An Extended BIC for Model Selection

at
the JSM meeting 2007 - Salt Lake City

Surajit Ray
Boston University (Dept of Mathematics and Statistics)

Joint work with
James Berger, Duke University; Susie Bayarri, University of Valencia;
Woncheol Jang, University of Georgia; Luis R. Pericchi, University of Puerto Rico
**Motivation and Outline**

- Initiated from a SAMSI Social Sciences working group. “getting the model right” in structural equation modeling.
- Introduced in the **Wald Lecture by Jim Berger**.

**Specific characteristics**
- They have ‘independent cases’ (individuals).
- Often ‘Multilevel’ or ‘random effects’ models ($p$ grows as $n$ grows).
- Regularity Conditions may not be satisfied.

**Problems with BIC.**
- Can a generalization of BIC work?
  - Goal: Use only Standard software output and close-form expressions.
- Present a generalization of BIC (‘extended BIC’ or $EBIC$) and some examples in linear regression scenario.
- Work in progress (comments are welcome!)
Bayesian Information Criterion: BIC

\[ BIC = 2l(\hat{\theta} | x) - p \log(n), \]

- \( l(\hat{\theta} | x) \) the log-likelihood of \( \hat{\theta} \) given \( x \)
- \( p \) the number of estimated parameters.
- \( n \) the cardinality of \( x \).

Swartz’s result: As \( n \to \infty \) (with \( p \) fixed) this is an approximation (up to a constant) to twice the Bayesian log likelihood for the model,

\[ m(x) = \int f(x | \theta) \pi(\theta) d\theta, \text{ so that} \]

\[ m(x) = c_{\pi} e^{BIC/2} (1 + o(1)). \]

BIC approaches to BF’s further ignore the \( c_{\pi} \)’s check this
First Ambiguity: What is \( n \)?

Consider time series \( y_{it} \),

\[
y_{it} = \beta y_{i(t-1)} + \varepsilon_{it}, \quad i = 1, \ldots, n^*, \quad t = 1, \ldots, T,
\]

\[
\varepsilon_{it} \sim N(0, \sigma^2).
\]

When the goal is to estimate the mean level \( \bar{y} \), consider two extreme cases:

1. \( \beta = 0 \); then the \( y_{it} \) are i.i.d., and the estimated parameter \( \bar{y} \) has sample size
   \[
n = n^* \times T
\]

2. \( \beta = 1 \); then clearly the effective sample size for \( \bar{y} \) equals \( n = n^* \)

Second Ambiguity: Can \( n \) be different for different parameters?

Consider data from \( p \) groups

\[
y_{ip} = \mu_p + \varepsilon_{ip}, \ i = 1 \ldots n^*, \ p = 1 \ldots r
\]

\( \varepsilon_{ip} \sim N(0, \sigma^2) \)

\( \mu_p \) and \( \sigma^2 \) both unknown.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>( y_{11} )</th>
<th>( y_{21} )</th>
<th>\ldots</th>
<th>( y_{n^*1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>( y_{12} )</td>
<td>( y_{22} )</td>
<td>\ldots</td>
<td>( y_{n^*2} )</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Group p</td>
<td>( y_{1p} )</td>
<td>( y_{2p} )</td>
<td>\ldots</td>
<td>( y_{n^*p} )</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Group r</td>
<td>( y_{1r} )</td>
<td>( y_{2r} )</td>
<td>\ldots</td>
<td>( y_{n^*r} )</td>
</tr>
</tbody>
</table>

\( \mu \) and \( \sigma^2 \) both unknown.
Second Ambiguity: Can $n$ be different for different parameters?

Consider data from $p$ groups

$$y_{ip} = \mu_p + \varepsilon_{ip}, \; i = 1 \ldots n^*, \; p = 1 \ldots r$$

$$\varepsilon_{ip} \sim N(0, \sigma^2)$$

$\mu_p$ and $\sigma^2$ both unknown.

Different parameters can have different sample sizes:

- the sample size for estimating each $\mu_p$ is $n^*$
- the sample size for estimating $\sigma^2$ is $n^* \times r$
Second Ambiguity: Can $n$ be different for different parameters?

Consider data from $p$ groups

$$y_{ip} = \mu_p + \varepsilon_{ip}, \quad i = 1 \ldots n^*, \quad p = 1 \ldots r$$

$$\varepsilon_{ip} \sim N(0, \sigma^2)$$

$\mu_p$ and $\sigma^2$ both unknown.

Computation for this example yields

$$\hat{I} = \begin{pmatrix} \frac{n^*}{\hat{\sigma}^2} I_r \times r & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix},$$

where $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{p} \sum_{l=1}^{r} (X_{il} - \bar{X}_i)^2$. 
Third Ambiguity: What is the number of parameters

Example: Random effects group means
In the previous Group Means example, with \( p \) groups and \( r \) observations \( X_{il} \) per group, with \( X_{il} \sim N(\mu_i, \sigma^2) \) for \( l = 1, \ldots, r \), consider the multilevel version in which

\[
\mu_i \sim N(\xi, \tau^2),
\]

with \( \xi \) and \( \tau^2 \) being unknown.

What is the number of parameters? (see also Pauler (1998))

(1) If \( \tau^2 = 0 \), there is only two parameters: \( \xi \) and \( \sigma^2 \).

(2) If \( \tau^2 \) is huge, is the number of parameters \( p + 3 \) ?

\( \bullet \) the means along with \( \xi, \tau^2 \) and \( \sigma^2 \)

Continued ...
Example: Random effects group means

(3) But, if one integrates out $\mu = (\mu_1, \ldots, \mu_p)$, then

$$f(x \mid \sigma^2, \xi, \tau^2) = \int f(x \mid \mu, \xi, \sigma^2) \pi(\mu \mid \xi, \tau^2) d\mu$$

$$\propto \frac{1}{\sigma^{-p(r-1)}} \exp \left( \frac{\hat{\sigma}^2}{2\sigma^2} \right) \prod_{i=1}^{p} \exp \left( -\frac{(\bar{x}_i - \xi)^2}{2\left( \frac{\sigma^2}{\tau} + \tau^2 \right)} \right),$$

so $p = 3$ if one can work directly with $f(x \mid \sigma^2, \xi, \tau^2)$.

Note: This seems to be common practice in Social Sciences: latent variables are integrated out, so remaining parameters are the only unknowns when applying BIC

Note: In this example the effective sample sizes should be $\approx pr$ for $\sigma^2$, $\approx p$ for $\xi$ and $\tau^2$, and $\approx r$ for the $\mu_i$’s.
Example: Common mean, differing variances:

(Cox mixtures)
Suppose each independent observation $X_i, i = 1, \ldots, n,$ has probability $1/2$ of arising from the $N(\theta, 1)$, and probability $1/2$ of arising from the $N(\theta, 1000)$.

Clearly half of the observations are worthless, so the ‘effective sample size’ is roughly $n/2$. 

Problem with $n$ and $p$

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

- Problems with $p$.
  - What is $p$ with random effects, latent variables, mixture model... etc. ?
  - Often $p$ grows with $n$
EBIC: a proposed solution

Based on a modified Laplace approximation to $m(x)$ for good priors:

- The good priors can deal with growing $p$
- The Laplace approximation
  - retains the constant $c_\pi$ in the expansion
  - often good even for moderate $n$

- Implicitly uses ‘effective sample size’ (discussion later)
- Computable using standard software using mle’s and observed information matrices.
EBIC: a proposed solution

- Based on a modified Laplace approximation to \( m(x) \) for good priors:
  - The *good* priors can deal with growing \( p \)
  - The Laplace approximation
    - retains the constant \( c_\pi \) in the expansion
    - often good even for moderate \( n \)

- Implicitly uses ‘effective sample size’ (discussion later)

- Computable using standard software using mle’s and observed information matrices.
EBIC: a proposed solution

- Based on a modified Laplace approximation to $m(x)$ for good priors:
  - The good priors can deal with growing $p$
  - The Laplace approximation
    - retains the constant $c_\pi$ in the expansion
    - often good even for moderate $n$
  - Implicitly uses ‘effective sample size’ (discussion later)
- Computable using standard software using mle’s and observed information matrices.
Laplace approximation

1. **Preliminary: ‘nice’ reparameterization.**

2. **Taylor expansion**
   By a Taylor’s series expansion of $e^{l(\theta)}$ about the mle $\hat{\theta}$,
   
   $$m(x) = \int f(x | \theta) \pi(\theta) d\theta = \int e^{l(\theta)} \pi(\theta) d\theta$$
   
   $$\approx \int \exp \left[ l(\hat{\theta}) + (\theta - \hat{\theta})^t \nabla l(\hat{\theta}) - \frac{1}{2} (\theta - \hat{\theta})^t \hat{I} (\theta - \hat{\theta}) \right] \pi(\theta) d\theta$$

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

3. **Approximation to marginal $m(x)$**
   If $\hat{\theta}$ occurs on the interior of the parameter space, so $\nabla l(\hat{\theta}) = 0$ (if not true, the analysis must proceed as in Haughton 1991, 1993), mild conditions yield
   
   $$m(x) = e^{l(\hat{\theta})} \int e^{-\frac{1}{2} (\theta - \hat{\theta})^t \hat{I} (\theta - \hat{\theta})} \pi(\theta) d\theta (1 + o_n(1)) .$$
Choosing a good prior

- Integrate out common parameters
- Orthogonalize the parameters.

Introduction and Motivation

EBIC

The prior

- Choosing a good prior
- Univariate priors
- Berger’s robust priors
- Proposal for EBIC
- Comparison of EBIC and BIC

EBIC*

Effective sample size

Simulation
Choosing a good prior

- Integrate out common parameters
- Orthogonalize the parameters.
  - Let \( \mathbf{O} \) be orthogonal and \( \mathbf{D} = \text{diag}(d_i), i = 1, \ldots, p_1 \) such that \( \Sigma = \mathbf{O}^t \mathbf{D} \mathbf{O} \)
Choosing a good prior

- Integrate out common parameters
- Orthogonalize the parameters.
  - Let $\mathbf{O}$ be orthogonal and $\mathbf{D} = \text{diag}(d_i), i = 1, \ldots, p_1$ such that $\Sigma = \mathbf{O}^t \mathbf{D} \mathbf{O}$
  - Make the change of variables $\xi = \mathbf{O} \theta(1)$, and let $\hat{\xi} = \mathbf{O} \hat{\theta}(1)$
Choosing a good prior

Integrate out common parameters

Orthogonalize the parameters.

Let $O$ be orthogonal and $D = \text{diag}(d_i), i = 1, \ldots p_1$ such that $\Sigma = O^t D O$

Make the change of variables $\xi = O\theta^{(1)}$, and let $\hat{\xi} = O\hat{\theta}^{(1)}$

If there is no $\theta^{(2)}$ to be integrated out, $\Sigma = \hat{I}^{-1} = O^t D O$
Choosing a good prior

Integrate out common parameters

Orthogonalize the parameters.

Let \( \mathbf{O} \) be orthogonal and \( \mathbf{D} = \text{diag}(d_i), i = 1, \ldots, p_1 \) such that 
\[
\Sigma = \mathbf{O}^t \mathbf{D} \mathbf{O}
\]

Make the change of variables \( \xi = \mathbf{O} \theta (1) \), and let \( \tilde{\xi} = \mathbf{O} \hat{\theta} (1) \)

If there is no \( \theta (2) \) to be integrated out, \( \Sigma = \mathbf{I}^{-1} = \mathbf{O}^t \mathbf{D} \mathbf{O} \)

We now use a prior that is independent in the \( \xi_i \) i.e., \( \pi (\xi) = \prod_{i=1}^{p_1} \pi_i (\xi_i) \)
Choosing a good prior

- Integrate out common parameters
- Orthogonalize the parameters.

Let $O$ be orthogonal and $D = \text{diag}(d_i), i = 1, \ldots, p_1$ such that $\Sigma = O^t DO$.

Make the change of variables $\xi = O\theta_{(1)}$, and let $\hat{\xi} = O\hat{\theta}_{(1)}$.

If there is no $\theta_{(2)}$ to be integrated out, $\Sigma = \hat{I}^{-1} = O^t DO$.

We now use a prior that is independent in the $\xi_i$ i.e, $\pi(\xi) = \prod_{i=1}^{p_1} \pi_i(\xi_i)$.

The approximation to $m(x)$ is then

$$m(x) \approx e^{l(\hat{\theta})}(2\pi)^{p/2} |\hat{I}|^{-1/2} \left[ \prod_{i=1}^{p_1} \int \frac{1}{\sqrt{2\pi d_i}} e^{-\frac{(\xi_i - \hat{\xi}_i)^2}{2d_i}} \pi_i(\xi_i) d\xi_i \right].$$

where we recall that $d_i$ are the diagonal elements of $D$ from $\Sigma = O^t DO$, that is, $\text{Var}(\xi_i)$. They will appear often in what follows.
Univariate priors

For the priors $\pi_i(\xi_i)$, there are several possibilities:

1. Jeffreys recommended the Cauchy $(0, b_i)$ density

$$
\pi_i^C(\xi_i) = \int_0^\infty N(\xi_i \mid 0, \frac{1}{\lambda_i} b_i) \ Ga(\lambda_i \mid \frac{1}{2}, \frac{1}{2}) d\lambda_i
$$

where $(b_i)^{-1} = \left(\frac{d_i}{n_i}\right)^{-1}$ is the unit information for $\xi_i$ with $n_i$ being the

effecteive sample size for $\xi_i$ (Kass & Wasserman, 1995).

Note: For i.i.d. scenarios, $d_i$ is like $\sigma^2_i/n_i$ so $b_i$ is like the

variance of one observation.

2. Use other sensible (objective) testing priors, like the intrinsic prior.

3. **Goal is to get closed form expression:** We use priors proposed in Berger (1985) which is very close to both the Cauchy and the intrinsic priors and

produces close form expressions for $m(x)$ for normal likelihoods.
Berger’s robust priors

For if \( b_i \geq d_i \), the prior is given by

\[
\pi_i^R(\xi_i) = \int_0^1 N \left( \xi_i \mid 0, \frac{1}{2\lambda_i} (d_i + b_i) - d_i \right) \frac{1}{2\sqrt{\lambda_i}} d\lambda_i ,
\]
Berger’s robust priors

For if $b_i \geq d_i$, the prior is given by

$$\pi_i^R(\xi_i) = \int_0^1 N \left( \xi_i \mid 0, \frac{1}{2\lambda_i} (d_i + b_i) - d_i \right) \frac{1}{2\sqrt{\lambda_i}} d\lambda_i,$$

- Introduced by Berger in the context of Robust Analyses
- No closed form expression.
- Not widely used in model selection
- Behaves like Cauchy priors with parameters $b_i$. 

**EBIC**

- Choosing a good prior
- Univariate priors
- Berger’s robust priors
- Proposal for EBIC
- Comparison of EBIC and BIC

**EBIC**

- Effective sample size
- Simulation
Berger’s robust priors

For if $b_i \geq d_i$, the prior is given by

$$
\pi^R_i(\xi_i) = \int_0^1 N \left( \xi_i \middle| 0, \frac{1}{2\lambda_i} (d_i + b_i) - d_i \right) \frac{1}{2\sqrt{\lambda_i}} d\lambda_i ,
$$

- Introduced by Berger in the context of Robust Analyses
- No closed form expression.
- Not widely used in model selection
- Behaves like Cauchy priors with parameters $b_i$.
- Gives closed form expression for approximation to $m(x)$

\[
\begin{align*}
    m(x) &\approx e^{i(\hat{\Theta})} (2\pi)^{p/2} |\hat{I}|^{-1/2} \left[ \prod_{i=1}^{p_1} \frac{1}{\sqrt{2\pi(d_i + b_i)}} \left( 1 - e^{-\xi_i^2/(d_i + b_i)} \right) \right ] \\
\end{align*}
\]

where $(d_i)^{-1}$ is global information about $\xi_i$, and $(b_i)^{-1}$ is unit information.
Finally, we have, as the approximation to $2 \log m(\mathbf{x})$,

$$\text{EBIC} \equiv 2l(\hat{\theta}) + (p - p_1) \log(2\pi) - \log |\hat{I}_{22}| - \sum_{i=1}^{p_1} \log(1 + n_i)$$

$$+ 2 \sum_{i=1}^{p_1} \log \left(\frac{1 - e^{-v_i}}{\sqrt{2}v_i}\right), \quad \text{where} \quad v_i = \frac{\hat{\xi}_i^2}{b_i + d_i}.$$ 

(error as approximation to $2 \log m(\mathbf{x})$ is $o_n(1)$; exact for Normals)

If no parameters are integrated out (so that $p_1 = p$), then

$$\text{EBIC} = 2l(\hat{\theta}) - \sum_{i=1}^{p} \log(1 + n_i) + 2 \sum_{i=1}^{p} \log \left(\frac{1 - e^{-v_i}}{\sqrt{2}v_i}\right).$$

If all $n_i = n$, the dominant terms in the expression (as $n \to \infty$) are $2l(\hat{\theta}) - p \log n$. The third term (the ‘constant’ ignored by Swartz) is negative.
Comparison of EBIC and BIC

Compare this EBIC and BIC:

\[
\text{EBIC} = 2l(\hat{\theta}) - \sum_{i=1}^{p} \log(1 + n_i) - k_{\pi}
\]

\[
\text{BIC} = 2l(\hat{\theta}) - p \log n
\]

where \(k_{\pi} > 0\) is the ‘ignored’ constant from \(\pi\)

It is easy to intuitively notice that ‘BIC makes two mistakes’:

- Penalizes too much with \(p \log n\). Note that \(n_i \leq n\)
- Penalizes too little with the prior (the third term, a new penalization, is absent)

The second ‘mistake’ actually helps the first one (in this sense, BIC is ‘lucky’), but often the ‘first mistake’ completely overwhelms this little help.
Priors centered at zero penalize complex models too much?

Priors centered around MLE (Raftery, 1996) favor complex models too much?

An attractive compromise: Cauchy-type priors centered at zero, but with the scales, $b_i$, chosen so as to maximize the marginal likelihood of the model.

Empirical Bayes alternative, popularized in the robust Bayesian literature (Berger, 1994)
EBIC*: Modification More Favorable to Complex Models

The $b_i$ that maximizes $m(x)$ can easily be seen to be

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

- Problem: when $\xi_i = 0$, this empirical Bayes choice can result in inconsistency as $n_i \to \infty$.
- Solution: prevent $\hat{b}_i$ from being less than $n_i d_i$. This results, after an accurate approximation (for fixed $p$) in

$$\text{EBIC}^* \approx 2l(\hat{\theta}) - \sum_{i=1}^{p} \log(1 + n_i) - \sum_{i=1}^{p} \log(3\nu_i + 2 \max\{\nu_i, 1\}).$$

again a very simple expression
Effective Sample size

Calculate \( \mathbf{I} = I_{jk} \), the *expected* information matrix, at the MLE,

\[
I_{jk} = -E_{X|\theta} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(x | \theta) \bigg|_{\theta=\hat{\theta}}.
\]
Calculate $I = I_{jk}$, the \textit{expected} information matrix, at the MLE,

$$I_{jk} = -E^X |_{\theta} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(x | \theta) = I_{jk}.\tag{1}$$

For each case $x_i$, compute $I_i = (I_{i,jk})$, having $(j,k)$ entry

$$I_{i,jk} = -E^X_i |_{\theta} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_i(x_i | \theta) = I_{i,jk}.\tag{2}$$
Effective Sample size

- Calculate $I = I_{jk}$, the *expected* information matrix, at the MLE,

$$I_{jk} = -E X \left| \theta \right. \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f (x \mid \theta) \left| \theta = \hat{\theta} \right..$$

- For each case $x_i$, compute $I_i = (I_{i,jk})$, having $(j, k)$ entry

$$I_{i,jk} = -E X_i \left| \theta \right. \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_i (x_i \mid \theta) \left| \theta = \hat{\theta} \right..$$

- Define $I_{jj} = \sum_{i=1}^{n} I_{i,jj}$ Define $n_j$ as follows:
  - Define information weights $w_{ij} = I_{i,jj} / \sum_{k=1}^{n} I_{k,jj}$
  - Define the effective sample size for $\theta_j$ as

$$n_j = \frac{I_{jj}}{\sum_{i=1}^{n} w_{ij} I_{i,jj}} = \frac{(I_{jj})^2}{\sum_{i=1}^{n} (I_{i,jj})^2}.$$
Effective Sample size: Continued

Intuitively, $\sum w_{ij} I_{i,j}$ is a weighted measure of the information ‘per observation’, and dividing the total information about $\theta_j$ by this information per case seems plausible as an effective sample size.

This is just a tentative proposal

Works for most of the examples we discussed

Research in progress
A small linear regression simulation:

\[ Y_{n \times 1} = X_{n \times 8} \beta_{8 \times 1} + \epsilon_{n \times 1} \quad \epsilon \sim N(0, \sigma^2), \]
\[ \beta = (3, 1.5, 0, 0, 2, 0, 0, 0), \]
\[ \text{and design Matrix } X \sim N(0, \Sigma_x), \quad \Sigma_x = \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix} \]
Linear regression Example

Percentage of true models selected under different situations
Linear regression Example

Average number of 0’s that were determined to be 0. (True Model)
**Linear regression Example**

<table>
<thead>
<tr>
<th>Method</th>
<th>n=30</th>
<th>σ=3</th>
<th>60</th>
<th>σ=3</th>
<th>200</th>
<th>σ=3</th>
<th>500</th>
<th>σ=3</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>abf2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>aic</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bic</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Gbic</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Gbic*</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>hbf</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Average number of non-zeros that were determined to be zero (small is good).
Conclusions and future work

**BIC can be improved** by
- keeping rather than ignoring the prior
- using a sensible, objective prior producing close-form expressions
- carefully acknowledging “effective sample size” for each parameter to guide choice of the scale of the priors

**Work in progress**
- A satisfactory definition of ‘effective sample size’ (if possible). Current proposal works in most cases.
- Extension to the non-independent case
References


