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dimension =10</th>
<th>Time/1000 iterations</th>
<th>#iterations to convergence</th>
<th>Convergence time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit and Run with log $V$</td>
<td>5.053(s)</td>
<td>$2.8 \times 10^7$</td>
<td>$1.415 \times 10^5$(s)</td>
</tr>
<tr>
<td>Metropolis</td>
<td>3.272(s)</td>
<td>$3.4 \times 10^7$</td>
<td>$1.112 \times 10^5$(s)</td>
</tr>
<tr>
<td>Hoff (+ new method for $D$)</td>
<td>20.495(s)</td>
<td>$4.3 \times 10^6$</td>
<td>$8.813 \times 10^4$(s)</td>
</tr>
<tr>
<td>New method</td>
<td>34.091(s)</td>
<td>$4 \times 10^5$</td>
<td>$1.363 \times 10^4$(s)</td>
</tr>
</tbody>
</table>
Table 3. The mean square error (MSE) of method $M_{ij}$

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

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

$k_1 = 4, m_1 = 10, k_2 = 5, m_2 = 15; \beta_1 = 1_k, \beta_2 = 501_k; V_1 = I_k, V_2 = diag\{8k - 7, \ldots, 9, 1\}$

<table>
<thead>
<tr>
<th></th>
<th>$k_1\beta_1V_1$</th>
<th>$k_1\beta_2V_1$</th>
<th>$k_1\beta_1V_2$</th>
<th>$k_1\beta_2V_2$</th>
<th>$k_2\beta_1V_1$</th>
<th>$k_2\beta_2V_1$</th>
<th>$k_2\beta_1V_2$</th>
<th>$k_2\beta_2V_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{11}$</td>
<td>68.481</td>
<td>71.552</td>
<td>76.541</td>
<td>84.039</td>
<td>111.507</td>
<td>128.434</td>
<td>134.340</td>
<td>145.854</td>
</tr>
<tr>
<td>$M_{12}$</td>
<td>53.346</td>
<td>58.592</td>
<td>73.334</td>
<td>87.801</td>
<td>91.378</td>
<td>120.832</td>
<td>131.415</td>
<td>147.922</td>
</tr>
<tr>
<td>$M_{13}$</td>
<td>50.649</td>
<td>56.135</td>
<td>72.512</td>
<td>82.752</td>
<td>85.465</td>
<td>115.641</td>
<td>130.210</td>
<td>142.433</td>
</tr>
<tr>
<td>$M_{14}$</td>
<td>44.816</td>
<td>51.743</td>
<td>65.373</td>
<td>79.331</td>
<td>79.972</td>
<td>108.060</td>
<td>127.290</td>
<td>138.479</td>
</tr>
<tr>
<td>$M_{21}$</td>
<td>64.535</td>
<td>70.355</td>
<td>73.514</td>
<td>83.743</td>
<td>113.912</td>
<td>131.937</td>
<td>131.491</td>
<td>148.484</td>
</tr>
<tr>
<td>$M_{22}$</td>
<td>50.565</td>
<td>61.872</td>
<td>74.752</td>
<td>90.341</td>
<td>99.409</td>
<td>125.294</td>
<td>135.927</td>
<td>156.847</td>
</tr>
<tr>
<td>$M_{23}$</td>
<td>48.989</td>
<td>62.930</td>
<td>71.839</td>
<td>86.867</td>
<td>96.968</td>
<td>124.577</td>
<td>128.472</td>
<td>149.775</td>
</tr>
<tr>
<td>$M_{24}$</td>
<td>44.669</td>
<td>56.133</td>
<td>65.563</td>
<td>80.857</td>
<td>100.694</td>
<td>112.786</td>
<td>126.933</td>
<td>141.599</td>
</tr>
<tr>
<td>$M_{31}$</td>
<td>64.897</td>
<td>69.076</td>
<td>73.717</td>
<td>81.565</td>
<td>108.946</td>
<td>125.860</td>
<td>131.537</td>
<td>143.387</td>
</tr>
<tr>
<td>$M_{32}$</td>
<td>48.725</td>
<td>54.799</td>
<td>70.865</td>
<td>82.310</td>
<td>88.961</td>
<td>114.740</td>
<td>128.969</td>
<td>143.568</td>
</tr>
<tr>
<td>$M_{33}$</td>
<td>47.030</td>
<td>53.319</td>
<td>70.706</td>
<td>80.075</td>
<td>83.801</td>
<td>112.450</td>
<td>127.643</td>
<td>140.915</td>
</tr>
<tr>
<td>$M_{34}$</td>
<td>42.735</td>
<td>44.746</td>
<td>63.311</td>
<td>76.338</td>
<td>77.129</td>
<td>107.277</td>
<td>123.973</td>
<td>134.529</td>
</tr>
</tbody>
</table>
Higher levels of a hierarchical model: The same priors for $\beta$ and $V$ 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 $k = ps$.

$$
\begin{align*}
\text{Level 1:} & \quad x_i = \theta_i + N_k(0, I_k), \quad i = 1, 2, \ldots, m; \\
\text{Level 2:} & \quad \theta_i = Z_i \beta + N_k(0, V), \quad \beta^t = (\beta^t_1, \ldots, \beta^t_s); \\
\text{Level 3:} & \quad \beta_j = \eta + N_p(0, W), \quad j = 1, 2, \ldots, s,
\end{align*}
$$

Here $Z_i$ is an $k \times sp$ known matrix, and $(\eta, V, W)$ are unknown parameters.

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

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