MARGINALIZED DEVIANCES FOR
ASSESSING THE NUMBER OF
HIERARCHICAL STAGES AND ALIGNING
MULTILEVEL MODELS

A.E. Gelfand, M. Trevisani

2000.7

Dipartimento di Scienze Statistiche
Università degli Studi
Via S. Francesco, 33
35121 Padova

Maggio 2000
Marginalized Deviances for Assessing the Number of Hierarchical Stages and Aligning Multilevel Models
by Alan E. Gelfand, and Matilde Trevisani

Abstract

Recent computational advances have made it feasible to fit hierarchical models in a wide range of complex applications. In the process, the question of model determination arises. With regard to model selection we consider the problem of comparing complex hierarchical models with differing number of levels. In hierarchical models various levels of marginalization can be considered, with complete marginalization leading to prior predictive distribution. The Bayes Factor, the prior predictive idea, in fact, marginalizes over all parameters and, then, prior predictive distributions can be compared. However, there are reasons not in favor of the Bayes factor and there are alternative model choice criteria based upon the likelihood. But, then, the question is which likelihood to use? It is certainly convenient the use of the first stage likelihood because no marginalization is required, but then how to compare fairly models with differing number of hierarchical levels? We illustrate a prior Deviance inequality that clarifies the direction of adjustment and the quantity which needs to be calculated in comparing models.

We finally consider posterior calculation of the mentioned Deviance inequality and compare it with the prior finding. Till now prior-posterior comparison has been used for outlier detection. Our perspective is, instead, to assess the need for additional hierarchy. This is connected with various

---

1 Alan E. Gelfand is Professor in the Department of Statistics at the University of Connecticut, Storrs, CT, 06269. Matilde Trevisani is a Ph.D. candidate in the Department of Statistical Sciences at the University of Padova, Padova, Italy, 35121. The work of the first author was supported in part by NSF grant DMS 96-25383, while that of the second author was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07750 and National Institutes of Allergy and Infectious Diseases (NIAID) Grant R01-AI41966.
recent works which also utilize the log likelihood to study adequacy or criticism of stages in a hierarchical model.

\textit{Key words:} hierarchical models; model selection; Bayes Factor; Deviance inequality; stage failure.

1 Correction via prior deviance inequality

The general idea can be described using a two-stage model as a starting point.

Let’s consider the two-stage hierarchical Bayesian model\(^2\)

\[
[Y \mid \Theta_1] \quad [\Theta_1 \mid \Theta_2] \quad [\Theta_2]
\]

where \(\Theta_1\) denotes the first stage parameters, \(\Theta_2\) the second stage parameters, and \([\Theta_2]\) is proper.

Also let \(L(\Theta_1; Y)\) denote the first stage likelihood and define the marginalized likelihood

\[
L(\Theta_2; Y) = \int L(\Theta_1; Y) [\Theta_1 \mid \Theta_2] d\Theta_1
\]

\[
= E_{[\Theta_1 \mid \Theta_2]} L(\Theta_1; Y)
\]

Now, if the first stage likelihood is log-concave in \(\Theta_1\) (Wedderburn, 1974; Dellaportas and Smith, 1993, [6]),\(^3\) then, applying the Jensen’s inequality, it follows that

\[
E_{[\Theta_2]} lnL(\Theta_2; Y) = E_{[\Theta_2]} lnE_{[\Theta_1 \mid \Theta_2]} L(\Theta_1; Y)
\]

\[
\geq E_{[\Theta_2]} E_{[\Theta_1 \mid \Theta_2]} lnL(\Theta_1; Y)
\]

\(^2\)Square brackets denote densities.

\(^3\)This happens, for instance, in one parameter exponential families with canonical link and more generally as well.
\[ E[\Theta_2] \ln L(\Theta_2; Y) - E[\Theta_1] \ln L(\Theta_1; Y) \geq 0 \] (4)

indicates the penalty or adjustment to make fair a comparison between models with differing number of levels. Indeed the above prior inequality clarifies that, if we use the marginal likelihood, we expect a larger log likelihood than if we use the first stage likelihood.

A comparison with the Bayes Factor We introduce, at this point, the (log) Bayes Factor as an alternative to the difference (4).

The Bayes Factor is a commonly used criterion in Bayesian model choice. It is a summary of the evidence provided by the data in favor of one statistical model, as opposed to another. In this context, both the two model components constituting a Bayesian model, \((f(y|\theta), \pi(\theta))\), can be varied. The case where the joint density \(f(y|\theta)\) or likelihood \(L(\theta; y)\) is held fixed while \(\pi(\theta)\) is varied to assess the sensitivity of the posterior to such prior variation is referred to as Bayesian robustness. Our intent here is, instead, model comparison with respect to different likelihoods.

The formal Bayesian model choice procedure goes as follows. Let \(w_i\) be the prior probability of model \(M_i, i = r, f\) with \(r\) standing for reduced and \(f\) standing for full, and \(f(y|M_i)\) be the (marginal or) predictive distribution for model \(M_i\), i.e.,

\[ f(y|M_i) = \int f(y|\theta_i, M_i) \pi(\theta_i|M_i) d\theta_i. \]
If \( w_i = 1/2 \), we use the Bayes factor (of \( M_r \) with respect to \( M_f \))

\[
BF = f(y|M_r)/f(y|M_f).
\]

We apply, now, the BF in a non proper way, that is, comparing two models incorporated in a single hierarchical model.

If the two-stage model (1) corresponds to \( M_f \), and the model \([y|\theta_2][\theta_2]\) to \( M_r \), then

\[
f(y|M_f) = \int f(y|\theta_1)|\theta_1|
\]

\[
= E_{\theta_1} L(\theta_1; y),
\]

and

\[
f(y|M_r) = \int f(y|\theta_2)|\theta_2|
\]

\[
= E_{\theta_2} L(\theta_2; y),
\]

(5)

but it is also true that

\[
f(y|M_f) = \int f(y|\theta_1) \pi(\theta_1) \, d\theta_1
\]

\[
= \int_{\theta_1} \int_{\theta_2} f(y|\theta_1) \pi(\theta_1|\theta_2) \pi(\theta_2) \, d\theta_2 \, d\theta_1
\]

\[
= \int_{\theta_2} \left[ \int_{\theta_1} f(y|\theta_1) \pi(\theta_1|\theta_2) \, d\theta_1 \right] \pi(\theta_2) \, d\theta_2
\]

\[
= \int_{\theta_1} f(y|\theta_2) \pi(\theta_2) \, d\theta_2
\]

\[
= f(y|M_r),
\]

(6)
hence,  \( \ln BF = 0 \), as it must hold because we are within one single model.

We put in comparison inequality (4) with this last result,

\[
E_{\theta_1, \theta_2} \ln \frac{f(y|\theta_2)}{f(y|\theta_1)} \geq 0, \quad \ln \frac{\int f(y|\theta_2)[\theta_2]}{\int f(y|\theta_1)[\theta_1]} = 0,
\]

in order to stress, again, the different content of the two diagnostics. (??) is more suited to model checking, in particular failure for each level. \( \ln BF \) is, correctly, a tool for model selection.

A hint on some penalized likelihood ratio tests  Prior BF resembles B information criterion (BIC), a particular case of penalized likelihood ratio test statistics. We need, for clarification, to turn back to some classical notions.

In the Newman-Pearson framework, optimal tests exist for the choice between two parametric fully specified models. The most common unambiguous specification of a null and alternative hypothesis, \( H_1 \) and, respectively, \( H_2 \), arises when model \( M_1 \) is a parametric submodel of \( M_2 \). We continue to indicate them as \( M_r \) and \( M_f \).

In the general case the LRT is commonly used: \( H_1 \) is rejected in favor of \( H_2 \) when \( LRT_{r/f} = \frac{L(\hat{\theta}_1)}{L(\hat{\theta}_2)} \) is less than some constant \( c \) (\( < 1 \)), chosen so that the test has level \( \alpha \) under \( H_1 \). If we write \( \lambda = -2 \ln(LRT_{r/f}) \), under mild conditions, as the sample size \( n \to \infty \) (with fixed \( p_i \)), \( \lambda \) is approximately distributed as \( \chi^2_{p_f - p_r} \), under \( H_1 \). With this approximation, inconsistency of the likelihood ratio test arises, i.e., \( \lambda \) tends to be too large. \( LRT_{r/f} \) gives too much weight to the full model.

As a result, various penalized versions of the likelihood ratio test have been proposed. We indicate the general class of these penalized forms as

\[
\lambda - k(n; p_f - p_r),
\]
where \( k(n; p_f - p_r) > 0 \) and increasing in both arguments. Then, the full model is clearly more penalized than the reduced one. In this class we recall the B (or Schwarz) information criterion (BIC) with \( k = \log(n) (p_f - p_r) \), that eliminates inconsistency, and other more common statistics of the form \( k = m(p_f - p_r) \) with, for instance, \( m = 1 \) (the proposal of Nelder and Wedderburn, 1972), \( m = \sqrt{2} \) (Aitkin, 1991), \( m = 2 \) (Akaike's information criterion, AIC, 1973).

The BF is closely related to the likelihood ratio statistics: the first eliminates the parameters \( \theta_i \) by integration, the second by maximization. Moreover there exists a parallel between some ‘Bayes Factors’ and some particular penalized LRTs, depending on the prior specification.

In fact, when the prior is proper normal with information proportional (in sample size) to that in the sample (Smith and Spiegelhalter, 1980)

\[ -2\ln BF = \lambda - \log(n) \cdot (p_f - p_r). \]

That is as \( n \to \infty \) (as well as \( v \to \infty \)), \( BF \to \infty \), favoring reduced models. More in general, given vague prior information, the BF provides too much support to the reduced model: its behavior is in complete contrast with that of the LRT. We will see one other parallel in section 4.6.1.

Our future intent is, therefore, to penalize somehow the \(-2\ln BF\) by means of our difference (4).

More precisely, we think to correct the ‘favor’ for reduced model by increasing the \( f(y|M_f) \) predictive measure, utilizing a marginalized form for it. Moreover, it should be of interest whether the left side of (4) is really a function, \( d(n; p_1 - p_2) \), which increases in both arguments. In this way, our correction would essentially eliminate typical inconsistency of (prior) BF.
Prior inequality with nuisance parameters  We can extend our basic result by incorporating a nuisance parameter, \( \eta \), in the model (1), e.g.,

\[
[Y \mid \Theta_1, \eta] \quad [\Theta_1 \mid \Theta_2, \eta] \quad [\Theta_2] \quad [\eta]
\]  

(7)

Define the marginal likelihood as

\[
L(\Theta_2, \eta; Y) = \int L(\Theta_1, \eta; Y) \mid \Theta_1 \mid \Theta_2, \eta \, d\Theta_1
\]

\[
= E_{[\Theta_1 \mid \Theta_2, \eta]} L(\Theta_1, \eta; Y).
\]  

(8)

Now

\[
E_{[\Theta_2, \eta]} \ln L(\Theta_2, \eta; Y) = E_{[\Theta_2, \eta]} \ln E_{[\Theta_1 \mid \Theta_2, \eta]} L(\Theta_1, \eta; Y)
\]

\[
\geq E_{[\Theta_2, \eta]} E_{[\Theta_1 \mid \Theta_2, \eta]} \ln L(\Theta_1, \eta; Y)
\]

\[
= E_{[\Theta_1, \eta]} \ln L(\Theta_1, \eta; Y),
\]  

(9)

again under log-concavity for \( L(\Theta_1, \eta; Y) \) in \( \Theta_1 \) for fixed \( \eta \).

Then, a priori

\[
E_{[\Theta_2, \eta]} \ln L(\Theta_2, \eta; Y) \geq E_{[\Theta_1, \eta]} \ln L(\Theta_1, \eta; Y).
\]

With this extension we can evaluate the criterion a priori in a case where the variance is unknown.

Multilevel version of prior inequality  Our inequality can be generalized for the multilevel hierarchical Bayesian models. Consider, for instance, a three-level structure

\[
[Y \mid \Theta_2] \quad [\Theta_1 \mid \Theta_2] \quad [\Theta_2 \mid \Theta_3] \quad [\Theta_3]
\]


We define the marginalized likelihood with respect to the two-level parameters as

\[
L(\Theta_2; Y) = \int L(\Theta_1; Y | \Theta_2) d\Theta_1 \\
= E_{\Theta_1 | \Theta_2} L(\Theta_1; Y),
\]

(10)

that is, with a formula identical to that for a two-stage model.

In addition, define the marginalized likelihood with respect to the three-level parameters as

\[
L(\Theta_3; Y) = \int L(\Theta_2; Y | \Theta_3) d\Theta_2 \\
= E_{\Theta_2 | \Theta_3} L(\Theta_2; Y),
\]

(11)

and, developing the integrand,

\[
= \int L(\Theta_1; Y | \Theta_2)[\Theta_2 | \Theta_3] d\Theta_1 d\Theta_2 \\
= E_{\Theta_1 | \Theta_2} L(\Theta_1; Y).
\]

(12)

Then, we obtain, from (11),

\[
E_{\Theta_3} \ln L(\Theta_3; Y) \geq E_{\Theta_2} \ln L(\Theta_2; Y),
\]

and, from (10),

\[
E_{\Theta_3} \ln L(\Theta_2; Y) \geq E_{\Theta_1} \ln L(\Theta_1; Y)
\]

We anticipate only that posterior computation of the difference is important as well.

In this regard our perspective is to assess the need for additional hierarchy through a posterior-prior
2 Computation of the prior deviance inequality

2.1 Analytic approach

The computation of the prior deviance inequality is analytically feasible with Normal first stage and, generally, with known variances.

**General Normal/Normal first/second stage** We consider the general linear Bayesian model of Lindley and Smith (1972, [5]), assuming all the variances are known. This general case is, then, applied to the Gaussian one-way ANOVA model.

Suppose, according to the notation of Lindley and Smith, that

\[ Y | \Theta_1 \sim N(A_1 \Theta_1, C_1) \quad \Theta_1, p_1 \times 1 \]
\[ \Theta_1 | \Theta_2 \sim N(A_2 \Theta_2, C_2) \quad \Theta_2, p_2 \times 1 \]
\[ \Theta_2 \sim N(\mu_\Theta, C_3) \]

Two marginalizations are, in the following, utilized,

\[ Y | \Theta_2 \sim N(A_1 A_2 \Theta_2, C_1 + A_1 C_2 A_1^T) \quad (13) \]
\[ \Theta_1 \sim N(A_2 \mu_\Theta, C_2 + A_2 C_3 A_2^T) \quad (14) \]
After some calculations, the prior difference results:

\[
E_{\Theta_2} \ln L(\Theta_2; Y) - E_{\Theta_1} \ln L(\Theta_1; Y) = \\
\log \frac{C_2}{C_2'} + \frac{1}{2} (Y - A_1 A_2 \mu) \Sigma^{-1} A_1 (C_2^{-1} + A_1^T \Sigma^{-1} A_1)^{-1} A_1^T C_2^{-1} (Y - A_1 A_2 \mu) + \frac{1}{2} tr(C_2^{-1} A_1 A_2) + C_2^{-1} A_1 (C_2^{-1} + A_1^T \Sigma^{-1} A_1)^{-1} A_1^T A_2^T \Sigma^{-1} A_1^T.
\]

**One sample Normal problem** Consider an illustrative example,

\[
Y_i \text{ i.i.d. } \sim N(\mu, \sigma^2), \\
\mu \sim N(\mu_0, \tau^2)
\]

where \( i = 1, \ldots, n \), \( \sigma^2 \) and \( \tau^2 \) are known, and \( \mu_0 \) is degenerate.

By sufficiency,

\[
\bar{Y} \sim N(\mu, \sigma^2/n).
\]

We make, in subsequence, the necessary steps to obtain the final result:

\[
\ln L(\mu_0; \bar{y}) = -\frac{1}{2} \ln (2\pi (\sigma^2/n + \tau^2)) - \frac{(\bar{y} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}
\]

\[
E\ln L(\mu_0) = -\frac{1}{2} \ln (2\pi (\sigma^2/n + \tau^2)) - \frac{(\bar{y} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}
\]

\[
\ln L(\mu; \bar{y}) = -\frac{1}{2} \ln (2\pi \sigma^2/n) - \frac{(\bar{y} - \mu)^2}{2\sigma^2/n}
\]

\[
E\ln L(\mu) = -\frac{1}{2} \ln (2\pi \sigma^2/n) - \frac{(\bar{y} - \mu)^2 + \tau^2}{2\sigma^2/n}
\]

Then, the prior inequality (4) is

\[
\frac{(\bar{y} - \mu_0)^2}{2} \left( \frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2} \right) + \frac{\tau^2}{2\sigma^2/n} \geq 1/2 \ln \left( 1 + \frac{\tau^2}{\sigma^2/n} \right),
\]
the inequality holding \( \ln(1 + x) \leq x \) if \( x > 0 \).

Note that, if we do not take expectation, but work directly with log-likelihood difference

\[
-\frac{1}{2} \ln(2\pi(\sigma^2/n + \tau^2)) - \frac{(\bar{y} - \mu_0)^2}{2(\sigma^2/n + \tau^2)} + \frac{1}{2} \ln(2\pi\sigma^2/n) + \frac{(\bar{y} - \mu)^2}{2\sigma^2/n} \\
= \frac{(\bar{y} - \mu)^2}{2\sigma^2/n} - \frac{(\bar{y} - \mu_0)^2}{2(\sigma^2/n + \tau^2)} - \frac{1}{2} \ln \left(1 + \frac{n\tau^2}{\sigma^2}\right),
\]

we obtain a distribution, which is \textit{a priori} of the general form \( C_0 + C_1 V \), where \( V \) has a noncentral \( \chi^2 \) distribution.

\textbf{ANOVA with replications} We consider, as simple example, a one-way ANOVA setting with \( n \) replications for each of \( g \) groups \( (N = gn) \). The model specification is supposed to be in the Lindley and Smith form. Thus, we have to substitute symbols as follow:

\[
A_1 = I_g \otimes 1_n, \quad C_1 = \sigma^2 I_N, \quad \theta_1 = (\mu_1, \ldots, \mu_g)^T = s\mu \\
A_2 = 1_g, \quad C_2 = \tau^2 I_g, \quad \theta_2 = \mu, \\
\mu_{\theta_2} = \mu_0, \quad C_3 = \omega^2.
\]

We only give the final result

\[
\frac{E \ln L(\mu) - E \ln L(s\mu)}{2} = \frac{g}{2} \left[ \log \frac{\sigma^2/n}{\sigma^2/n + \tau^2} + \frac{\tau^2}{\sigma^2/n} \left(1 + \frac{\omega^2}{\sigma^2/n + \tau^2}\right) \right] \\
+ \frac{\tau^2}{2\sigma^2/n} \frac{1}{\sigma^2/(n + \tau^2)} \sum_{i=1}^{g}(y_i - \mu_0)^2
\]

(16)

Note that the difference (16) is positive since \( \log(1 + x) \leq x \), with \( x = n\tau^2/\sigma^2 \). The difference is larger when \( \tau^2/(\sigma^2/n) \) is larger, as well as \( \omega^2/(\sigma^2/N + \tau^2/g) \) is larger.
The interpretation is rather direct if we consider for instance the simplest ANOVA model with

\( n = 1 \). In this case, we have

\[
\begin{align*}
\ln L(\mu; y) - \ln L(s\mu; y) &= n/2 \ln \frac{\sigma^2}{\sigma^2 + \tau^2} \\
&+ 1/2 \left( \frac{\sum_i (y_i - \mu_i)^2}{\sigma^2} - \frac{\sum_i (y_i - \mu)^2}{\sigma^2 + \tau^2} \right),
\end{align*}
\]

that is, the difference between the 2nd-stage likelihood and the 1st-stage likelihood is essentially

the difference between the deviances (with respect to inverted stages).

Then, at prior, before having estimated the individual means, the likelihood (deviance) for the

reduced model is expected to be higher (lower) than the one for the full model, since it is scaled

with respect to the total prior variance. Thus, (17) is higher as \( \tau^2 \) is larger with respect to \( \sigma^2/n \).

We try to imagine the behavior of the same deviance difference at posterior. When the likelihood

(deviance) is studied relatively to the posterior mean estimates, the full model, probably, is more

supported when the between-variance \( \tau^2 \) is large with respect to the within-variance \( \sigma^2/n \).

2.2 Simulation Approach

When closed-forms for the various likelihoods do not exist, the computing of expectations needs

simulation. We, briefly, sketch the simulation algorithm.

In the (7), both at prior and at posterior, the required samples are obtained under the fullest model.

Prior samples can be generated following the 'bottom-to top' order of the hierarchy; however, we

will always need proper priors at all stages. Posterior samples can be supplied, for example, by

Gibbs sampling output. Let see in detail:
• At prior, sample $\theta^*_2, \eta^*_1$, then $\theta^*_1$ from $[\theta_1|\theta^*_2, \eta^*_1]$, to obtain $\{(\theta^*_1, \theta^*_2, \eta^*_1), l = 1, \ldots, L\}$ from $[\theta_1, \theta_2, \eta]$. In this case $\{(\theta^*_1, \eta^*_1), l = 1, \ldots, L\}$ is a sample from $[\theta_1, \eta]$; $\{(\theta^*_2, \eta^*_1), l = 1, \ldots, L\}$ is a sample from $[\theta_2, \eta]$;

• At posterior, similarly, if we run a Gibbs sampling for the full model, we obtain $\{(\theta^*_1, \theta^*_2, \eta^*_1), l = 1, \ldots, L\}$ essentially from $[\theta_1, \theta_2, \eta|y]$. Therefore, we have $\{(\theta^*_1, \eta^*_1), l = 1, \ldots, L\}$, a sample from $[\theta_1, \eta|y]$; $\{(\theta^*_2, \eta^*_1), l = 1, \ldots, L\}$, a sample from $[\theta_2, \eta|y]$;

Then, different kinds of Monte Carlo integration are performed.

We need, at prior,

$$E_{\theta_1, \eta_1} lnL(\theta_1, \eta_1; y) \approx \frac{1}{L} \sum_{l=1}^{L} lnL(\theta^*_1, \eta^*_1; y),$$

$$E_{\theta_2, \eta_2} lnL(\theta_2, \eta_2; y) \approx \frac{1}{L} \sum_{l=1}^{L} lnL(\theta^*_2, \eta^*_1; y),$$

i.e., a Monte Carlo integration using the prior samples.

Similarly at posterior, we perform a Monte Carlo integration for $E_{\theta_1, \eta_1|y} lnL(\theta_1, \eta_1; y)$ and $E_{\theta_2, \eta_2|y} lnL(\theta_2, \eta_2; y)$, now, using posterior samples.

So, the remaining part of the computing is to obtain

$$lnL(\theta_1, \eta_1; y), \quad lnL(\theta_2, \eta_2; y),$$

The first quantity is immediately available as first-stage likelihood. To obtain the second quantity, instead, we again need a MC integration.

That is, given $\theta_2, \eta$, we must draw $\theta^*_1, b = 1, \ldots, B$ from $[\theta_1|\theta_2, \eta]$, after which we will approximate

$$lnL(\theta_2, \eta; y) \quad \text{by} \quad ln \left[ \frac{1}{B} \sum_{b=1}^{B} L(\theta^*_1, \eta; y) \right],$$
by Cauchy-Schwarz inequality. Hence,

$$\int \left( \int \frac{f(y|\theta_1)}{\pi(\theta_1|\theta_2, y)} \pi(\theta_2|y) d\theta_1 \right) \pi(\theta_2|y) d\theta_2 \geq \int f(y|\theta_2) \pi(\theta_2|y) d\theta_2,$$

i.e.,

$$\int f(y|\theta_1) \pi(\theta_1|y) d\theta_1 \geq \int f(y|\theta_2) \pi(\theta_2|y) d\theta_2,$$

so, in our case, the PoBF is always $\leq 1$, or, equivalently, $\ln\text{PoBF}$ is always $\leq 0$, regardless of data. The fact that PoBF is not identically equal to 0 within a single model constitutes a new criticism of it.

Turning back to our specific purpose of fairly aligning models in model comparison, we could attempt some penalty function analysis using the (21) inequality.

In this regard we recall that, asymptotically,

$$-2\ln\text{PoBF} = \lambda - (p_f - p_r) \cdot \log 2,$$

i.e., $-2\log\text{PoBF}$ behavior is similar to that of Akaike's information criterion (AIC, Akaike, 1973).

AIC = $\lambda - mv$ with $m = 2$ is another penalized version of the ratio of maximized likelihoods (see section 4.3). Nevertheless, because $m = \log 2 = 0.693 < 1$, $-2\log\text{PoBF}$ tends to favour complex models 'unduly', contrary to AIC (Aitkin, [11], p.117)

Then, to correct the asymptotic bias of the PoBF, a possible remedy could be to use the posterior predictive distribution with respect to the higher stage parameters of the hierarchical model when compared with a model with lower number of levels.

\(^4\)The above disequality results applying $[E(z)]^2 \leq E(z^2)E(z^2)$, after setting $x = f(y|\theta_1)$ and $z = 1.$
4.2 Analytic calculations

One sample Normal problem In this paragraph we take back the analysis of the example in section 1.4.

What we only need to obtain the (18) quantity, is the posterior distribution

$$\mu|\bar{y}, \mu_0 \sim N\left(\frac{n\tau^2\bar{y} + \sigma^2\mu_0}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

At posterior the difference (15), is again connected to a noncentral $\chi^2$ distribution. In fact, we can drop $C_0$ and $C_1$ and compare the prior for $(\mu - \bar{y})^2$ with the posterior for the same quantity.

The prior is $\tau^2\chi_1 ((\mu - \bar{y})^2/2\tau^2)$, the posterior is $\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\chi_1^2 ((\mu - \bar{y})^2/2\tau^2 \cdot \sigma^2/(n\tau^2 + \sigma^2))$.

With this change of scale the prior expectation is $\tau^2 + (\bar{y} - \mu_0)^2$, while the posterior expectation is

$$\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} + (\bar{y} - \mu_0)^2 \cdot \left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\right)^2$$

$$= \alpha\tau^2 + \alpha^2(\bar{y} - \mu_0)^2,$$

where $0 < \alpha = \sigma^2 + n\tau^2 < 1$.

So, in this simple case, expectation is always smaller under posterior than prior, and becomes smallest as $\alpha \to 0$, largest as $\alpha \to 1$. This finding agrees with intuition: if $\sigma^2 >> \tau^2$ ($\alpha \to 1$), then it is harder to criticize $\mu = \mu_0$; in the opposite extreme case, it is easier. In other words, the deviance with respect to the 1-stage parameter increases as $\sigma^2 >> \tau^2$, and tends to its value at prior.
ANOVA with replications  We only give the formula of posterior difference (18) for ANOVA models with replications. $E_{\mu|y}(\mu) - E_{\nu|y}(\nu)$ is

$$
\frac{g}{2} \left[ \log \frac{\gamma_n^2}{\sigma^2/n + \tau^2} + \frac{\tau^2}{\sigma^2/n + \tau^2} \frac{\sigma^2/n + \tau^2 + (g - 1)\omega^2}{\sigma^2/n + \tau^2 + g\omega^2} \right]
$$

$$
- \frac{1}{2} \left[ \frac{\tau^2}{(\sigma^2/n + \tau^2)^2} \sum_{i=1}^{g} (\bar{y}_i - \bar{y})^2 + \frac{g\tau^2}{(\sigma^2/n + \tau^2 + g\omega^2)^2} (\bar{y} - \mu_0)^2 \right]
$$

(22)

As it is apparent, the sign of our disequality is inverted at posterior. That is, it becomes negative supporting anyhow the full model. Moreover, the larger $\tau^2/(\sigma^2/n)$, $g$ and $\tau^2/(\sigma^2/n + g\omega^2)$ are, the larger is the evidence in favor of the full (or -multistage) hierarchical model.

Then the prior-posterior comparison of our deviance difference can be a flexible tool to decide whether an additional level is of benefit for model adequacy.

References


