

Extensions and generalizations of BIC

**Susie Bayarri, Jim Berger, Woncheol Jang, Luis Pericchi,
Surajit Ray, Raul Rueda, Ingmar Visser**

(U. Valencia, SAMSI, and others)

*Sixth Workshop on Objective Bayesian Methodology
Roma, Italia, June 9-12 2007*

Motivation and Outline

- Arose from a SAMSI Social Sciences working group. They wanted to “get the model right” in structural equation modeling. Some specific characteristics
 - They have ‘independent cases’ (individuals)
 - Often ‘Multilevel’ or ‘random effects’ models (p grows as n grows)
 - Standard software and close-form expressions absolutely required
- Expose problems with BIC
- Present a generalization of BIC (‘extended BIC’ or *EBIC*) and some results about its consistency as $n \rightarrow \infty$ and as $p \rightarrow \infty$
- Work still very much in progress (comments are welcome!)

The Original BIC

Data: Independent vectors $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ with distribution

$$\mathbf{x}_i \sim f_i(\mathbf{x}_i | \boldsymbol{\theta}), \text{ for } i = 1, \dots, n$$

Log-likelihood function of unknown $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ is

$$l(\boldsymbol{\theta}) = \log f(\mathbf{x} | \boldsymbol{\theta}) = \log \left(\prod_{i=1}^n f_i(\mathbf{x}_i | \boldsymbol{\theta}) \right)$$

$$\text{where } \mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$$

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

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

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

BIC approaches to BF's further ignore the c_π 's

Some of the problems with BIC

- It ignores the constant c_π from the prior
- Problems with p .
 - What is p with random effects, latent variables, ... etc. ?
 - Often p grows with n
- Problems with n .
 - 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, ...)

Example: Group means

For $i = 1, \dots, p$ and $l = 1, \dots, r$, let

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

- Since we have r observations in each of p groups the total number of real observations ('cases') is pr and it might seem that, in BIC, $n = pr$
- Following Schwarz, however, one would have vector observations $\mathbf{X}_l = (X_{1l}, \dots, X_{pl})^t$, so that defining $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^t$

$$\mathbf{X}_l \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \quad l = 1, \dots, r$$

so that the 'sample size' appearing in BIC should be r .

- So Swartz definition of n for BIC is here number of real observations divided by number of parameters.
- While the 'effective sample size' for each μ_i is r , the effective sample size for σ^2 is pr (much larger than r for large p), so effective sample size is parameter-dependent.
- One could easily be in the situation where $p \rightarrow \infty$ but the effective sample size r (for μ) is fixed.

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 $l = 1, \dots, r$, consider the multilevel version in which

$$\mu_i \sim N(\xi, \tau^2),$$

with ξ and τ^2 being unknown.

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

- (1) If $\tau^2 = 0$, there is only two parameters: ξ and σ^2 .
- (2) If τ^2 is huge, is the number of parameters $p + 3$? (the means along with ξ, τ^2 and σ^2)

(3) But, if one integrates out $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$, then

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

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

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

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 σ^2 , $\approx p$ for ξ and τ^2 , and $\approx r$ for the μ_i 's.

Example ANOVA models:

- Let $\mathbf{Y} = (Y_1, \dots, Y_n)^t \sim N_n(\mathbf{X} \boldsymbol{\beta}, \sigma^2 \mathbf{I})$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$ and σ^2 are unknown
- \mathbf{X} is a given $n \times p$ matrix of 1's and -1's with orthogonal columns
- The information matrix for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is

$$\hat{\mathbf{I}} = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} I_{p \times p} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix},$$

so that now the effective sample size appears to be n for all parameters.

Example: Common mean, differing variances:

(Cox mixtures)

Suppose each independent observation X_i , $i = 1, \dots, n$, 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 $n/2$.

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

EBIC: a proposed solution

- Based on a modified Laplace approximation to $m(\boldsymbol{x})$ for *good* priors:
 - The *good* priors can deal with growing p
 - The Laplace approximation
 - * retains the constant c_π in the expansion
 - * often good for moderate n
- Needs 'effective sample size' (see later)
- Ideally, should be computable with standard software deriving mle's and observed information matrices.

Laplace approximation

1. Preliminary 'nice' reparameterization.

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

2. Taylor expansion

By a Taylor's series expansion of $e^{l(\boldsymbol{\theta})}$ about the mle $\hat{\boldsymbol{\theta}}$,

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

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

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

3. Approximation to marginal $m(\mathbf{x})$

If $\hat{\boldsymbol{\theta}}$ 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

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

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 close form.

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

Choosing a 'good' prior

The procedure involves integrating out the 'common parameters', orthogonalizing and assigning appropriate univariate robust priors to each component

1. Integrating out 'common' parameters

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

- Let $\theta = (\theta_{(1)}, \theta_{(2)})$, where $\theta_{(2)}$ denotes the parameters that are to be integrated out

- Partition the observed information matrix accordingly as

$$\hat{\mathbf{I}} = \begin{pmatrix} \hat{\mathbf{I}}_{11} & \hat{\mathbf{I}}_{12} \\ \hat{\mathbf{I}}_{21} & \hat{\mathbf{I}}_{22} \end{pmatrix}$$

- Let Σ be the Covariance matrix of $\boldsymbol{\theta}_{(1)}$, that is

$$\Sigma^{-1} = \hat{\mathbf{I}}_{11} - \hat{\mathbf{I}}_{12} \hat{\mathbf{I}}_{22}^{-1} \hat{\mathbf{I}}_{12}^t.$$

- Integrating out $\boldsymbol{\theta}_{(2)}$ yields

$$\begin{aligned} m(\mathbf{x}) &\approx e^{l(\hat{\boldsymbol{\theta}})} \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \hat{\mathbf{I}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right) d\boldsymbol{\theta}_{(2)} \pi(\boldsymbol{\theta}_{(1)}) d\boldsymbol{\theta}_{(1)} \\ &\approx e^{l(\hat{\boldsymbol{\theta}})} (2\pi)^{\frac{p}{2}} |\hat{\mathbf{I}}|^{-\frac{1}{2}} \int \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\theta}_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^t \Sigma^{-1} (\boldsymbol{\theta}_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})\right)}{(2\pi)^{p_1/2} |\Sigma|^{1/2}} \pi(\boldsymbol{\theta}_{(1)}) d\boldsymbol{\theta}_{(1)} \end{aligned}$$

where p_1 is the dimension of $\boldsymbol{\theta}_{(1)}$.

Remarks:

- We integrate out common parameters to minimize the effect of the prior. Improper priors can be used (Berger, Peric. & Vars.)
- Constant priors produce closed form expression (and the previous “Preliminary Reparameterization” helps)

2. Orthogonalization

Next we orthogonalize the remaining parameter, $\boldsymbol{\theta}_{(1)}$ as follows:

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

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

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

- We now use a prior that is independent in the ξ_i i.e,

$$\pi(\xi) = \prod_{i=1}^{p_1} \pi_i(\xi_i)$$

- The approximation to $m(\mathbf{x})$ is then

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

where we recall that d_i are the diagonal elements of \mathbf{D} from $\boldsymbol{\Sigma} = \mathbf{O}^t \mathbf{D} \mathbf{O}$, that is, $\text{Var}(\xi_i)$ (for orthogonalized problems with not common parameters integrated out, $\mathbf{D} = \hat{\mathbf{I}}^{-1}$).

They will appear often in what follows.

3. Univariate testing priors

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

- (a) Jeffreys recommended the Cauchy(0, b_i) density

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

where $(b_i)^{-1} = \frac{(d_i)^{-1}}{n_i}$ is the *unit information* for ξ_i with n_i being the “effective sample size” for ξ_i (Kass & Wasserman, 1995).

Note: For i.i.d. scenarios, d_i is like σ_i^2/n_i so b_i is like the variance of one observation.

- (b) Use other sensible (objective) testing priors, like the *intrinsic prior*.
- (c) We propose use of a prior proposed in Berger (1985) which is very close to both the Cauchy and the intrinsic priors and produces close form expressions for $m(\boldsymbol{x})$ for normal likelihoods

4. Berger's robust priors

Berger's priors were introduced in the context of robustness analyses and we refer to them as "robust priors". Their important role in model selection has gone unnoticed so far.

These priors do not have closed form expressions. The univariate robust prior is given, for if $b_i \geq d_i$, by

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

Interestingly, it is within 25% of the Cauchy and 5% of the intrinsic densities.

We will interpret this prior (and b_i) exactly as we would the Cauchy prior.

There are multivariate versions which we we'll explore elsewhere.

5. **Approximation to $m(\mathbf{x})$**

Remarkably enough, Berger robust priors, in spite of not having themselves close form expressions, produce close form expressions for the integrals we are interested in. It suffices to integrate first over ξ_i and then over λ_i to get

$$m(\mathbf{x}) \approx e^{l(\hat{\boldsymbol{\theta}})} (2\pi)^{p/2} |\hat{\mathbf{I}}|^{-1/2} \left[\prod_{i=1}^{p_1} \frac{1}{\sqrt{2\pi(d_i + b_i)}} \frac{\left(1 - e^{-\hat{\xi}_i^2/(d_i + b_i)}\right)}{\sqrt{2} \hat{\xi}_i^2 / (d_i + b_i)} \right]$$

where $(d_i)^{-1}$ is global information about ξ_i , and $(b_i)^{-1}$ is unit information.

Recall, the approximation is of order $o_n(1)$; it is *exact* for normal likelihoods.

6. Proposal for EBIC

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

$$\text{EBIC} \equiv 2l(\hat{\boldsymbol{\theta}}) + (p - p_1) \log(2\pi) - \log |\hat{\boldsymbol{\Gamma}}_{22}| - \sum_{i=1}^{p_1} \log(1 + n_i) \\ + 2 \sum_{i=1}^{p_1} \log \frac{(1 - e^{-v_i})}{\sqrt{2} v_i}, \quad \text{where } v_i = \frac{\hat{\xi}_i^2}{b_i + d_i}.$$

(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

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

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

- Compare this EBIC and BIC:

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

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

where $k_{\pi} > 0$ is the 'ignored' constant from π

- 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.

EBIC*: A Modification More Favorable to Complex Models

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

Raftery (1996) proposed unit-information normal priors centered at the mle's for the parameters, but this seems to favor complex models too much.

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

This can be viewed as the empirical Bayes alternative, popularized in the robust Bayesian literature (Berger, 1994)

The b_i that maximizes $m(\boldsymbol{x})$ can easily be seen to be

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

Problem: when $\xi_i = 0$, this empirical Bayes choice can result in inconsistency as $n_i \rightarrow \infty$.

Solution: prevent \hat{b}_i from being less than $n_i d_i$. This results, after an accurate approximation (for fixed p) in

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

again a very simple expression

Defining the 'effective sample size' n_j for ξ_j :

Here, we only have tentative proposals. Many more options are being explored as we speak and still much work is needed. The proposal, however, has been applied to many challenging examples with satisfactory results (in particular to all mentioned when exposing problems with BIC at the beginning)

Our motivating idea arose in the easiest situations when

1. The Information matrices (both observed and expected) are diagonal (so no orthogonalization is needed), and
2. We have independent 'cases' x_i (possibly vectors), $i = 1, \dots, n$, from some population. (This seems to be the situation in many social science problems).

Let $\mathbf{I} = I_{jk}$ be the *expected* information matrix, evaluated at the MLE, with (j, k) entry

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

Compute also, for each case \mathbf{x}_i , the associated expected information matrix $\mathbf{I}_i = (I_{i,jk})$, having (j, k) entry

$$I_{i,jk} = -E^{\mathbf{X}_i | \boldsymbol{\theta}} \left. \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_i(\mathbf{x}_i | \boldsymbol{\theta}) \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} .$$

Note that this is evaluated at the MLE $\hat{\boldsymbol{\theta}}$ from the full data.

The *ad hoc* preliminary proposed definition of the effective sample size n_j for θ_j relies on the property that the global information about

θ_j is the sum of the information in each observation:

$$I_{jj} = \sum_{i=1}^n I_{i,jj} .$$

When this is satisfied, it seems natural to 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 θ_j as

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

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

Stone counter example

Stone(1979) showed in an example that BIC can be inconsistent when p grows with n . Here we use the simplified version in Berger et. al (2003). Consistency of EBIC is discussed later.

Data consist of r replications $X_{ij}, j = 1, \dots, r$ in each of i groups, $i = 1, \dots, p$, with $X_{ij} \sim N(\mu_i, 1)$; want to test:

$$M_1 : \boldsymbol{\mu} = \mathbf{0} \quad \text{vs} \quad M_1 : \boldsymbol{\mu} \in \mathfrak{R}^p \setminus \mathbf{0}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$.

Computation gives $\hat{\mathbf{I}} = \mathbf{I} = r\mathbf{I}_{p \times p}$ so $d_i = 1/r$, the scale of the robust prior is $b_i = 1$ and $n_i = r$. Also, $\hat{\xi}_i = \hat{\mu}_i = \bar{x}_i$, so that v_i is

$$v_i = \frac{r\bar{x}_i^2}{r+1}.$$

We now derive the approximation (exact in this case) to

$$2 \log \text{BF}_{21} \approx \Delta \text{EBIC} = \text{EBIC}_2 - \text{EBIC}_1$$

by using the proposals for EBIC before.

Note that in both expressions the leading term is the same as in the usual BIC:

$$2 \log l_2(\hat{\boldsymbol{\mu}}) - 2 \log l_1 = r \sum \bar{x}_i^2,$$

but penalties can be different.

EBIC.— easy computation shows that

$$\begin{aligned}\Delta\text{EBIC} &= r \sum \bar{x}_i^2 - \sum \log(1 + n_i) + 2 \sum \log \frac{1 - e^{-v_i}}{v_i \sqrt{2}} \\ &= r \sum \bar{x}_i^2 - 2 \sum \log(r \bar{x}_i^2) + 2 \sum \log(1 - e^{-\frac{r \bar{x}_i^2}{r+1}}) + p \log \frac{r+1}{2}\end{aligned}$$

EBIC* (approximation).— computation gives

$$\Delta\text{EBIC}^* = r \sum_{i=1}^p \bar{x}_i^2 - p \log(1 + r) - \sum_{i=1}^p \log(3v_i + 2 \max\{v_i, 1\}).$$

Assuming the \bar{x}_i^2 ordered increasingly in magnitude, and letting p_s be such that $\bar{x}_i^2 \leq (r+1)/r$ iff $i \leq p_s$, an alternative expression is

$$\Delta\text{EBIC}^* = r \sum_{i=1}^p \bar{x}_i^2 - \sum_{i=1}^{p_s} \log(3r\bar{x}_i^2 + 2r + 2) - \sum_{i=p_s+1}^p \log(5r\bar{x}_i^2),$$

a remarkably simple expression.

A small comparative simulation

Berger, Ghosh and Mukhopadhyay (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 EBIC'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 $p \rightarrow \infty$.

We next compare our EBIC'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 Δ GBIC (denoted Δ_O), our Δ EBIC proposal (denoted Δ_N), the robust modification (denoted Δ_R) and its approximation Δ EBIC* (denoted Δ_*).

		$\mu = 0$		$\mu_i \sim N(2, 1)$		$\mu_i \sim Ex(\mu = 2)$		$\mu_1 = 10, \mu_i = 0$	
p, r	Δ EBIC	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$p = 2$	Δ_o	0.383	1.38	17.89	9.17	7.839	6.18	180.56	27.65
$r = 2$	Δ_N	-2.155	1.42	16.54	8.64	7.350	6.12	187.08	27.96
(p=5 last column)	Δ_R	-2.117	1.54	19.58	9.89	8.911	7.07	194.33	28.26
	Δ_*	-2.226	1.45	17.58	9.18	7.826	6.52	190.22	28.10
$p = 15$	Δ_o	-1.64	1.65	92.96	20.31	257.01	33.54	157.6	25.93
$r = 2$	Δ_N	-16.47	4.18	87.66	19.65	258.02	33.41	175.64	27.44
	Δ_R	-16.20	4.58	103.56	22.28	281.31	34.93	183.05	27.75
	Δ_*	-17.00	4.28	93.00	20.77	267.21	34.10	178.42	27.59
$p = 200$ $r = 2$	Δ_o							56.63	17.13
	Δ_N							-27.82	31.05
	Δ_R							-17.28	31.73
	Δ_*							-31.60	31.24

Table 1: For the group means problem, the means and standard deviations of various Δ EBIC \equiv EBIC $_{\mu \neq 0}$ - EBIC $_{\mu = 0}$ for sets of 500 replications, under different assumptions about the group means.

Δ_o : Cauchy, Δ_N : new EBIC, Δ_R : robust EBIC, Δ_* : approx. to robust

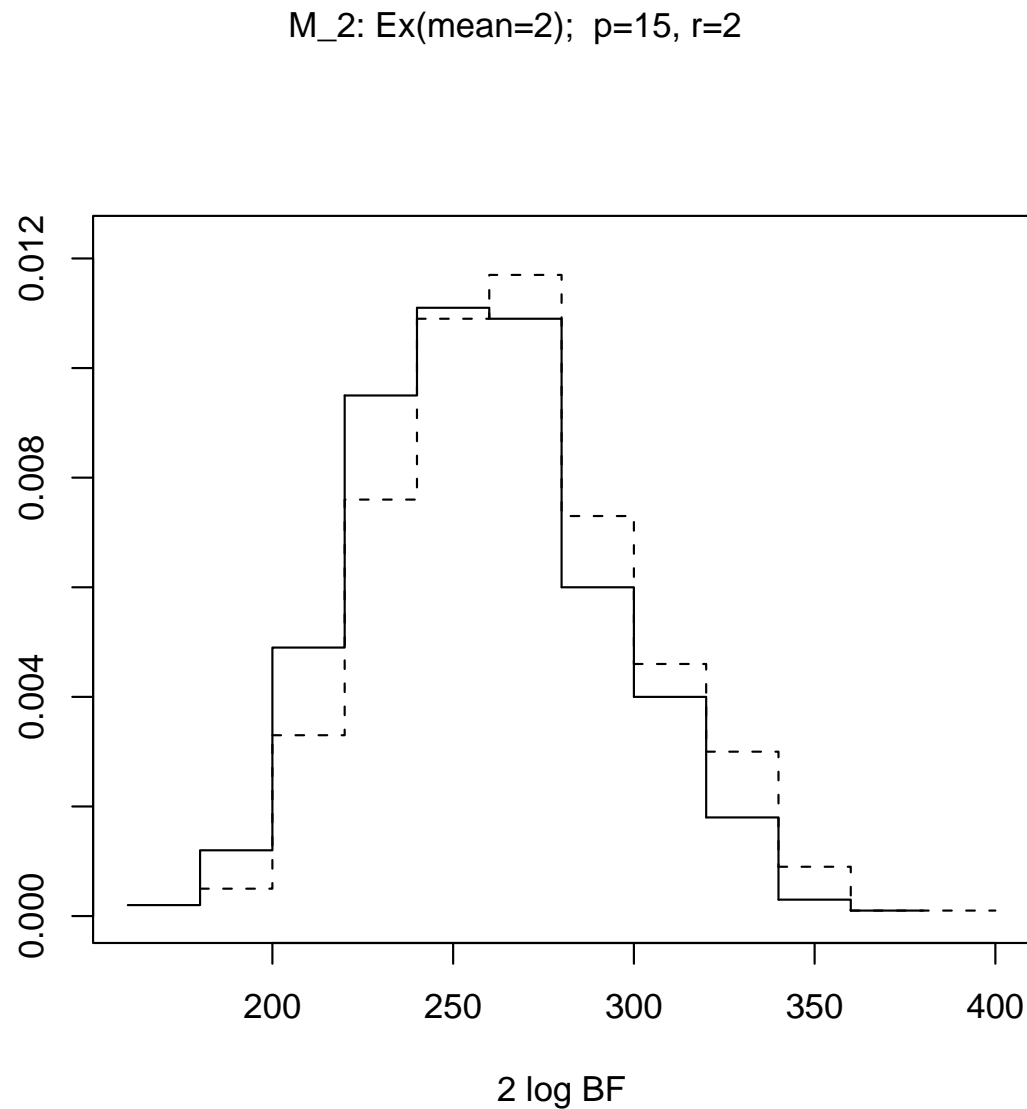


Figure 1: Simulated distributions of ΔGBIC and ΔGBIC^* when the group means are i.i.d exponential with mean 2, $p = 15$ and $r = 2$. 500 simulations

Consistency

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

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

Example 1 *Group means problem with known $\sigma^2 = 1$ and effective sample size $n_i = r$ fixed. Compare the null model $M_0 : \mu_1 = \cdots = \mu_p = 0$ with the full model $M_1 : \text{all } \mu_i \text{ nonzero. If the } \mu_i \text{ are independently assigned } N(0, \tau_i^2) \text{ priors, consistency obtains under } M_1 \text{ as } p \rightarrow \infty \text{ if and only if } V \equiv \lim_{p \rightarrow \infty} \frac{1}{p} \sum_i^p \mu_i^2 \text{ satisfies } V \geq \lim_{p \rightarrow \infty} \frac{1}{p} \sum_i^p \tau_i^2 - 1, \text{ assuming the limits exist (shown to Jim by J.K Ghosh).}$*

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 $M_1 : \text{all } \mu_i \text{ nonzero}$. EBIC and EBIC* are consistent under M_0 as $p \rightarrow \infty$. Under M_1 and assuming $V \equiv \lim_{p \rightarrow \infty} \frac{1}{p} \sum_i \mu_i^2$ exists, EBIC and EBIC* are

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

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

Note 1: Inconsistency results only when M_1 is close to M_0 .

(Mukhopadhyay, Ghosh, and Berger, 2005 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 linear regression simulation:

$$Y_{n \times 1} = X_{n \times 8} \beta_{8 \times 1} + \epsilon_{n \times 1} \quad \epsilon \sim N(0, \sigma^2),$$

$$\beta = (\mathbf{3}, \mathbf{1.5}, 0, 0, \mathbf{2}, 0, 0, 0),$$

and design Matrix $X \sim N(0, \Sigma_x)$, $\Sigma_x =$

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

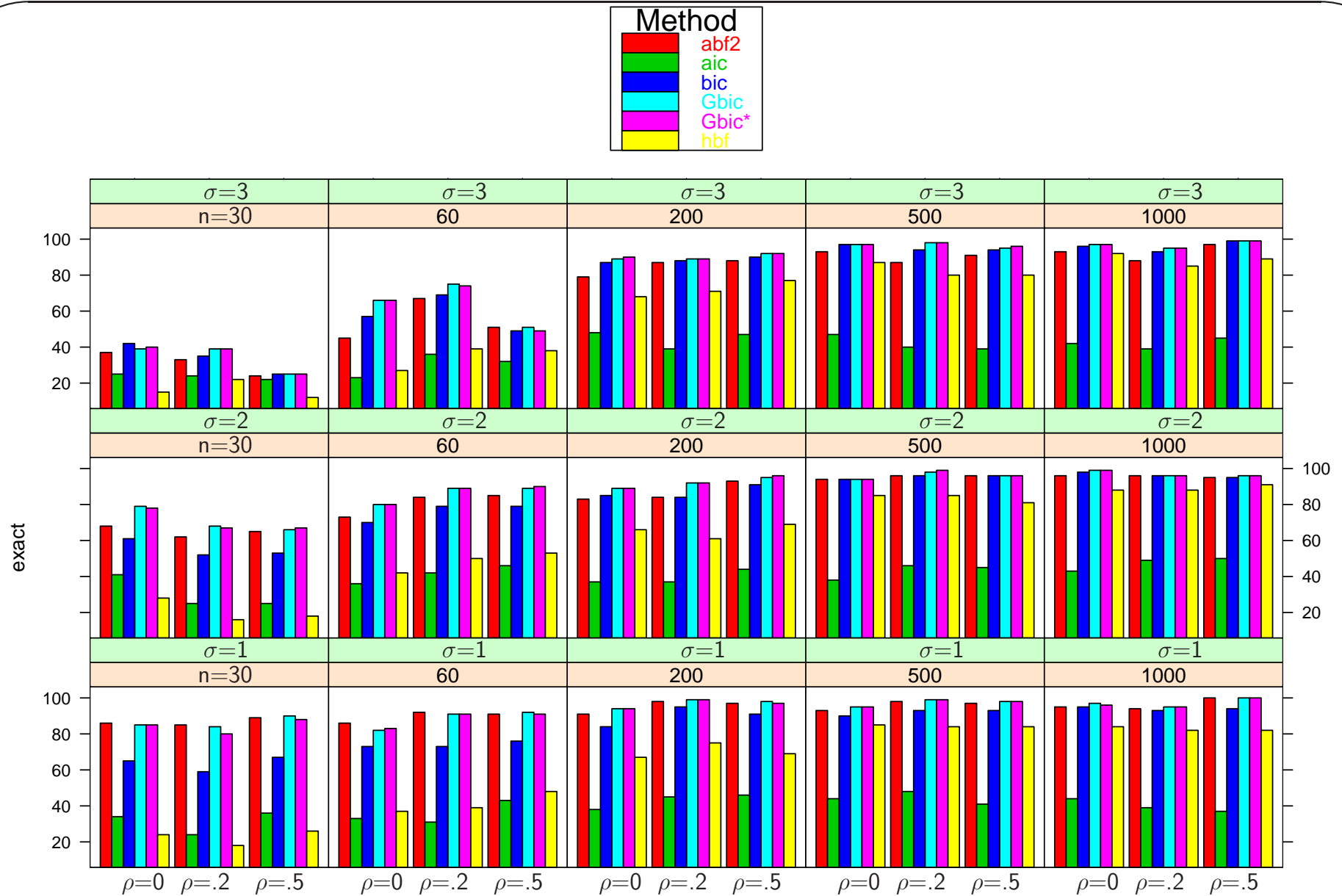


Figure 2: Results of linear regression displaying percentage of true models selected under different situations

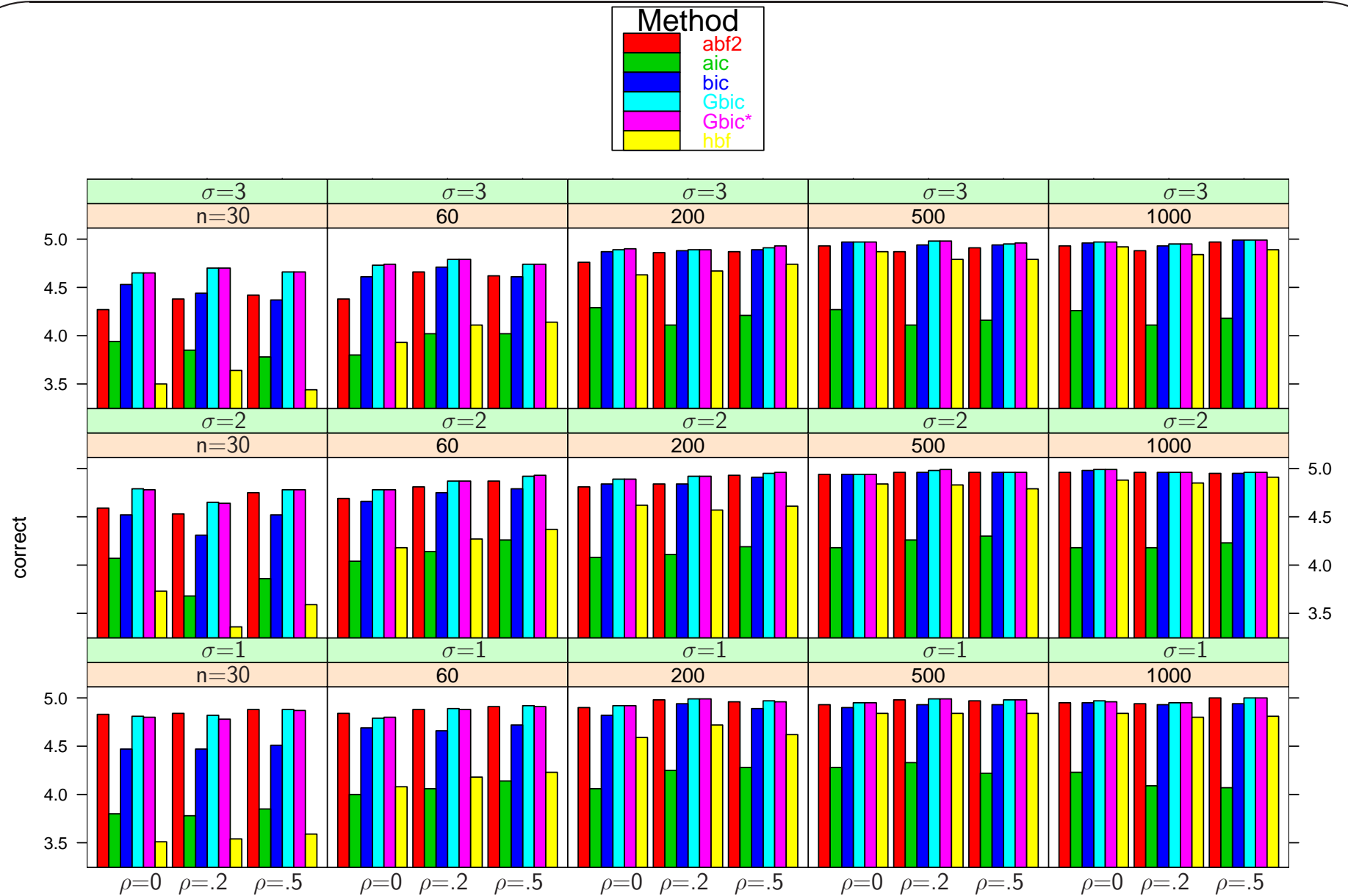


Figure 3: Results of linear regression displaying the average number of 0's that were determined to be 0.

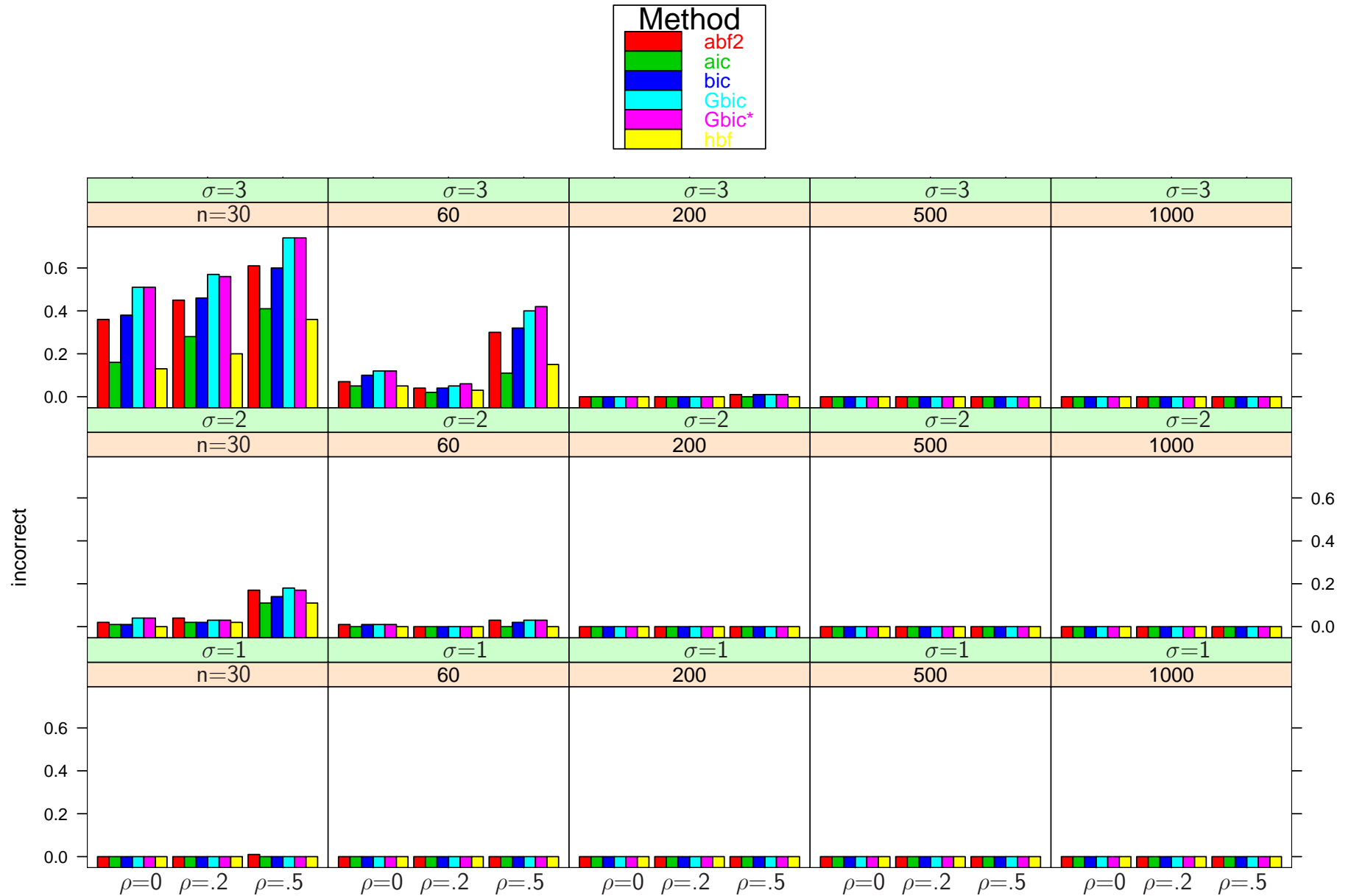


Figure 4: Results of linear regression displaying the average number of non-zeroes that were determined to be zero (small is good).

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).
Still fighting . . .
- Extension to the non-independent case

... and that's all for today

to be continued in O'Bayes 7

THANKS!