Lecture 7: Approximations

Jim Berger

Duke University

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Laplace approximation
- BIC and AIC
- Prior-based BIC
- The effective sample size

I. Laplace Approximation

Goal: Analytically approximate the marginal density

$$m(\boldsymbol{x}\ ) = \int f(\boldsymbol{x}\ \ |\ \boldsymbol{\theta}\ )\pi(\boldsymbol{\theta}\ )d\boldsymbol{\theta}\ \ .$$

Preliminary 'nice' reparameterization: Choose a 'good' transformation to make the Laplace approximation as accurate as possible. In particular, all parameters should lie in $(-\infty, \infty)$.

- For variances, transform to $\nu = \log \sigma^2$ as the parameter.
- For a probability p, transform to, e.g., $\nu = \log \frac{p}{1-p}$.

Definitions: Let $\mathcal{L}(\boldsymbol{\theta}\ ) = \log f(\boldsymbol{x}\ \ |\ \boldsymbol{\theta}\ )$ denote the log-likelihood, maximized at the mle $\hat{\boldsymbol{\theta}}$, and let $\hat{\boldsymbol{I}}$ denote the *observed* Fisher Information matrix $I(\hat{\boldsymbol{\theta}}\ )$, where the Fisher Information matrix $\boldsymbol{I}\ (\boldsymbol{\theta}\ )$ has (i, j) element

$$I_{ij}(\boldsymbol{\theta}\ ) = -\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}\ )}{\partial\theta_i\, \partial\theta_j}\ .$$

Expanding $\mathcal{L}(\boldsymbol{\theta}\,) = \log f(\boldsymbol{x}\ \,|\,\boldsymbol{\theta}\,)$ about its maximum $\hat{\boldsymbol{\theta}}$, yields [a]

$$\mathcal{L}(\boldsymbol{\theta}\,) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}\,) - \frac{1}{2}\,(\boldsymbol{\theta}\ - \hat{\boldsymbol{\theta}}\,)^t \hat{\boldsymbol{I}}\ (\boldsymbol{\theta}\ - \hat{\boldsymbol{\theta}}\,)\,.$$

If $\pi(\boldsymbol{\theta}\,)$ is relatively flat near $\hat{\boldsymbol{\theta}}$, where $\mathcal{L}(\boldsymbol{\theta}\,)$ is non negligible[b],

$$\begin{aligned}
m(\boldsymbol{x}\,) &= \int f(\boldsymbol{x}\ \,|\,\boldsymbol{\theta}\,)\pi(\boldsymbol{\theta}\,)d\boldsymbol{\theta} \\
&\approx \pi(\hat{\boldsymbol{\theta}}\,) \int f(\boldsymbol{x}\ \,|\,\boldsymbol{\theta}\,)d\boldsymbol{\theta} \\
&\approx \pi(\hat{\boldsymbol{\theta}}\,)\, f(\boldsymbol{x}\ \,|\,\hat{\boldsymbol{\theta}}\,) \int \exp\left\{-\frac{1}{2}\,(\boldsymbol{\theta}\ - \hat{\boldsymbol{\theta}}\,)^t \hat{\boldsymbol{I}}\ (\boldsymbol{\theta}\ - \hat{\boldsymbol{\theta}}\,)\right\} d\boldsymbol{\theta} \\
&= \pi(\hat{\boldsymbol{\theta}}\,)\, f(\boldsymbol{x}\ \,|\,\hat{\boldsymbol{\theta}}\,)\,(2\,\pi)^{\frac{p}{2}}\,|\hat{\boldsymbol{I}}\,|^{-1/2}\,,
\end{aligned}$$

where p is the dimension of $\boldsymbol{\theta}$.

[a] we assume that $\mathcal{L}$ has continuous second partial derivatives and that the first partial derivative vanishes at $\hat{\boldsymbol{\theta}}$

[b] This will be true if $\mathcal{L}(\boldsymbol{\theta}\,)$ is highly peaked in a small neighborhood around $\hat{\boldsymbol{\theta}}$, which is typically de case for large n

Improved approximation: Define $\mathcal{L}_\pi(\boldsymbol{\theta}\,) = \log[f(\boldsymbol{x}\;\,|\,\boldsymbol{\theta}\,)\pi(\boldsymbol{\theta}\,)]$, and let $\hat{\boldsymbol{\theta}}_{\,\pi}$ and $\hat{\boldsymbol{I}}_{\,\pi}$ be the maximium and Hessian for this function. Then

$$m(\boldsymbol{x}\,) \approx \pi(\hat{\boldsymbol{\theta}}_{\,\pi})\, f(\boldsymbol{x}\;\,|\,\hat{\boldsymbol{\theta}}_{\,\pi})\, (2\,\pi)^{\frac{p}{2}}\, |\hat{\boldsymbol{I}}_{\,\pi}|^{-1/2}\,.$$

Comments:

The approximation is of order $n^{-1/2}$ (under regularity conditions).

Kass and Raftery (95) $\rightsquigarrow$ samples of size less than $5p$ worrisome, larger than $20p$ fine for 'usual' problems with 'good' parameterizations.

Use of *Fisher information*) itself, instead of $\hat{\boldsymbol{I}}$, is worse.

If applied to both the numerator and denominator of a Bayes factor, the approximation can be much better still (errors canceling).

Another approximation: For an objective estimation prior $\pi^O(\boldsymbol{\theta}\ )$,

$$\begin{aligned} m(\boldsymbol{x}\ ) &= \int f(\boldsymbol{x}\ \mid \boldsymbol{\theta}\ )\pi(\boldsymbol{\theta}\ )d\boldsymbol{\theta} \\ &= \int f(\boldsymbol{x}\ \mid \boldsymbol{\theta}\ )\frac{\pi(\boldsymbol{\theta}\ )}{\pi^O(\boldsymbol{\theta}\ )}\pi^O(\boldsymbol{\theta}\ )d\boldsymbol{\theta} \\ &\approx \frac{\pi(\hat{\boldsymbol{\theta}}\ )}{\pi^O(\hat{\boldsymbol{\theta}}\ )}\int f(\boldsymbol{x}\ \mid \boldsymbol{\theta}\ )\pi^O(\boldsymbol{\theta}\ )d\boldsymbol{\theta} \\ &= \frac{\pi(\hat{\boldsymbol{\theta}}\ )}{\pi^O(\hat{\boldsymbol{\theta}}\ )}\ m^O(\boldsymbol{x}\ )\,. \end{aligned}$$

This is useful when $m^O(\boldsymbol{x}\ ) = \int f(\boldsymbol{x}\ \mid \boldsymbol{\theta}\ )\pi^O(\boldsymbol{\theta}\ )d\boldsymbol{\theta}$ is available in closed form, as it applies with virtually no regularity conditions (Berger and Pericchi, 1996).

(First) Laplace approximation to Bayes factors: Apply Laplace expansion to numerator and denominator of B_{21} to get the Laplace approximation to B_{21}:

$$\begin{aligned} B_{21}^L &= \frac{\int f_2(\boldsymbol{x} \mid \boldsymbol{\theta}_2)\pi_2(\theta_2)d\theta_2}{\int f_1(\boldsymbol{x} \mid \boldsymbol{\theta}_1)\pi_1(\theta_1)d\theta_1} \\ &\approx \frac{\pi_2(\hat{\boldsymbol{\theta}}_2)\ f_2(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_2)\ |\hat{\mathbf{I}}_2|^{-1/2}\ (2\pi)^{p_2/2}}{\pi_1(\hat{\boldsymbol{\theta}}_1)\ f_1(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_1)\ |\hat{\mathbf{I}}_1|^{-1/2}\ (2\pi)^{p_1/2}}\,, \end{aligned}$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the m.l.e.'s for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (which have dimensions p_1 and p_2) and $\hat{\mathbf{I}}_1$ and $\hat{\mathbf{I}}_2$ are observed information matrices.

Large n and i.i.d. data: Then $\hat{\mathbf{I}}_i \approx n\mathbf{I}_i^*$, where $\mathbf{I}_i^*$ is the expected Fisher information for a single observation in M_i, and

$$B_{21}^L \approx \frac{f_2(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_2)}{f_1(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_1)} \cdot n^{-\frac{1}{2}(p_2-p_1)} \cdot \frac{|\mathbf{I}_2^*|^{-1/2}(2\pi)^{p_2/2}\pi_2(\hat{\boldsymbol{\theta}}_2)}{|\mathbf{I}_1^*|^{-1/2}(2\pi)^{p_1/2}\pi_1(\hat{\boldsymbol{\theta}}_1)}\,. \tag{1}$$

II. BIC and AIC

BIC (Bayes Information Criterion)

The Schwarz (78) approximation to Bayes factors is based on simply ignoring the last term in (1), because it is constant in n and so not as important as the first two terms:

$$B_{21}^{BIC} \approx \frac{f_2(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_2)}{f_1(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_1)} \cdot n^{-\frac{1}{2}(p_2 - p_1)} .$$

Raftery notes that, if one takes $\pi_i(\boldsymbol{\theta}_i)$ to be $N(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i, \mathbf{I}_i^{*(-1)})$ (a *unit information* prior centered at the mle), the third term in (1) is exactly 1.

Ref.: Schwarz (1978), Kass and Wasserman (1995), Dudley and Haughton (1997), Kass and Vaidyanathan (1992), Pauler (1998).

Typically, instead of using the Bayes factor directly, one uses the BIC criterion

$$BIC_i = -2 \log f_i(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_i) + p_i \log n \quad (\approx -2 \log m_i(\boldsymbol{x})),$$

so that

$$BIC_2 - BIC_1 = -2 \log B_{21}^{BIC}.$$

For multiple models, one just chooses that model with minimal BIC_i.

The main justification that Schwarz gave for BIC is that it is consistent, i.e. will select the correct model as $n \to \infty$. (The constant terms that BIC ignores are irrelevant asymptotically.)

Akaike's Information Criterion (AIC)

Criterion: AIC chooses the model M_i minimizing

$$AIC_i = -2 \log f(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_i) + 2\, p_i$$

here, the 'penalty' for dimension p_i is $2\, p_i$, so AIC has 'penalty' 2, whereas BIC has 'penalty' $\log n \rightsquigarrow$ AIC tends to choose larger models.

Bayes factor: AIC criterion corresponds to using

$$B_{21}^{AIC} \approx \frac{f_2(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_2)}{f_1(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_1)} \cdot e^{-2\,(p_2 - p_1)},$$

which cannot arise from any reasonable prior.

AIC versus BIC. Roughly:

AIC can be better than BIC if

- Complexity of models grows with n ($p_i \to \infty$ as $n \to \infty$)
- None of the models is correct and the goal is good prediction rather than deciding which of the models is true.

BIC is usually better than AIC if

- There is a set of fixed models and n is large, since then AIC is not even consistent.

  Example Testing a normal mean
  - $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$
  - To test $M_1 : \theta = 0$ vs $M_2 : \theta \neq 0$, $z = \sqrt{n}\,\bar{x}$
  - $B_{21}^{AIC} = e^{(\frac{1}{2}z^2 - 2)}$
  - Note that, as $n \to \infty$ under $M_1 : \theta = 0$, $z = \sqrt{n}\,\bar{x} \sim N(0, 1)$
  - Thus $B_{21}^{AIC} > 1$ with positive probability as $n \to \infty$, so that AIC is not consistent under M_1
- One of the models is (approximately) true.
- Simple models are desired for other reasons.

III. Prior-Based BIC

Prior-based BIC (PBIC)

(done with a SAMSI Social Sciences working group - Susie Bayarri, Woncheol Jang, Luis Pericchi, Surajit Ray, and Ingmar Visser; the context was "getting the model right, in structural equation modeling.")

Data: Independent vectors $\mathbf{x}_i \sim g_i(\mathbf{x}_i \mid \boldsymbol{\theta})$, for $i = 1, \ldots, n$.

Unknown: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$; $\hat{\boldsymbol{\theta}}$ is the MLE

Log-likelihood function: $l(\boldsymbol{\theta}) = \log f(\mathbf{x} \mid \boldsymbol{\theta}) = \log\left(\prod_{i=1}^{n} g_i(\mathbf{x}_i \mid \boldsymbol{\theta})\right)$ where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Usual BIC: $\text{BIC} \equiv -2l(\hat{\boldsymbol{\theta}}) + p \log n$ (Schwarz, 1978 AOS)

As $n \to \infty$ (with p fixed) this is an approximation (up to a constant) to twice the log of the Bayesian log likelihood for the model, $m(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, so that

$$m(\mathbf{x}) = c_\pi e^{-BIC/2}(1 + o(1)) .$$

Some of the problems with BIC

- Can the constant c_π from the prior be ignored?
- Problems with p.
  - What is p with random effects or latent variables?
  - What if p grows with n?
- Problems with n.
  - Is n the number of vector observations or the number of real observations?
  - What if different θ_i have different effective sample sizes?
  - What if observations vary significantly in information (as possible in mixture contexts, models with mixed continuous and discrete observations, ...)?

Example - Group means: For $i = 1, \ldots, p$ and $l = 1, \ldots, r$,

$$X_{il} = \mu_i + \epsilon_{il}, \quad \text{where} \quad \epsilon_{il} \sim N(0, \sigma^2)\,.$$

- It might seem that $n = pr$ but, if one followed Schwarz, one would have (defining $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^t$) that $\boldsymbol{X}_l = (X_{1l}, \ldots, X_{pl})^t \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, $l = 1, \ldots, r$, so that the 'sample size' appearing in BIC should be r.
- The 'effective sample size' for each μ_i is r, but the effective sample size for σ^2 is pr, so effective sample size is parameter-dependent.
- One could easily be in the situation where $p \to \infty$ but the effective sample size r is fixed.

Example - Random effects group means: $\mu_i \sim N(\xi, \tau^2)$, with ξ and τ^2 being unknown. What is the number of parameters? (see Pauler, 1998 Biometrika)

Example - Common mean, differing variances: Suppose $n/2$ of the Y_i are $N(\theta, 1)$, while $n/2$ are $N(\theta, 1000)$. Clearly the 'effective sample size' is roughly $n/2$.

Example - ANOVA: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{X}$ is a given $n \times p$ matrix of 1's and -1's with orthogonal columns, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ and σ^2 are unknown. Then the information matrix for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is $\hat{\boldsymbol{I}} = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} I_{p \times p} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}$, so that now the effective sample size appears to be n for all parameters.

Note: The group means problem and ANOVA are linear models, so one can have effective sample sizes from $r = 1$ to n for parameters in the linear model.

PBIC: a proposed solution

Preliminary 'nice' reparameterization.

Choose a 'good' transformation to make the Laplace approximation as accurate as possible. In particular, all parameters should lie in $(-\infty, \infty)$. For variances, it is typical to define $\nu = \log \sigma^2$ as the parameter.

By a Taylor's series expansion about the mle $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned} m(\mathbf{x}) &= \int f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int e^{l(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx \int \exp\left[l(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \nabla l(\hat{\boldsymbol{\theta}}) - \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \hat{\boldsymbol{I}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

where ∇ denotes the gradient and $\widehat{\boldsymbol{I}} = (\hat{I}_{jk})$ is the observed information matrix, with (j, k) entry

$$\hat{I}_{jk} = -\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(\mathbf{x} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} .$$

If $\boldsymbol{\theta}$ occurs on the interior of the parameter space, so $\nabla l(\hat{\boldsymbol{\theta}}) = 0$ (if not true, the analysis must proceed as in Haughton (1991,1993)), mild conditions yield

$$m(\mathbf{x}) = e^{l(\hat{\boldsymbol{\theta}})} \int e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^t \hat{\boldsymbol{I}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} (1 + o_n(1)) \,.$$

Note 1. Usually $\pi(\boldsymbol{\theta})$ is also included in the expansion. We will instead choose $\pi(\boldsymbol{\theta})$ to be a 'good' prior for which the integral above is closed form.

Note 2. The term $o_n(1)$ is absent in normal likelihoods, so all expressions will be *exact* in normal scenarios.

If there are any *common* parameters in all models (as in regression, when all models usually have the intercept), integrate them out $d\boldsymbol{\theta}$.

Assuming no common parameters (for convenience) we choose the prior $\pi(\boldsymbol{\theta})$ as follows:

- Let $\boldsymbol{O}$ be orthogonal and $\boldsymbol{D} = \text{diag}(d_i)$ such that $\hat{\boldsymbol{I}}^{-1} = \boldsymbol{O}^t \boldsymbol{D} \boldsymbol{O}$ and make the change of variables $\boldsymbol{\xi} = \boldsymbol{O}\boldsymbol{\theta}$, $\hat{\boldsymbol{\xi}} = \boldsymbol{O}\hat{\boldsymbol{\theta}}$.
- For each ξ_i and following Kass and Wasserman (1995 JASA), let $(b_i)^{-1} = (n_i d_i)^{-1} = \frac{1/d_i}{n_i}$ be the "unit information" for ξ_i, with n_i being the "effective sample size" for ξ_i.
- Instead of using the unit information Cauchy or intrinsic priors, choose the prior (from Berger 1985, generalizing the Strawderman prior)

$$\pi(\xi) = \prod_{i=1}^{p} \pi_i^R(\xi_i)\,, \quad \pi_i^R(\xi_i) = \int_0^1 N\left(\xi_i \,\middle|\, 0, \frac{1}{2\lambda_i}(d_i + b_i) - d_i\right) \frac{1}{2\sqrt{\lambda_i}}\, d\lambda_i\,,$$

which is very close to the unit information Cauchy or intrinsic prior.

Then

$$m(\mathbf{x}) \approx e^{l(\hat{\boldsymbol{\theta}})}(2\pi)^{p/2}|\hat{\boldsymbol{I}}|^{-1/2}\left[\prod_{i=1}^{p}\frac{1}{\sqrt{2\pi(d_i+b_i)}}\frac{\left(1-e^{-\hat{\xi}_i^2/(d_i+b_i)}\right)}{\sqrt{2}\,\hat{\xi}_i^2/(d_i+b_i)}\right].$$

and, we have, as the approximation to $-2\log m(\mathbf{x})$,

$$\text{PBIC} = -2l(\hat{\boldsymbol{\theta}})+\sum_{i=1}^{p}\log(1+n_i)-2\sum_{i=1}^{p}\log\frac{(1-e^{-v_i})}{\sqrt{2}\,v_i}, \quad \text{where } v_i = \frac{\hat{\xi}_i^2}{b_i+d_i}.$$

The error, as an approximation to $-2\log m(\mathbf{x})$, is $o_n(1)$. (Note that it is exact for normal likelihoods.)

If all $n_i = n$, the dominant terms in the expression (as $n \to \infty$) are $-2l(\hat{\boldsymbol{\theta}}) + p\log n$. The third term is negative.

PBIC*: A Modification More Favorable to Complex Models

Concern: Do unit-information Cauchy-type priors centered at zero penalize complex models too much?

- Raftery (1996 Biometrika) proposed unit-information normal priors centered at the mle's for the parameters, but this can be argued to favor complex models too much.
- An attractive compromise is to use the robust priors centered at zero, but with the scales, b_i, chosen to maximize $m(\boldsymbol{x})$. This is the empirical Bayes alternative, popularized in the robust Bayesian literature (see, e.g., Berger, 1994 Test). The b_i that maximizes $m(\boldsymbol{x})$ is

$$\hat{b}_i = \max\{d_i, \frac{\hat{\xi}_i^2}{w} - d_i\}, \text{ with } w \text{ s.t. } e^w = 1 + 2w, \text{ or } w \approx 1.3 \ .$$

- *Problem:* When $\xi_i = 0$, this empirical Bayes choice can result in inconsistency as $n_i \to \infty$.
- *Solution:* prevent b_i from being less than $n_i d_i$, using $\tilde{b}_i = \max\{n_i d_i, \hat{b}_i\}$.

Consistency of PBIC

PBIC and PBIC* are consistent as the effective sample sizes $n_i \to \infty$ with p fixed, since the priors are then essentially fixed priors.

Much harder is consistency as $p \to \infty$, with n_i fixed.

Theorem 1 *For the group means problem with fixed r and known σ^2, consider comparison of $M_0 : \mu_1 = \cdots = \mu_p = 0$ with the full model M_1 :* all μ_i nonzero. *PBIC and PBIC* are consistent under M_0 as $p \to \infty$. Under M_1 and assuming $V \equiv \lim_{p\to\infty} \frac{1}{p} \sum_i^p \mu_i^2$ exists, they are*

$$\text{consistent if } V > \frac{1}{r}\left[\log 2 + \log(1+r) + 1\right];$$

$$\text{inconsistent if } V < \frac{1}{r}\left[\log 2 + \log(1+r) - 1\right].$$

Note 1: Inconsistency results only when M_1 is close to M_0. (Mukhopadhyay, Ghosh, and Berger, 2005 SPL, showed a multivariate Cauchy prior is always consistent.)

Note 2: The theorem applies to any two models for which the difference in dimensions goes to ∞.

A small comparative simulation:

Berger, Ghosh and Mukhopadhayay (2003) computed Laplace approximations to the marginal density with a multivariate Cauchy prior; they called GBIC the resulting $\log m(\boldsymbol{x}\ )$ and showed that it was consistent.

This original GBIC, which inspired our PBIC's, does not have closed form expression. Berger et al. (2003) give an approximation valid when $\sum \bar{x}_i^2 > r^{-1} + \epsilon$ for some $\epsilon > 0$ as $p \to \infty$.

We next compare our PBIC's and this approximated, closed-form expression GBIC (note, however that the condition is likely to be violated when sampling from the null model, or whenever it is likely to get many x_i^2 near 0, and then the simplified expression used would not be a good approximation to Berger et al. (2003) proposal.)

We generate 500 sets of observations with several values for p and r, under the following conditions:

a) All observations $X_{ir} \sim N(0,1)$ (null model);

b) the p group means (the μ_i) were generated from a N(2,1), (and then the 500 sets of X_{ir} from the $N(\mu_i, 1)$);

c) similar to the previous one, but the μ_i generated from an exponential with mean 2

d) one μ_i is set to 10, and the rest to 0 (note neither the null nor the alternative are true)

The following table gives the mean and standard deviation of $\triangle$GBIC (denoted $\triangle_O$), our $\triangle$PBIC proposal (denoted $\triangle_N$), and the robust modification (denoted $\triangle_R$).

		$\boldsymbol{\mu} = \mathbf{0}$		$\mu_i \sim N(2,1)$		$\mu_i \sim Ex(\mu = 2)$		$\mu_1 = 10, \mu_i = 0$	
p, r	$\triangle$PBIC	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$p = 2$	$\triangle_o$	0.383	1.38	17.89	9.17	7.8	6.2	180	27
$r = 2$	$\triangle_N$	-2.155	1.42	16.54	8.64	7.4	6.1	187	28
(p=5 last)	$\triangle_R$	-2.117	1.54	19.58	9.9	8.9	7.07	194	28
$p = 15$	$\triangle_o$	-1.64	1.65	92.96	20.31	257	33.5	157	26
$r = 2$	$\triangle_N$	-16.47	4.18	87.66	19.65	258	33.4	175	27
	$\triangle_R$	-16.20	4.58	103.56	22.28	281	34.9	183	28
$p = 200$	$\triangle_o$							56	17
$r = 2$	$\triangle_N$							-27	31
	$\triangle_R$							-17	32

Table 1: For the group means problem, the means and standard deviations of various $\Delta\text{PBIC} \equiv \text{PBIC}_{\boldsymbol{\mu}\,=\mathbf{0}} - \text{PBIC}_{\boldsymbol{\mu}\,\neq\mathbf{0}}$ for sets of 500 replications, under different assumptions about the group means.
Δ_0 :Cauchy, Δ_N :new PBIC, Δ_R :robust PBIC.

IV. The Effective Sample Size (in Linear Models)

with Susie Bayarri and Luis Pericchi

Recall the question: what is the effective sample size n for a parameter?

- Is n the number of vector observations or the number of real observations?
- Different θ_i can have different effective sample sizes
- Some observations can be more informative than others (as in mixture contexts, models with mixed continuous and discrete observations, ...)

A Solution for Linear Models

Assume that:

- all linear models under consideration are of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Gamma}), \quad \boldsymbol{\Gamma} \ \text{ known}\,,$$

  with dimensions $\boldsymbol{Y}_{[n\times 1]}$, $\boldsymbol{\beta}_{[p\times 1]}$

- $\boldsymbol{\beta}$ is the original parameter of interest to the investigator.

- no component of $\boldsymbol{\beta}$ can be considered 'common' to all models.

  In a preliminary step, 'common' parameters (appearing in all models) are orthogonalized to $\boldsymbol{\beta}$, and do not require assessment of effective sample size.

TESS defines the effective sample size for any *scalar* linear transformation $\xi = \boldsymbol{v}\,\boldsymbol{\beta}$ ($\boldsymbol{v}$ is $[1 \times p]$) of $\boldsymbol{\beta}$ to be

$$n^e = \frac{|\boldsymbol{v}|^2}{\boldsymbol{v}\,\boldsymbol{C}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{C}\,\boldsymbol{v}^t}$$

- $\boldsymbol{C}_{[p\times p]}$ is diagonal with entries $c_{ii} = \max_j\{|X_{ji}|/\sigma_j\}$
- $\boldsymbol{\Gamma} = \boldsymbol{\sigma}\boldsymbol{R}\boldsymbol{\sigma}$, with $\boldsymbol{\sigma} = \text{diag}\{\sigma_1, \ldots, \sigma_p\}$, $\boldsymbol{R}$ is correlation matrix

The **"unit information" prior scale** for ξ is then $b = d\,n^e$, where $d = \boldsymbol{v}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{v}^t$ is the variance of $\widehat{\xi}$

Group means example

Assume $Y_{ij} = \mu_i + \varepsilon_{ij}$ for $i = 1, .., p$ groups, $j = 1, .., r_i$ replicates in ith group, $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. Here

$$\boldsymbol{n}^e = \begin{pmatrix} r_1 & & \\ & \ddots & \\ & & r_p \end{pmatrix} \quad \text{and TESS for } \mu_i \text{ is } n_i^e = r_i$$

- r_i could be 1, which can be seen to be the lower bound on TESS for linear models when $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}$, which is intuitively reasonable
- the prior scale for μ_i is $b_i = \sigma^2$.

Simple Linear Regression

$$\boldsymbol{Y} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \qquad \text{where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}) \ .$$

The effective sample size and prior scale for β are, respectively,

$$n^e = \frac{\sum_{i=1}^n X_i^2}{(\max_i |X_i|)^2} \qquad \text{and} \quad b = \sigma^2 \ . \tag{2}$$

Let's consider some particular cases

Case 1: $\boldsymbol{X} = (X_1, \delta, \ldots, \delta)^t$, with δ very small. Here

$n^e = 1 + \dfrac{(n-1)\delta}{X_1^2} \approx 1$, for small δ, an intuitive result

Case 2: $\boldsymbol{X} = (1, \ldots, 1)^t$. Here $n^e = n$, which agrees with intuition.

Case 3: $\boldsymbol{X} = (X_1, \ldots, X_n)^t$, with $X_i \overset{i.i.d.}{\sim} N(k, 1)$. For large n

- if $k = 0$, $n^e \approx n/(2 \log n - 3)$
- at the other extreme, if k large compared to $\log n$, $n^e \approx n$

Case 4: $X_i = 1/\sqrt{i},\ i = 1 \ldots, n$. This is Findley's counter example to consistency of BIC. Here $n^e = \sum_{i=1}^{n} 1/i \approx \log(n+1)$ which behaves like $\log n$ and the inconsistency observed by Findley disappears.

Orthogonal and Related Designs

Assume that $\boldsymbol{X}$ has orthogonal columns with entries $\pm a_i \neq 0$, and that $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}$.

- Here TESS gives $n_i^e = n$ for each β_i
- When $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}$ and any other design matrix $\boldsymbol{X}$ is used, the effective sample sizes will be less than n

NOTE: This along with the result for the group means example, establishes that when $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}$, TESS will always be between 1 and n, with both limits attainable

Heteroscedastic independent observations

Assume $Y_i = \mu + \varepsilon_i$, ε_i independent, $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, n$. Here the effective sample size and prior scale for μ are

$$n^e = \frac{\sum_{i=1}^n 1/\sigma_i^2}{\max_i\{1/\sigma_i^2\}}, \qquad b = \min_i\{\sigma_i^2\}\,.$$

Particular Case: observations with little information. Suppose that, for $i = 1, \ldots, n_1$, we have $Y_i \sim N(\mu, \sigma_1^2)$, whereas for the remaining $n_2 = n - n_1$ observations, $Y_i \sim N(\mu, \sigma_2^2)$, where σ_2^2 is much larger than σ_1^2, so that intuitively only the first n_1 observations count. Then, unless n_2 is large,

$$n^e = \frac{n_1/\sigma_1^2 + n_2/\sigma_2^2}{1/\sigma_1^2} = n_1 + n_2\frac{\sigma_1^2}{\sigma_2^2} \approx n_1\,.$$

Correlated observations

Let $Y_i = \mu + \varepsilon_i,\ i = 1, \ldots, n$, but where the ε_i are *not* independent, with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$ with $\boldsymbol{\Gamma}$ non diagonal. Here

$$n^e = \frac{\mathbf{1}^t \boldsymbol{\Gamma}^{-1} \mathbf{1}}{\max_i \{\frac{1}{\sigma_i^2}\}}\ , \qquad \text{and} \quad b = \min_i \sigma_i^2$$

Particular case 1. Consider $\boldsymbol{\Gamma} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & & \cdots & 1 \end{pmatrix}$.

Then $n^e = \dfrac{n}{1 + (n-1)\rho}$ and $b = \sigma^2$

Note that

$$n^e \longrightarrow \begin{cases} 1, & \text{as } \rho \to 1 \\ n, & \text{as } \rho \to 0 \\ \infty, & \text{as } \rho \to -1/(n-1) \end{cases}, \qquad \text{and} \quad b = \sigma^2 \,,$$

all intuitively reasonable results (in the last case, we know $\mu = \bar{x}$)

Note: effective sample size can be larger than n in the presence of negative correlation, but prior scales remains constant (σ^2) $\rightsquigarrow$ Bayesian analysis will automatically adjust for correlations.

Particular case 2. Consider now a general $\boldsymbol{R}$ for $n = 2$, and assume $\sigma_1^2 < \sigma_2^2$, so that

$$\boldsymbol{Y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mu + \boldsymbol{\varepsilon}\,, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Gamma}), \text{ with } \quad \boldsymbol{\Gamma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

$$\text{Here} \quad n^e = 1 + \frac{(\frac{\sigma_1}{\sigma_2} - \rho)^2}{1 - \rho^2} \qquad \text{and} \quad b = \sigma_1^2$$

- minimum value for TESS is $n^e = 1$, when $\rho = \sigma_1/\sigma_2$
- $n^e \to \infty$ as $|\rho| \to 1$

  recall: when $\sigma_1 = \sigma_2$, $n^e \to 1$ as $\rho \to 1$

  Here, when $\rho = \pm 1$ we know μ *perfectly*, corresponding to 'infinite sample information'.

Argument for TESS

The precision d^{-1} of $\widehat{\xi}$ is roughly the effective sample size, except that it has three type of scale factors in it that need to be removed:

(i) the scales arising from the σ_j in $\mathbf{\Gamma}$,

(ii) the scales arising from possible arbitrariness in the scaling of the columns of $\boldsymbol{X}$.

(iii) the scales arising from possible arbitrariness in the definition of ξ

Step 1. Remove σ scales $\rightsquigarrow$ divide original observation Y_i by its standard deviation σ_i, $i = 1, \ldots, n$. That is:

The original model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{\Gamma})$,

transforms to: $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}, \quad \widetilde{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \boldsymbol{R})$, where

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{\sigma}^{-1}\,\boldsymbol{Y}, \text{ and } \widetilde{\boldsymbol{X}} = \boldsymbol{\sigma}^{-1}\,\boldsymbol{X}$$

.

Step 2. Remove X scales $\rightsquigarrow$ divide columns of $\widetilde{\boldsymbol{X}}$ by their maximum (other scalings are possible: to be pursued)

After Steps 1 and 2, we have transformed the original model to the following "scale free" model

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{X}^* \boldsymbol{\beta}^* + \widetilde{\boldsymbol{\varepsilon}}, \quad \widetilde{\boldsymbol{\varepsilon}} \sim N(\boldsymbol{0}, \boldsymbol{R}), \quad \text{where}$$

- $\boldsymbol{X}^* = \boldsymbol{X}\ \boldsymbol{C}^{-1}$

  $\boldsymbol{C}$ is diagonal with $c_{ii} = \max_j \left\{ \dfrac{|X_{ji}|}{\sigma_j} \right\} \quad i = 1, \ldots, p.$

- $\boldsymbol{\beta}^* = \boldsymbol{C}\boldsymbol{\beta} = \begin{pmatrix} c_{11}\beta_1 \\ \vdots \\ c_{pp}\beta_p \end{pmatrix}$

  is like a 'scale free' version of the original parameter $\boldsymbol{\beta}$.

Step 3. Compute TESS for original parameters. We *define* the effective sample size matrix for the original parameter $\boldsymbol{\beta}$ as the precision of $\widehat{\boldsymbol{\beta}^*}$ the MLE in the scale free formulation, giving

$$\boldsymbol{n}_o^e = \boldsymbol{C}^{-1}(\boldsymbol{X}^t\boldsymbol{\Gamma}^{-1}\boldsymbol{X})\boldsymbol{C}^{-1}\,.$$

Step 4. TESS for the parameters of interest. We *define* TESS for any scalar transformation $\boldsymbol{v}\,\boldsymbol{\beta}$ by

$$[\widetilde{\boldsymbol{v}}\,(\boldsymbol{n}_o^e)^{-1}\,\widetilde{\boldsymbol{v}}^t]^{-1}\ \text{, where}\quad \widetilde{\boldsymbol{v}} = \boldsymbol{v}/|\boldsymbol{v}|$$

Note: we have removed arbitrariness in the scale of $\boldsymbol{v}$ so TESS for ξ is the same as TESS for $k\,\xi$

Current Status

- We are happy with PBIC, although both PBIC and PBIC* should typically be considered.
  - Note that these are *exact* expressions if the likelihoods are normal and can, hence, even be used as $p \to \infty$.
- We are happy with TESS in linear models, in that
  - it has desirable scale-free properties;
  - it produces pleasant surprises;
  - we have no examples of it failing to provide a sensible answer.
- We are not happy with the following possible definition of effective sample size in non-linear models.

Effective Sample Size in Nonlinear Models

A possible general definition for the 'effective sample size' follows from considering the information associated with observation $\boldsymbol{x}_i$ arising from the single-observation expected information matrix $\boldsymbol{I}_i^* = \boldsymbol{O}'(I_{i,jk}^*)\boldsymbol{O}$, where

$$I_{i,jk}^* = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f_i(\boldsymbol{x}_i \mid \boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Since $I_{jj}^* = \sum_{i=1}^n I_{i,jj}^*$ is the expected information about ξ_j, a reasonable way to define n_j is

- define information weights $w_{ij} = I_{i,jj}^* / \sum_{k=1}^n I_{k,jj}^*$;
- define the effective sample size for ξ_j as

$$n_j = \frac{I_{jj}^*}{\sum_{i=1}^n w_{ij} I_{i,jj}^*} = \frac{\left(I_{jj}^*\right)^2}{\sum_{i=1}^n \left(I_{i,jj}^*\right)^2}.$$

Intuitively, $\sum w_{ij} I_{i,jj}^*$ is a weighted measure of the information 'per observation', and dividing the total information about ξ_j by this information per case seems plausible as an effective sample size.

References

[1] Dudley, R. and Haughton, D. (1997). Information Criteria for Multiple Data Sets and Restricted Parameters. *Statistica Sinica*, **7**, 265–284.

[2] Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *JASA* **90**, 773–795.

[3] Kass, R.E., and Vaiyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with applications to testing equality of two binomial proportions. *JRRS B* **54**, 129–144.

[4] Kass, R.E., and Wasserman, L. (1995). A Reference test for neste hypothesis and its relationship to the Schwarz criterion. *JASA* **90**, 928–934.

[5] Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

[6] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

[7] Tirney, L., and Kadane, J.B. (1986). Accurate approximations of posterior moments and marginal densities. *JASA* **81**, 82–86.