An Objective Prior for Hyperparameters in Normal Hierarchical Models

Dongchu Sun, University of Missouri

Jointly with

James Berger, Duke University;

Chengyuan Song, East China Normal University

October 14, 2016

Motivation

- Hierarchical models are the workhorse of much of Bayesian analysis, yet there is uncertainty as to which objective priors to use for hyperparmeters (parameters at higher levels of the hierarchical model).
- Formal approaches to objective Bayesian analysis, such as the Jeffreys-rule or reference prior approach, are only implementable in simple hierarchical settings (one-way model).
- It is common to use less formal approaches, i.e., utilize formal priors from non-hierarchical models in hierarchical settings.
- This can be fraught with danger, however. For instance, non-hierarchical Jeffreys-rule priors for variances or covariance matrices result in improper posterior if they are used at higher levels of a hierarchical model.
- Such less formal approaches must be carefully evaluated, and not just from the perspective of posterior propriety.

Normal Hierarchical Model

Consider the hierarchical model,

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\theta}_i &\sim N_k(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i); && (1) \\ \boldsymbol{\theta}_i \mid \boldsymbol{\beta}, \boldsymbol{V} &\sim N_k(\boldsymbol{\beta}, \boldsymbol{V}), \quad i = 1, \ldots, m, && (2) \end{aligned}$$

where

- $\boldsymbol{y}_i$ are $k \times 1$ observation vectors,
- $\boldsymbol{\Sigma}_i$ are $k \times k$ known covariance matrices,
- $\boldsymbol{\theta}_i$ are $k \times 1$ unknown mean vectors,
- $\boldsymbol{\beta}$ is a $k \times 1$ unknown "hyper-mean" vector and
- $\boldsymbol{V}$ is an unknown $k \times k$ "hyper-covariance matrix".
- Here m is the block number and k is the dimension of $\boldsymbol{y}_i$.

A Note

- Often the data arise from linear models $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\theta}_i + \boldsymbol{e}_i$, where the $\boldsymbol{X}_i$ are known design matrices of covariates.
- In this situation, one can simply transform to $\boldsymbol{y}_i^* = (\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i$, which will be distributed as in (1), with transformed covariance matrix $\boldsymbol{\Sigma}_i^* = (\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)^{-1}$.

Hospital Example, I

- At hospital i, the observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ik})'$ are the sample averages of the costs of k different medical treatments;
- $\boldsymbol{\theta}_i$ is the corresponding unknown vector of true mean costs of the treatments at the hospital;
- $\boldsymbol{\Sigma}_i$ is the covariance matrix associated with these sample averages. If $\boldsymbol{\Sigma}_i$ is not known, it can typically be estimated from the observed data.

Hospital Example, II

- The means $\boldsymbol{\theta}_i$ are then modeled, conditional on a $l \times 1$ unknown 'hyper-mean' vector $\boldsymbol{\beta}$ and $k \times k$ unknown 'hyper-covariance matrix' $\boldsymbol{V}$, as

$$\boldsymbol{\theta}_i = \boldsymbol{z}_i\boldsymbol{\beta} + \boldsymbol{e}_i^*, \qquad \boldsymbol{e}_i^* \sim N_k(\mathbf{0}, \boldsymbol{V}), \tag{3}$$

- $\boldsymbol{z}_i$ are specified $k \times l$ covariate matrices and $l \geq 1$.

Hospital Example, III

- Consider the j^{th} coordinate, θ_{ij} of each $\boldsymbol{\theta}_i$; it refers to the cost of a certain medical treatment at i^{th} hospital.
- It is natural to model this as depending on p hospital characteristics, such as the number of patients receiving each treatment, the average severity of the condition of the patients for each of the treatments, the average income of the patients, etc.
- Denoting these characteristics for the j^{th} treatment at the i^{th} hospital as $\boldsymbol{v}_{ij} = (v_{ij1}, \ldots, v_{ijp})$;

Hospital Example, IV

- A reasonable model would be a regression model

$$\theta_{ij} = \boldsymbol{v}_{ij}\boldsymbol{\alpha}_j + e_{ij}, \quad \text{for } i = 1, \ldots, m\,,$$

- where $\boldsymbol{\alpha}_j$ is a column vector of weights determining the effect of hospital characteristics on the cost of treatment j, and e_{ij} is normal error. There would typically be a separate regression of this form for each treatment j, and stacking these regressions vertically leads to equation (3), where $l = kp$,

$$\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{v}_{i1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}_{i2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{v}_{ik} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{pmatrix}. \qquad (4)$$

Hospital Example, V

- If it were thought that each of the treatment cost means were independent, $\boldsymbol{V}$ could be taken as diagonal,
- but it is far more likely that there is considerable dependence, so that completely unknown $\boldsymbol{V}$ is reasonable.

Hospital Example, VI

- The above example suggests another possible level to the hierarchical modeling.
- In (3) we stop with the unknown hyperprior mean $\boldsymbol{\beta}$, but (4) suggests that it might be sensible to further model this mean, with a third level hierarchy, as

$$\boldsymbol{\alpha}_i \sim N_p(\cdot \mid \xi, \boldsymbol{\Omega})\,. \tag{5}$$

- Whether or not this is appropriate depends on precise context, but we will see that this analysis can be done with the methodology herein.

The hyperprior distributions: $\pi(\boldsymbol{\beta}, \boldsymbol{V}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{V})$

For $\boldsymbol{V}$, satisfies the following condition :

Condition 1. for $0 \leq c \leq 1$,

$$\frac{C_1}{|\mathbf{I} + \boldsymbol{V}|^{a_2 - a_1} |\boldsymbol{V}|^{a_1} [\prod_{i<j}(d_i - d_j)]^{1-c}}$$

$$\leq \pi(\boldsymbol{V}) \leq \frac{C_2}{|\boldsymbol{I} + \boldsymbol{V}|^{a_2 - a_1} |\boldsymbol{V}|^{a_1} [\prod_{i<j}(d_i - d_j)]^{1-c}}$$

Where C_1 and C_2 are positive constants, $d_1 > d_2 > \ldots > d_k > 0$ are the ordered eigenvalues of $\boldsymbol{V}$.

Hyperprior of $\boldsymbol{V}$: Common Cases

- Constant prior $(a_1 = a_2 = 0, c = 1)$: $\pi(\boldsymbol{V}) = 1$;
- Nonhierarchical indep. Jeffreys priors $(a_1 = a_2 = (k+1)/2,\ c = 1) : \pi(\boldsymbol{V}) = |\boldsymbol{V}|^{-(k+1)/2}$.
- Hierarchical indep. Jeffreys priors $(a_1 = 0,\ a_2 = (k+1)/2,\ c = 1)$: $\pi(\boldsymbol{V}) = |\boldsymbol{I} + \boldsymbol{V}|^{-(k+1)/2}$.
- Nonhierarchical indep. reference priors $(a_1 = a_2 = 1, c = 0)$: $\pi(\boldsymbol{V}) = [|\boldsymbol{V}| \prod_{i<j}(d_i - d_j)]^{-1}$; Yang & Berger (1994).
- Hierarchical indep. reference priors
  1. $a_1 = 0, a_2 = 1, c = 0$: $\pi(\boldsymbol{V}) = [|\boldsymbol{I} + \boldsymbol{V}| \prod_{i<j}(d_i - d_j)]^{-1}$;
  2. $a_1 = a_2 = 1 - \frac{1}{2k}$, $c = 0$.
     $\pi(\boldsymbol{V}) = [|\boldsymbol{V}|^{-(1-1/2k)} \prod_{i<j}(d_i - d_j)]^{-1}$; (recommended)

Hyperprior of $\boldsymbol{\beta}$

Three commonly considered priors for hyperparameter $\boldsymbol{\beta}$ are:

- Case 1 Constant prior: $\pi(\boldsymbol{\beta}) = 1$.
- Case 2 Conjugate prior: $\pi(\boldsymbol{\beta})$ is $N_k(\boldsymbol{\beta}^0, \boldsymbol{A})$, where $\boldsymbol{\beta}^0$ and $\boldsymbol{A}$ are subjectively specified.
- Case 3 Hierarchical prior:

$$\boldsymbol{\beta} \mid \lambda \sim N_k(\boldsymbol{\beta}^0, \lambda \boldsymbol{I}), \quad \lambda \sim \pi(\lambda), \lambda > 0,$$

where $\boldsymbol{\beta}^0$ is specified, $\pi(\lambda) \propto \lambda^{-1/2} e^{-1/2\lambda}$,
recommended default prior:

$$\pi(\boldsymbol{\beta}) \sim \frac{1}{[1 + \|\boldsymbol{\beta}\|^2]^{(k-1)/2}}.$$

Frequentist Justification in estimating $\boldsymbol{\theta}$

- Frequentist coverage problility of Bayesian credible intervals of a function of $\boldsymbol{\theta}$.
  What functions of $\boldsymbol{\beta}$? Each component of $\boldsymbol{\theta}$?
- Admissibility of Estimation of $\boldsymbol{\theta}$ under a loss function.

The definition of admissibility and inadmissibility.

The talk gives conditions under which the hierarchical Bayes estimate $\delta^{\pi}(\boldsymbol{y})$ (posterior mean) of $\boldsymbol{\theta}$ is admissible and inadmissible for quadratic loss

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}) = (\boldsymbol{\theta} - \boldsymbol{\delta}^{\pi}(\boldsymbol{y}))^{t} \boldsymbol{Q} (\boldsymbol{\theta} - \boldsymbol{\delta}^{\pi}(\boldsymbol{y})).$$

where $\boldsymbol{Q}$ is a known positive-definite matrix.
The performance of an estimator will be evaluated by the usual frequentist risk function

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = E_{\boldsymbol{\theta}}^{\boldsymbol{x}}[L(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi})]$$

The estimator $\boldsymbol{\delta}$ is inadmissible if there exists another estimator with risk function nowhere bigger and somewhere smaller. If no such better estimator exists, $\boldsymbol{\delta}$ is admissible.

Admissiblilty

- Berger, Strawderman and Tang (2005) approached the question of choice of hyperpriors in normal hierarchical models by looking at the frequentist notion of admissibility of resulting estimators.
- The motivation was that hyperpriors that are too diffuse result in inadmissible estimators, while hyperpriors that are concentrated enough result in admissible estimators.
- Hyperpriors that are 'on the boundary of admissibility' are sensible choices for objective priors, being as diffuse as possible without resulting in inadmissible procedures.
- The admissibility (and propriety) properties of a number of priors were considered in the paper, but no overall conclusion was reached as to a specific prior to recommend, in part because they were not able to prove admissibility for the leading candidate prior.

Two powerful results from Brown (1971)

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 $r = \|x\|$.

Result 1. If $\delta^\pi(x) - x$ is uniformly bounded and

$$\int_c^\infty [r^{mk-1} \overline{m}(r)]^{-1} dr = \infty \tag{6}$$

for some $c > 0$, then $\delta^\pi(x)$ is admissible.

Result 2. If

$$\int_c^\infty r^{1-mk} \underline{m}(r) dr < \infty \tag{7}$$

for some $c > 0$, then $\delta^\pi(x)$ is inadmissible.

A Theorem on Admissibility

Theorem 0.1

Consider a class of indepependent objective priors,

$$\pi_b(\boldsymbol{\beta}) \propto \frac{1}{[1+||\boldsymbol{\beta}||^2]^{(k+2b-2)/2}}, \quad \pi_a(\boldsymbol{V}) = \frac{1}{|\boldsymbol{V}|^a \prod_{i<j}(d_i - d_j)}. \quad (8)$$

If $m \geq 2$, $0 \leq b < 1$ *and* $1 - b/k \leq a < 1$, *the posterior mean of* $\boldsymbol{\theta}$ *is admissible.*

- The theorem extends the corresponding result in Berger, Strawderman and Tang (2005).
- It proves admissibility for the boundary $a = 1 - b/k$, that is crucial as it is priors on the boundary of admissibility which are sought.

Computation: Gibbs sampling for $\boldsymbol{\beta}$

To compute with the recommended prior, use the representation

$$\boldsymbol{\beta} \mid \lambda \sim N_k(\cdot \mid \mathbf{0}, \lambda \boldsymbol{I}),\ \pi(\lambda) \propto \lambda^{-1/2} e^{-1/2\lambda}\,.$$

- Sample λ from its full conditional, the Inverse Gamma$((k-1)/2, 2/[1+||\boldsymbol{\beta}||^2])$ density; if $k = 1$, this step is not needed as the hyperprior is constant.
- Given λ (and $\boldsymbol{V}$ and the $\boldsymbol{\theta}_i$), Gibbs sampling of $\boldsymbol{\beta}$ can be done from its full conditional, (when $k = 1$, set $\lambda = \infty$)

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

Note that the only cost in using the shrinkage prior as opposed to the constant prior is the need to sample λ, but this is an insignificant cost.

Previously suggested sampling methods for $\boldsymbol{V}$

The full conditional for $\boldsymbol{V}$ can be written

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

where $\boldsymbol{W}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{m}(\boldsymbol{\theta}_i - \boldsymbol{z}_i'\boldsymbol{\beta})(\boldsymbol{\theta}_i - \boldsymbol{z}_i'\boldsymbol{\beta})^t$.

Method 1. The method in Yang and Berger using the log transformation of a covariance matrix. This relies on sampling all of the covariance matrix at one time and hence can only work for modest dimensions (≤ 5).

Previously suggested Metropolis sampling methods for $\boldsymbol{V}$.

- *Step 0.* Start with $\boldsymbol{V}^0 = \boldsymbol{I}$ or the marginal MLE.
- *Step 1.* At iteration r, generate $\boldsymbol{V}^* \sim$ Inv-Wishart$(\boldsymbol{W}(\boldsymbol{\theta}, \boldsymbol{\beta}), m)$.
- *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.* After iterating Steps 1 and 2 N times, throw away the first M iterations keeping $(\boldsymbol{V}^{M+1}, \ldots, \boldsymbol{V}^N)$ as the samples from the posterior distribution.

This also samples the entire covariance matrix and hence will only work for modest dimensions.

A New Method

This is a potentially powerful new method that might work for higher dimensions (which the others do not).
Defining $r = (\frac{m}{2} + 1 - \frac{1}{k})$, the full conditional for $\boldsymbol{V}$ can be written

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

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

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

To sample from the full conditional dist. of $\boldsymbol{O}$ given $\boldsymbol{D}$, write

$$\boldsymbol{O} = (\boldsymbol{O}_{12}\boldsymbol{O}_{13}\cdots\boldsymbol{O}_{1p})(\boldsymbol{O}_{23}\cdots\boldsymbol{O}_{2p})\cdots(\boldsymbol{O}_{p-1,p})\boldsymbol{D}_\epsilon\,,$$

with $\boldsymbol{O}_{ij}$ being a simple orthogonal matrix such as

$$\boldsymbol{O}_{ij} = \boldsymbol{O}_{ij}(o_{ij}) = \begin{array}{c} \\ i \\ \\ j \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & i & & j \end{array} \\ \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \cos o_{ij} & 0 & -\sin o_{ij} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sin o_{ij} & 0 & \cos o_{ij} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \end{array}, \qquad (10)$$

where $-\pi/2 < o_{ij} \leq \pi/2$, and $\boldsymbol{D}_\epsilon$ being a diagonal matrix with diagonal elements ± 1 (see Anderson, Olkin, and Underhill, 1987). Insert this in (9), and consider Gibbs sampling of the o_{ij}.

- No need for sampling $\boldsymbol{D}_\epsilon$.
- The full cond. pdf of o_{ij} can be written in the form (for constants a, b, c depending on the other parameters and $\boldsymbol{W}$),

$$\pi(o_{ij}) = \pi(o_{ij} \mid \boldsymbol{D}, \{o_{jk} : jk \neq ij\}) \propto e^{[a\cos^2(o_{ij}) + b\cos(o_{ij}) + c\sin(o_{ij})]}.$$

- A simple rejection sampler to draw from this can be constructed as follows:
- Find the mle $\hat{o}_{ij}$. This requires solving a quartic equation.
- Compute the observed Fisher information $\hat{I}_{ij}$ (i.e., compute the second derivate of $-\log \pi(o_{ij})$ and plug in the mle).
- 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 $\pi(o_{ij})/[Kp(o_{ij})]$.

Comments

- We have compare the computational methods. The Gibbs Sampling of the new method converge to the target distribution quite fast.
- We have done the simulation study to see the frequentist performance of the recommend priors with other priors.
- We complete the story and propose a particular objective prior for use in normal hierarchical models, based on considerations of admissibility, ease of implementation (including computational considerations), and performance.
- The method will be extended to three level normal hierarchical models.