J. R. Statist. Soc. B (2017)
79, *Part* 2, *pp.* 1–38

A Bayesian information criterion for singular models

Mathias Drton

University of Washington, Seattle, USA

and Martyn Plummer

International Agency for Research on Cancer, Lyon, France

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 5th, 2016*, Professor C. Leng *in the Chair*]

Summary. We consider approximate Bayesian model choice for model selection problems that involve models whose Fisher information matrices may fail to be invertible along other competing submodels. Such singular models do not obey the regularity conditions underlying the derivation of Schwarz's Bayesian information criterion BIC and the penalty structure in BIC generally does not reflect the frequentist large sample behaviour of their marginal likelihood. Although large sample theory for the marginal likelihood of singular models has been developed recently, the resulting approximations depend on the true parameter value and lead to a paradox of circular reasoning. Guided by examples such as determining the number of components of mixture models, the number of factors in latent factor models or the rank in reduced rank regression, we propose a resolution to this paradox and give a practical extension of BIC for singular model selection problems.

Keywords: Bayesian information criterion; Factor analysis; Mixture model; Model selection; Reduced rank regression; Schwarz information criterion; Singular learning theory

1. Introduction

Information criteria are valuable tools for model selection (Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Konishi and Kitagawa, 2008). At a high level, they fall into two categories (Yang, 2005; van Erven *et al*., 2012; Wit *et al*., 2012). On one side, there are criteria that target good predictive behaviour of the selected model. For instance, cross-validation-based scores assess the quality of out-of-sample predictions by splitting available data into test and training cases, and Akaike's information criterion AIC provides an estimate of an out-of-sample prediction (or generalization) error that is justified via asymptotic distribution theory for large samples (Akaike, 1974). Following a different philosophy that will be the focus of this paper, the Bayesian information criterion BIC of Schwarz (1978) draws motivation from Bayesian inference. Schwarz's criterion aims to capture key features of posterior model uncertainty via a penalty that is motivated by the large sample properties of the marginal likelihood (which is also commonly referred to as integrated likelihood or evidence). In a nutshell, under suitable regularity conditions, a quadratic approximation to the log-likelihood function can be used to relate the marginal likelihood to a Gaussian integral in which the sample size acts as an inverse variance. This dependence of the Gaussian integral on the sample size leads to the familiar

Address for correspondence: Mathias Drton, Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA.
E-mail: md5@uw.edu

 1369–7412/17/79000

BIC penalty term that, on the log-scale, consists of the product of model dimension and the logarithm of the sample size.

BIC penalizes model complexity more heavily than predictive criteria such as AIC. From the frequentist perspective, it has been shown that BIC's penalty depends on the sample size in a way that makes the criterion consistent for a wide range of problems. In other words, when optimizing BIC the probability of selecting a fixed most parsimonious true model tends to 1 as the sample size tends to ∞ (e.g. Nishii (1984), Haughton and 1988, 1989). However, a wide range of penalties would yield consistency of a model selection score, and it is instead the aim of capturing the asymptotic scaling of the marginal likelihood that leads to the familiar dependence on dimension and log-sample size. Indeed, from a Bayesian point of view, BIC supplies rather crude but computationally inexpensive proxies to otherwise difficult-to-calculate posterior model probabilities, which form the basis for Bayesian model choice and averaging; see Kass and Wasserman (1995), Raftery (1995), DiCiccio *et al*. (1997), Hoeting *et al*. (1999) or Hastie *et al*. (2009), chapter 7.7.

In this paper, we are concerned with BICs in the context of singular model selection problems, i.e. problems that involve models with Fisher information matrices that may fail to be invertible. For example, owing to the breakdown of parameter identifiability, the Fisher information matrix of a mixture model with three components is singular at a distribution that can be obtained by mixing only two components. This clearly presents a fundamental challenge for selection of the number of components. In particular, when the Fisher information matrix is singular, the log-likelihood function does not admit a large sample approximation by a quadratic form. Rotnitzky *et al*. (2000) illustrated some of the resulting difficulties in asymptotic distribution theory under an assumption of identifiability. Non-identifiability of parameters, as present in the examples that we shall consider, leads to considerably more complicated scenarios as discussed, for instance, by Liu and Shao (2003) and Azaïs *et al*. (2006, 2009). The key obstruction to justifying BIC is that in singular models there need no longer be a connection between the Bayesian marginal likelihood and a Gaussian integral. In particular, a parameter count or model dimension may fail to capture the asymptotic scaling of the marginal likelihood (Watanabe, 2009). We illustrate this fact in the following example.

1.1 Example 1

Suppose that $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nn})$ is a sample of independent and identically distributed observations whose unknown distribution is modelled as a mixture of two normal distributions. Specifically, the data-generating distribution is assumed to be of the form

$$\pi(\alpha, \mu_1, \mu_2) := \alpha \mathcal{N}(\mu_1, 1) + (1-\alpha)\mathcal{N}(\mu_2, 1),$$

where $\alpha \in [0, 1]$ is an unknown mixture weight, $\mu_1, \mu_2 \in \mathbb{R}$ are two unknown means and the variances are known and equal to 1. To exemplify later notation, we write out the likelihood function of the mixture model $\mathcal{M}$ considered, which maps the parameter vector (α, μ_1, μ_2) to

$$P\{\mathbf{Y}_n | \pi(\alpha, \mu_1, \mu_2), \mathcal{M}\} = \prod_{i=1}^{n} \{\alpha\varphi(Y_{ni} - \mu_1) + (1-\alpha)\varphi(Y_{ni} - \mu_2)\}.$$

Here, φ denotes the standard normal density. As a prior for Bayesian inference, consider a uniform distribution for α, and take μ_1 and μ_2 to be independent $\mathcal{N}(0, 16)$. Then the marginal likelihood of model $\mathcal{M}$ is

$$L(\mathcal{M}) = \int_{[0,1]\times\mathbb{R}^2} P\{\mathbf{Y}_n | \pi(\alpha, \mu_1, \mu_2), \mathcal{M}\}\, \varphi(\mu_1/4)\, \varphi(\mu_2/4)\, \mathrm{d}(\alpha, \mu_1, \mu_2).$$

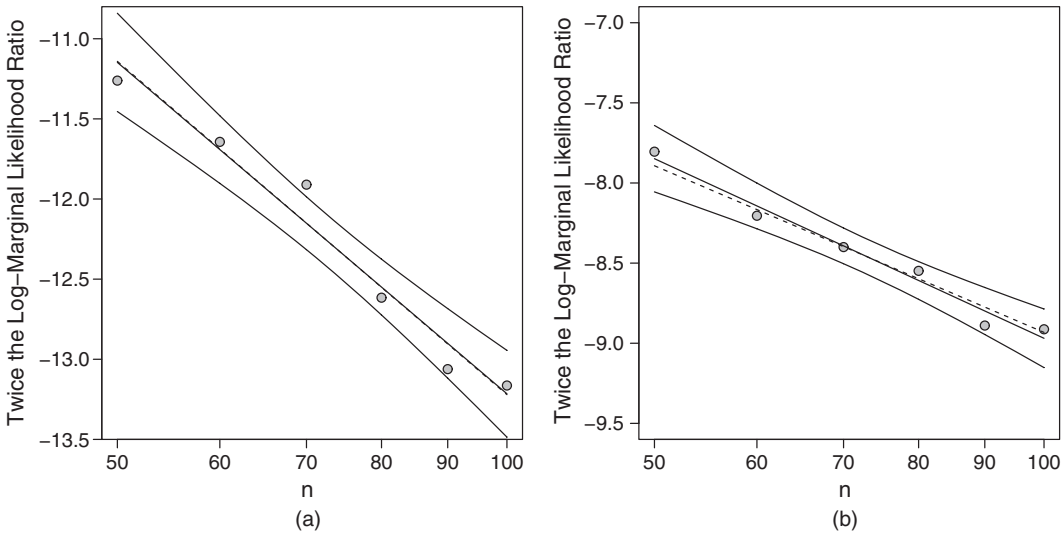

Fig. 1. Averages of twice the log-marginal-likelihood ratio for a Gaussian mixture model, least squares line, simultaneous confidence bands and theory-based slope (– – –): (a) data from a two-component mixture; (b) standard normal data

We now simulate values of the random variable $L(\mathcal{M})$. For each choice of a sample size $n \in \{50, 60, \ldots, 100\}$, we generate 200 independent realizations of $L(\mathcal{M})$, drawing the sample $\mathbf{Y}_n$ from the normal mixture π_0 given by $\alpha = 0.4$, $\mu_1 = -2$ and $\mu_2 = 2$. Following Neal (1999), we compute each value of $L(\mathcal{M})$ by standard Monte Carlo sampling with 10^7 draws from the prior. To allow for comparisons across different samples, we consider the marginal likelihood ratio $L_0(\mathcal{M})$ that is obtained by dividing $L(\mathcal{M})$ by $P(\mathbf{Y}_n|\pi_0)$, which is the likelihood of the sample under the true distribution. The results are summarized in Fig. 1(a), which plots average values of $2\log\{L_0(\mathcal{M})\}$ together with a least squares line relating $2\log\{L_0(\mathcal{M})\}$ to $\log(n)$. We also show 90% simultaneous confidence bands and a line with slope determined by large sample theory. We emphasize that Fig. 1(a)'s horizontal axis has the sample size on the log-scale. The slope of the least squares line comes out to be -2.98 and is close to the slope of -3 that is predicted by the parameter count from Schwarz's BIC.

A different picture emerges, however, when we repeat the simulations changing the data-generating distribution π_0 to the standard normal distribution $\mathcal{N}(0, 1)$; see Fig. 1(b). In this case, the slope of the least squares line is no longer close to the negated parameter count. Instead, it is about -1.62. In Section 2, we discuss asymptotic theory that addresses the issue that the Fisher information matrix of $\mathcal{M}$ is singular at the standard normal distribution. For $\mathcal{N}(0, 1)$ data in this example, the theory predicts a slope of -1.5 (Aoyagi, 2010a). This large sample line is contained in the simultaneous confidence bands that we give in Fig. 1(b).

As we shall review in Section 2, refined mathematical knowledge about the asymptotic scaling of the marginal likelihood of singular models has been obtained in recent years. It is desirable to leverage this knowledge when defining an information criterion that is inspired by Bayesian methods. However, it is not immediately clear how to cope with the fact that even the most basic features of the asymptotics for the marginal likelihood depend on the unknown data-generating distribution. The generalization of BIC that we introduce in this paper resolves this issue by averaging different approximations in a data-dependent way.

As conveyed by the above example, the selection of the number of mixture components constitutes a singular model selection problem. Other important examples of this type include determining the rank in reduced rank regression, the number of factors in factor analysis or the number of states in latent class or hidden Markov models. More generally, all the classical hidden or latent variable models are singular, which expresses itself also in complicated geometry of the parameter space or set of distributions (Geiger *et al*., 2001; Drton *et al*., 2007; Zwiernik and Smith, 2012; Allman *et al*., 2015; Gassiat and van Handel, 2014).

Despite the possible disconnect between penalization based on model dimension alone and the large sample behaviour of Bayesian methods, the standard BIC is a state of the art method for many singular model selection problems; see for example McLachlan and Peel (2000), section 6.9, Steele and Raftery (2010) and Baudry and Celeux (2015) for mixture models and Lopes and West (2004) for factor analysis. From the frequentist perspective, BIC is known to be consistent in many singular settings (Keribin, 2000; Drton *et al*. (2009), chapter 5.1). However, as mentioned earlier, consistency can be achieved with many penalization schemes, which would not need to depend logarithmically on the sample size.

In this paper, we propose a generalization of the BIC that utilizes refined mathematical information about the marginal likelihood of the statistical models considered: information that goes beyond mere model dimension. Schwarz's BIC is Bayesian in the sense that it differs from the log-marginal-likelihood only by terms that are bounded. The new criterion, sBIC, maintains this connection to Bayesian model choice also in singular settings. Our sBIC-criterion preserves consistency properties of BIC and is an honest generalization of the standard criterion in the sense that sBIC coincides with Schwarz's BIC when the model is regular. sBIC is designed to capture the key features of posterior model uncertainty, but our numerical work shows that it can also lead to improved frequentist model selection properties.

The new criterion is presented in Section 3, which is preceded by a review of the theory that sBIC is built on (Section 2). This theory was developed over the last decade by Watanabe (2001, 2009). The large sample properties of sBIC are shown in Section 4. We first show consistency and then clarify the connection to the log-marginal-likelihood. Numerical examples demonstrating the use of sBIC are given in Sections 5 and 6. The former section focuses on problems from multivariate analysis, namely, reduced rank regression and factor analysis. The latter section treats mixture models, where it becomes particularly apparent that the choice of the prior distribution in each (singular) model has an influence on the form of sBIC—as a reader familiar with the work of Rousseau and Mengersen (2011) may suspect. We conclude the paper with a discussion of the strengths and the limitations of the proposed methodology in Section 7.

2. Background

Let $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nn})$ denote a sample of n independent and identically distributed (IID) observations, and let $\{\mathcal{M}_i : i \in I\}$ be a finite set of candidate models for the distribution of these observations. For a Bayesian treatment, suppose that we have positive prior probabilities $P(\mathcal{M}_i)$ for the models and that, in each model $\mathcal{M}_i$, a prior distribution $P(\pi_i|\mathcal{M}_i)$ is specified for the probability distributions $\pi_i \in \mathcal{M}_i$. Write $P(\mathbf{Y}_n|\pi_i, \mathcal{M}_i)$ for the likelihood of $\mathbf{Y}_n$ under data-generating distribution π_i from model $\mathcal{M}_i$, and let

$$L(\mathcal{M}_i) := P(\mathbf{Y}_n|\mathcal{M}_i) = \int_{\mathcal{M}_i} P(\mathbf{Y}_n|\pi_i, \mathcal{M}_i)\, \mathrm{d}P(\pi_i|\mathcal{M}_i) \tag{2.1}$$

be the marginal likelihood of model $\mathcal{M}_i$. Bayesian model choice is then based on the posterior model probabilities

$$P(\mathcal{M}_i|\mathbf{Y}_n) \propto P(\mathcal{M}_i)\, L(\mathcal{M}_i), \qquad i \in I.$$

The probabilities $P(\mathcal{M}_i|\mathbf{Y}_n)$ can be approximated by various Monte Carlo procedures—see Friel and Wyse (2012) for a recent review—but practitioners also often turn to computationally inexpensive proxies that are suggested by large sample theory. These proxies are based on the asymptotic properties of the sequence of random variables $L(\mathcal{M}_i)$ that are obtained when $\mathbf{Y}_n$ is drawn from a data-generating distribution $\pi_0 \in \mathcal{M}_i$, and we let the sample size n grow.

In practice, a prior distribution $P(\pi_i|\mathcal{M}_i)$ is typically specified by placing a distribution on the vector of parameters appearing in a parameterization of $\mathcal{M}_i$; recall example 1. So assume that

$$\mathcal{M}_i = \{\pi_i(\boldsymbol{\omega}_i) : \boldsymbol{\omega}_i \in \Omega_i\} \tag{2.2}$$

with d_i-dimensional parameter space $\Omega_i \subseteq \mathbb{R}^{d_i}$, and that $P(\pi_i|\mathcal{M}_i)$ is the transformation of a distribution $P(\boldsymbol{\omega}_i|\mathcal{M}_i)$ on Ω_i under the map $\boldsymbol{\omega}_i \mapsto \pi_i(\boldsymbol{\omega}_i)$. The marginal likelihood is then the d_i-dimensional integral

$$L(\mathcal{M}_i) = \int_{\Omega_i} P\{\mathbf{Y}_n|\pi_i(\boldsymbol{\omega}_i), \mathcal{M}_i\}\, \mathrm{d}P(\boldsymbol{\omega}_i|\mathcal{M}_i). \tag{2.3}$$

Now the observation of Schwarz and other subsequent work is that, under suitable technical conditions on the model $\mathcal{M}_i$, the parameterization $\boldsymbol{\omega}_i \mapsto \pi_i(\boldsymbol{\omega}_i)$ and the prior distribution $P(\boldsymbol{\omega}_i|\mathcal{M}_i)$, it holds for all $\pi_0 \in \mathcal{M}_i$ that

$$\log\{L(\mathcal{M}_i)\} = \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \frac{d_i}{2}\log(n) + O_p(1). \tag{2.4}$$

Here, $P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)$ is the maximum of the likelihood function, and $O_p(1)$ stands for a sequence of remainder terms that is bounded in probability. The first two terms on the right-hand side of equation (2.4) can be evaluated in statistical practice and may be used as a model score or a proxy for the logarithm of the marginal likelihood. The resulting *Bayesian* or *Schwarz's information criterion* for model $\mathcal{M}_i$ is

$$\mathrm{BIC}(\mathcal{M}_i) = \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \frac{d_i}{2}\log(n). \tag{2.5}$$

Briefly put, the large sample behaviour from equation (2.4) relies on the following properties of regular problems. First, with high probability, the integrand in equation (2.3) is negligibly small outside a neighbourhood of the maximum likelihood estimator of $\boldsymbol{\omega}_i$. Second, in such a neighbourhood, the log-likelihood function $\log[P\{\mathbf{Y}_n|\pi_i(\boldsymbol{\omega}_i\}, \mathcal{M}_i\}]$ can be approximated by a negative definite quadratic form, whereas a smooth prior $P(\boldsymbol{\omega}_i|\mathcal{M}_i)$ is approximately constant. The integral in equation (2.3) may thus be approximated by the product of $P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)$ and a Gaussian integral, in which the inverse covariance matrix equals n times the Fisher information. This d_i-dimensional Gaussian integral depends on n via the multiplicative factor $n^{-d_i/2}$, and taking logarithms we arrive at equation (2.4). We remark that this approach also allows for estimation of the remainder term in equation (2.4), giving a Laplace approximation with error $O_p(n^{-1/2})$ as discussed, for instance, in Tierney and Kadane (1986), Haughton (1988), Kass and Wasserman (1995) or Wasserman (2000).

A large sample quadratic approximation to the log-likelihood function is not possible, however, when the Fisher information matrix is singular. Consequently, the classical theory that was alluded to above does not apply to singular models. Indeed, model (2.4) is generally false in singular models. Nevertheless, asymptotic theory for the marginal likelihood of singular models has been developed over the last decade, culminating in Watanabe (2009). Indeed, theorem 6.7

in Watanabe (2009) shows that a wide variety of singular models have the property that, for $\mathbf{Y}_n$ drawn from $\pi_0 \in \mathcal{M}_i$,

$$\log\{L(\mathcal{M}_i)\} = \log\{P(\mathbf{Y}_n|\pi_0, \mathcal{M}_i)\} - \lambda_i(\pi_0)\log(n) + \{m_i(\pi_0) - 1\}\log\{\log(n)\} + O_p(1); \quad (2.6)$$

see also the introduction to the topic in Drton *et al.* (2009), chapter 5.1. In this paper, we follow the terminology of Watanabe, (2009) and refer to the quantity $\lambda_i(\pi_0)$ as the *learning coefficient*. However, other terminology such as *real log-canonical-threshold* or *stochastic complexity* is in use. The number $m_i(\pi_0)$ is the *multiplicity* of the learning coefficient or real log-canonical-threshold. In contrast with the regular case, it is generally very difficult to estimate the $O_p(1)$ remainder term in equation (2.6). We are not aware of any successful work on higher order approximations in statistically relevant singular settings.

Remark 1. The theorem giving equation (2.6) is developed under the 'fundamental conditions (I) and (II)' from definitions 6.1 and 6.3 in Watanabe (2009). Although the precise nature of these conditions is not important for the developments in this paper, we would like to summarize them briefly. Under conditions (I) and (II), the distributions in $\mathcal{M}_i$ share a common support and have densities with respect to a dominating measure. The parameter space Ω_i in equation (2.2) is compact and defined by real analytic constraints. (An assumption of compactness is needed only when the set of parameter vectors representing the true distribution $\{\boldsymbol{\omega}_i \in \Omega_i : \pi(\boldsymbol{\omega}_i) = \pi_0\}$ is not already compact.) Watanabe's conditions further require that the log-likelihood ratios of π_0 with respect to the distributions $\pi(\boldsymbol{\omega}_i)$ can be bounded by a function that is square integrable under π_0. Moreover, the log-likelihood ratios satisfy a requirement of analyticity that allows for power series expansions in $\boldsymbol{\omega}_i$. Finally, the prior distribution $P(\boldsymbol{\omega}_i|\mathcal{M}_i)$ has a density that is the product of a smooth positive function and a non-negative analytic function.

Watanabe's result applies to models such as reduced rank regression, factor analysis, binomial mixtures and latent class analysis, which we shall consider in the numerical experiments of Sections 5 and 6. Via suitable analytic bounds on the log-likelihood ratios, the result can also be extended to other 'non-analytic models', such as mixtures of normal distributions with known common variance as we considered in example 1 (Watanabe (2009), section 7.8). Although the case of Gaussian mixtures with unknown variance has not yet been treated explicitly in the literature, we show experiments with such models in Section 6.

2.1. Example 2

Let $\mathcal{M}_2$ be the Gaussian mixture model with $i = 2$ components that we considered in Example 1. If π_0 is a normal distribution $\mathcal{N}(\mu, 1)$ then $\lambda_2(\pi_0) = \frac{3}{4}$. If π_0 is an honest mixture of two normal distributions with variance 1 then $\lambda_2(\pi_0) = \frac{3}{2}$. In either case $m_2(\pi_0) = 1$. The values can be found in example 3.1 of Aoyagi (2010a). (The formula for the multiplicity in theorem 3.4 in Aoyagi (2010a) applies only if $r < H$, in the notation that is used there. If $r = H$, the multiplicity is 1, as confirmed in private communication with Aoyagi.)

Reduced rank regression, factor analysis and latent class analysis are all singular submodels of an exponential family, which is either the normal or the multinomial family. It follows that the sequence of likelihood ratios $P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)/P(\mathbf{Y}_n|\pi_0, \mathcal{M}_i)$ converges in distribution and, in particular, is bounded in probability (Drton, 2009). In this case, we can plug the maximum likelihood estimator into the first term of equation (2.6) and obtain that

$$\log\{L(\mathcal{M}_i)\} = \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \lambda_i(\pi_0)\log(n) + \{m_i(\pi_0) - 1\}\log\{\log(n)\} + O_p(1). \quad (2.7)$$

For more complicated models, such as Gaussian mixture models, likelihood ratios can often be shown to converge in distribution under compactness assumptions on the parameter space; see for instance Azaïs *et al.* (2006, 2009) who also reviewed much of the relevant literature. Without compactness, the log-likelihood ratios in mixture models need not be bounded in probability; for example, they would be of order $O_p[\log\{\log(n)\}]$ for Gaussian mixtures (Hartigan, 1985; Bickel and Chernoff, 1993).

Having estimated the log-likelihood by estimating the unknown data-generating distribution π_0, it seems tempting similarly to estimate the learning coefficient $\lambda_i(\pi_0)$ and its multiplicity $m_i(\pi_0)$. However, in contrast with the likelihood function, the learning coefficient and multiplicity are not continuous functions of π_0. Hence, substituting an estimate for π_0 is of little interest as the resulting expression fails to capture the behaviour of the marginal likelihood at (or near) model singularities. Instead, we shall make the fact from equation (2.7) the point of departure in the definition of our singular BIC, which is the topic of Section 3.

As in the original work of Schwarz (1978), our general treatment will focus on prior distributions with smooth densities that are bounded and positive. On a compact set, such a density will be bounded away from zero. In the analytic settings that were considered in Watanabe (2009), it then holds that $\lambda_i(\pi_0)$ is a rational number in $[0, d_i/2]$ and $m_i(\pi_0)$ is an integer in $\{1, \ldots, d_i\}$. However, as mentioned above, priors with densities that are zero in parts of the parameter space can be accommodated in the framework as long as the prior density vanishes in an 'analytic fashion'. In this case, the learning coefficient may depend on the prior $P(\omega_i|\mathcal{M}_i)$ in important ways. In particular, if the prior has a density that is zero at model singularities then $\lambda_i(\pi_0)$ could exceed $d_i/2$; compare the discussion of Jeffreys's prior in theorem 7.4 in Watanabe (2009). We shall revisit the role of the prior distribution in experiments with mixture models in Sections 6.2 and 6.3.

2.2. Example 3

Reduced rank regression is multivariate linear regression subject to a rank constraint on the matrix of regression coefficients (Reinsel and Velu, 1998). Suppose that we observe n independent copies of a partitioned zero-mean Gaussian random vector $Y = (Y_R, Y_C)$, where $Y_R \in \mathbb{R}^N$ and $Y_C \in \mathbb{R}^M$. Keeping only with the most essential structure, assume that the covariance matrix of Y_C and the conditional covariance matrix of Y_R given Y_C are both the identity matrix. The reduced rank regression model $\mathcal{M}_i$ that is associated with an integer $i \geqslant 0$ then postulates that the $N \times M$ matrix π in the conditional expectation $\mathbb{E}[Y_R|Y_C] = \pi Y_C$ has rank at most i.

In a Bayesian treatment, consider the parameterization $\pi = \omega_{i2}\omega_{i1}$, with smooth and positive prior densities for $\omega_{i2} \in \mathbb{R}^{N\times i}$ and $\omega_{i1} \in \mathbb{R}^{i\times M}$. Note that, whereas the matrix π is in one-to-one correspondence with the joint distribution of Y, this is not true for the pair of matrices $\omega_i = (\omega_{i1}, \omega_{i2})$ that is used to parameterize the model. For this set-up, Aoyagi and Watanabe (2005) derived the learning coefficients $\lambda_i(\pi_0)$ and their multiplicities $m_i(\pi_0)$, where the true data-generating distribution is given by an $N \times M$ matrix π_0 of rank $j \leqslant i$. In particular, $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ depend on π_0 only through the true rank j.

For a concrete instance, take $N = 5$ and $M = 3$. Then the values of $\lambda_{ij} := \lambda_i(\pi_0)$ are listed in Table 1, and the multiplicity $m_i(\pi_0) = 1$ unless $i = 3$ and $j = 0$ in which case $m_i(\pi_0) = 2$. Note that the table entries for $j = i$ are equal to $\dim(\mathcal{M}_i)/2$, where $\dim(\mathcal{M}_i) = i(N + M - i)$ is the dimension of $\mathcal{M}_i$, which can be identified with the set of $N \times M$ matrices of rank at most i. The dimension is also the maximal rank of the Jacobian of the map $(\omega_{i1}, \omega_{i2}) \mapsto \omega_{i2}\omega_{i1}$. The singularity issues that were addressed in Watanabe's theory arise at points where the Jacobian of the parameterization fails to have maximal rank. These have $\operatorname{rank}(\omega_{i2}\omega_{i1}) < i$ and thus define a

Table 1. Learning coefficients for reduced rank regression (five responses, three covariates)†

i	*Coefficients for the following values of j:*			
	$j=0$	$j=1$	$j=2$	$j=3$
0	0			
1	3/2	7/2		
2	3	9/2	6	
3	9/2	11/2	13/2	15/2

†Model postulates rank i; the true rank is j.

distribution that also belongs to a submodel $\mathcal{M}_j \subset \mathcal{M}_i$ given by a lower rank $j < i$. This presents a challenge for model selection, which here amounts to selection of an appropriate rank.

Although the regularity conditions in its derivation are not met, it is common practice to apply the standard BIC for selection of the rank. In doing so, one typically takes $d_i = \dim(\mathcal{M}_i)$ in equation (2.5). Simulation studies on rank selection have shown that this criterion has a tendency to favour overly small ranks; for a recent example see Cheng and Phillips (2012). The quoted values of $\lambda_i(\pi_0)$ give a theoretical explanation for this empirical phenomenon, as the use of dimension in BIC leads to overpenalization of models that contain the true data-generating distribution but are not minimal in that regard.

In other models, determining learning coefficients can be a challenging problem, but progress has been made. For some of the examples that have been treated, we refer the reader to Aoyagi (2010a, b, 2009), Drton *et al*. (2016), Rusakov and Geiger (2005), Watanabe and Amari (2003), Watanabe and Watanabe (2007), Yamazaki and Watanabe (2003, 2004, 2005) and Zwiernik (2011). The use of techniques from computational algebra and combinatorics was emphasized in Lin (2011); see also Arnol'd *et al*. (1988) and Vasil'ev (1979).

Progress in large sample theory, however, does not readily translate into practical statistical methodology because we face the obstacle that the learning coefficients depend on the unknown data-generating distribution π_0, as indicated in our notation in equation (2.7). For instance, for the problem of selecting the rank in reduced rank regression (example 2), the Bayesian measure of model complexity that is given by the learning coefficient and its multiplicity depends on the rank that we wish to determine in the first place; recall also example 1 that is about a mixture model. It is for this reason that there is currently no statistical method that takes advantage of theoretical knowledge about the values of learning coefficients. In the remainder of this paper, we propose a solution for how to overcome the problem of circular reasoning and give a practical extension of BIC to singular models.

3. New Bayesian information criterion for singular models

3.1. Averaging approximations

As previously stated, our point of departure is the large sample result from equation (2.7). If the learning coefficient $\lambda_i(\pi_0)$ and its multiplicity $m_i(\pi_0)$ that appear in this equation were known, then we could directly adopt the ideas in Schwarz (1978), omit the remainder term in equation (2.7) and define a proxy for the marginal likelihood $L(\mathcal{M}_i)$ as

$$L'_{\pi_0}(\mathcal{M}_i) := P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\, \frac{\log(n)^{m_i(\pi_0)-1}}{n^{\lambda_i(\pi_0)}}. \tag{3.1}$$

However, in practice, $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ are unknown. We thus propose to apply standard Bayesian thinking and to average the different possible approximations $L'_{\pi_0}(\mathcal{M}_i)$ from expression (3.1) by assigning a probability measure Q_i to the distributions in model $\mathcal{M}_i$. In other words, we eliminate the unknown distribution π_0 by marginalization and compute an approximation to $L(\mathcal{M}_i)$ as

$$L'_{Q_i}(\mathcal{M}_i) := \int_{\mathcal{M}_i} L'_{\pi_0}(\mathcal{M}_i)\, \mathrm{d}Q_i(\pi_0). \tag{3.2}$$

The crux of the matter now becomes choosing an appropriate probability measure Q_i.

Remark 2. Before discussing particular choices for Q_i, we stress that any choice for Q_i reduces to Schwarz's criterion in the regular case. Here, regularity refers to the setting in which model $\mathcal{M}_i$ with parameterization $\boldsymbol{\omega}_i \mapsto \pi_i(\boldsymbol{\omega}_i)$ has a Fisher information matrix that is invertible at all $\boldsymbol{\omega}_i$ in the parameter space Ω_i. For a model with d_i parameters, it then holds that $\lambda_i(\pi_0) = d_i/2$ and $m_i(\pi_0) = 1$ for all data-generating distributions $\pi_0 \in \mathcal{M}_i$. Hence, $L'_{\pi_0}(\mathcal{M}_i) = \exp\{\mathrm{BIC}(\mathcal{M}_i)\}$ for all $\pi_0 \in \mathcal{M}_i$. As the integrand in expression (3.2) is constant we have

$$\log\{L'_{Q_i}(\mathcal{M}_i)\} = \mathrm{BIC}(\mathcal{M}_i)$$

irrespectively of the choice of Q_i.

Returning to the singular case, one possible candidate for Q_i is $P(\pi_0|\mathcal{M}_i, \mathbf{Y}_n)$: the posterior distribution in $\mathcal{M}_i$. Under this distribution, however, the singular models that are encountered in practice have the learning coefficient $\lambda_i(\pi_0)$ almost surely equal to $\dim(\mathcal{M}_i)/2$ with multiplicity $m_i(\pi_0) = 1$; recall example 2. (We assume that the set $\mathcal{M}_i$ corresponds to a subset of Euclidean space with well-defined dimension.) We obtain that

$$\log\{L'_{Q_i}(\mathcal{M}_i)\} = \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \frac{\dim(\mathcal{M}_i)}{2}\log(n),$$

which is the usual BIC, albeit with the possibility that $\dim(\mathcal{M}_i) < d_i$, where d_i is the dimension of the parameter space Ω_i when $\mathcal{M}_i$ is presented as in equation (2.2). From a pragmatic point of view, this choice of Q_i is not attractive as it merely recovers the adjustment from d_i to $\dim(\mathcal{M}_i)$ that is standard practice when applying Schwarz's BIC to singular models. More importantly, however, averaging with respect to the posterior distribution $P(\pi_0|\mathcal{M}_i, \mathbf{Y}_n)$ involves conditioning on the single model $\mathcal{M}_i$, which clearly ignores the model uncertainty that is inherent to model selection problems.

In most practical problems, the finite set of models $\{\mathcal{M}_i : i \in I\}$ has interesting structure with respect to the partial order given by inclusion. (In the examples that we consider in this paper the order is always a total order. For an example where this is not so, see Drton *et al.* (2016).) For notational convenience, we define the poset structure on the index set I and write $i \preceq j$ when $\mathcal{M}_i \subseteq \mathcal{M}_j$. Instead of conditioning on a single model, we then advocate the use of the posterior distribution

$$Q_i(\pi_0) := P(\pi_0|\{\mathcal{M} : \mathcal{M} \subseteq \mathcal{M}_i\}, \mathbf{Y}_n) = \frac{\sum_{j \preceq i} P(\pi_0|\mathcal{M}_j, \mathbf{Y}_n)\, P(\mathcal{M}_j|\mathbf{Y}_n)}{\sum_{j \preceq i} P(\mathcal{M}_j|\mathbf{Y}_n)} \tag{3.3}$$

which is obtained by conditioning on the family of all submodels of $\mathcal{M}_i$. Intuitively, the proposed choice of Q_i introduces the knowledge that the data-generating distribution π_0 is in $\mathcal{M}_i$ all the while capturing remaining posterior uncertainty with respect to submodels of $\mathcal{M}_i$. This does not yet escape from the problem of circular reasoning, since expression (3.3) involves the posterior probabilities $P(\mathcal{M}_j|\mathbf{Y}_n)$ that we are trying to approximate. However, this problem can be overcome as we argue in Section 3.2.

For simpler notation, let $L'(\mathcal{M}_i) := L'_{Q_i}(\mathcal{M}_i)$ when Q_i is chosen according to our proposal from expression (3.3). We obtain from expressions (3.2) and (3.3) that

$$L'(\mathcal{M}_i) = \frac{1}{\sum_{j \preceq i} P(\mathcal{M}_j|\mathbf{Y}_n)} \sum_{j \preceq i} L'_{ij} P(\mathcal{M}_j|\mathbf{Y}_n), \tag{3.4}$$

with

$$L'_{ij} = P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\, \Lambda_{ij}(\mathbf{Y}_n) \tag{3.5}$$

and

$$\Lambda_{ij}(\mathbf{Y}_n) = \int_{\mathcal{M}_j} \frac{\log(n)^{m_i(\pi_0)-1}}{n^{\lambda_i(\pi_0)}} \mathrm{d}P(\pi_0|\mathcal{M}_j, \mathbf{Y}_n). \tag{3.6}$$

The integral $\Lambda_{ij}(\mathbf{Y}_n)$ is the expectation of a term measuring the complexity of model $\mathcal{M}_i$ under the posterior distribution given the submodel $\mathcal{M}_j$. The integration problem in equation (3.6) may seem complicated at this point but, in fact, for the statistical problems that we have in mind, the computation of integral (3.6) is trivial because the integrand is (almost surely) constant. Indeed, all singular model selection problems that we know satisfy the following condition.

For any $i \in I$ and $j \preceq i$, there are two constants λ_{ij} and m_{ij} such that

$$\lambda_i(\pi_0) = \lambda_{ij} \quad \text{and} \quad m_i(\pi_0) = m_{ij} \tag{3.7}$$

for all π_0 in a set $A_{ij} \subseteq \mathcal{M}_j$ with $P(A_{ij}|\mathcal{M}_j, \mathbf{Y}_n) = 1$.

With such generic values for learning coefficient and multiplicity, the integral $\Lambda_{ij}(\mathbf{Y}_n)$ does not depend on the data $\mathbf{Y}_n$ and equals

$$\Lambda_{ij}(\mathbf{Y}_n) = \frac{\log(n)^{m_{ij}-1}}{n^{\lambda_{ij}}}.$$

In this case,

$$L'_{ij} = P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\, \frac{\log(n)^{m_{ij}-1}}{n^{\lambda_{ij}}} \tag{3.8}$$

becomes easy to evaluate in statistical practice.

3.1.1. Example 4

Consider again the reduced rank regression model from example 3. As mentioned, Aoyagi and Watanabe (2005) have shown that the learning coefficient $\lambda_i(\pi_0)$ and its multiplicity $m_i(\pi_0)$ depend only on the rank j that is associated with π_0. Hence, for any pair $j \leqslant i$, there are constants λ_{ij} and m_{ij} such that condition (3.7) holds for all π_0 in $\mathcal{M}_j \setminus \mathcal{M}_{j-1}$. The exceptional set $\mathcal{M}_{j-1}$ corresponds to the matrices of rank at most $j-1$ and is a null set among the matrices of rank at most j. In Table 1, we listed numerical values of λ_{ij} for the special case of $N = 5$ responses and $M = 3$ covariates.

The fact that condition (3.7) holds for reduced rank regression would also be clear if we did not know explicit formulae for the learning coefficients and their multiplicities. Consider model $\mathcal{M}_i$, and let $j \leqslant i$. Our claim is then that $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ are functions of π_0 that are constant on a set that has probability 1 under $P(\pi_0|\mathcal{M}_j, \mathbf{Y}_n)$. The pair $(\lambda_i(\pi_0), m_i(\pi_0))$ is determined by the asymptotics of a Laplace integral. Using that $\mathcal{M}_i$ is a submodel of the regular family of *all* Gaussian distributions and that $\mathcal{M}_i$ is parameterized by a polynomial map, the phase function of the Laplace integral can be taken to be a polynomial; compare for example, section 2 in Drton *et al.* (2016) or lemma 1 in Aoyagi and Watanabe (2005). Moreover, this polynomial has coefficients that are polynomial functions of π_0. When making this statement, we identify π_0 with the $N \times M$ matrix of regression coefficients. By the theory that is discussed in Watanabe (2009), if $\pi_0 = \pi_0(\boldsymbol{\omega}_j) \in \mathcal{M}_j$ then there are generic values λ_{ij} and m_{ij} such that $(\lambda_i(\pi_0), m_i(\pi_0)) \neq (\lambda_{ij}, m_{ij})$ if and only if $\boldsymbol{\omega}_j$ satisfies a polynomial equation $g_{ij}(\boldsymbol{\omega}_j) = 0$ that does not hold for all points in Ω_j. Here, $\Omega_j = \mathbb{R}^{N \times i} \times \mathbb{R}^{i \times M}$ is the parameter space of $\mathcal{M}_j$. Since g_{ij} is a non-zero polynomial, its zero set has measure zero by the lemma in Okamoto (1973). Consequently, $(\lambda_i(\pi_0), m_i(\pi_0)) = (\lambda_{ij}, m_{ij})$ holds almost surely under $P(\pi_0|\mathcal{M}_j, \mathbf{Y}_n)$.

The reasoning just given applies *verbatim* to the factor analysis model that is treated later, and to models for categorical data such as latent class models or binomial mixtures. For mixtures of Gaussians some additional insights are needed to arrive at a polynomial set-up but condition (3.7) still holds (section 7.8 in Watanabe (2009)). We note that the model from examples 1 and 2 has $\lambda_{21} = \frac{3}{4}$ and $\lambda_{22} = \frac{3}{2}$. Outside the algebraic realm, it is more difficult to make a general statement about generic values of learning coefficients. Nonetheless, we expect condition (3.7) to hold in all model selection problems of practical interest; compare also remark 1.8 in Watanabe (2009).

3.2. Singular Bayesian information criterion

Even if we can evaluate all the integrated approximations L'_{ij} for $j \preceq i$ in the generic situation from equation (3.8), our proposed approximation $L'(\mathcal{M}_i)$ remains impractical because it is a weighted average with the weights being the posterior model probabilities $P(\mathcal{M}_j|\mathbf{Y}_n)$ that we seek to approximate in the first place. To make this fact more transparent, we rewrite equation (3.4) by using that $P(\mathcal{M}_j|\mathbf{Y}_n) \propto L(\mathcal{M}_j)\, P(\mathcal{M}_j)$, which gives

$$L'(\mathcal{M}_i) = \frac{1}{\sum_{j \preceq i} L(\mathcal{M}_j)\, P(\mathcal{M}_j)} \sum_{j \preceq i} L'_{ij}\, L(\mathcal{M}_j)\, P(\mathcal{M}_j). \tag{3.9}$$

We see explicitly that $L'(\mathcal{M}_i)$, which is the supposed proxy to marginal likelihood, is a function of the actual marginal likelihood $L(\mathcal{M}_i)$ as well as the marginal likelihood $L(\mathcal{M}_j)$ of any submodel indexed by $j \prec i$. Of course, there would hardly be any interest in a proxy $L'(\mathcal{M}_i)$ once the marginal likelihood $L(\mathcal{M}_i)$ has been computed.

This said, equation (3.9) also leads to a way out of this dilemma. Observe that in equation (3.9), the marginal likelihood for model $\mathcal{M}_i$ appears twice: first in approximation on the left-hand side and then as an exact value on the right-hand side when considering summation index $j = i$. This motivates building a 'fix point equation system' by replacing each marginal likelihood $L(\mathcal{M}_j)$ on the right-hand side of equation (3.9) by its approximation $L'(\mathcal{M}_j)$. We arrive at the equation system

$$L'(\mathcal{M}_i) = \frac{1}{\sum_{j \preceq i} L'(\mathcal{M}_j)\, P(\mathcal{M}_j)} \sum_{j \preceq i} L'_{ij}\, L'(\mathcal{M}_j)\, P(\mathcal{M}_j), \qquad i \in I, \tag{3.10}$$

where the L'_{ij} and the $P(\mathcal{M}_j)$ are known constants and the desired marginal likelihood approximations $L'(\mathcal{M}_i)$ are the unknowns that we wish to solve for. We emphasize that equation (3.10) is not mathematically deduced from equation (3.9); it is simply an equation system that we heuristically motivated. Now, if we can solve the non-linear equation system in expression (3.10) and obtain a solution with all $L'(\mathcal{M}_i) > 0$ then we have computed a practical approximation to the marginal likelihood of each considered model $\mathcal{M}_i$, $i \in I$.

Our next observation is that equations (3.10) indeed have a positive solution and that this solution is unique. To show this, we clear denominators and consider the polynomial equation system

$$\sum_{j \preceq i} \{L'(\mathcal{M}_i) - L'_{ij}\} L'(\mathcal{M}_j) P(\mathcal{M}_j) = 0, \qquad i \in I. \tag{3.11}$$

Proposition 1. The equation system in expression (3.11) has a unique solution with all unknowns $L'(\mathcal{M}_i) > 0$.

Proof. Let i be any minimal element of the poset I. Then $j = i$ is the only choice for the index j, and the equation from expression (3.11) reads

$$\{L'(\mathcal{M}_i) - L'_{ii}\} L'(\mathcal{M}_i) P(\mathcal{M}_i) = 0.$$

With $P(\mathcal{M}_i) > 0$, the equation has the unique positive solution

$$L'(\mathcal{M}_i) = L'_{ii} > 0,$$

which coincides with the exponential of the usual BIC for model $\mathcal{M}_i$.

Consider now a non-minimal index $i \in I$. Proceeding by induction, assume that positive solutions $L'(\mathcal{M}_j)$ have been computed for all $j \prec i$, where $j \prec i$ if $\mathcal{M}_j \subsetneq \mathcal{M}_i$. Then $L'(\mathcal{M}_i)$ solves the quadratic equation

$$L'(\mathcal{M}_i)^2 + b_i L'(\mathcal{M}_i) - c_i = 0 \tag{3.12}$$

with

$$b_i = -L'_{ii} + \sum_{j \prec i} L'(\mathcal{M}_j) \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_i)}, \tag{3.13}$$

$$c_i = \sum_{j \prec i} L'_{ij} L'(\mathcal{M}_j) \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_i)}. \tag{3.14}$$

Since $c_i > 0$ by the induction hypothesis, equation (3.12) has the unique positive solution

$$L'(\mathcal{M}_i) = \tfrac{1}{2}\{-b_i + \sqrt{(b_i^2 + 4c_i)}\}. \tag{3.15}$$

On the basis of proposition 1, we make the following definition in which we consider the equation system from expression (3.11) under the default of a uniform prior on models, i.e. $P(\mathcal{M}_i) = 1/|I|$ for $i \in I$.

Definition 1. The singular BIC for model $\mathcal{M}_i$ is

$$\mathrm{sBIC}(\mathcal{M}_i) = \log\{L'(\mathcal{M}_i)\},$$

where $(L'(\mathcal{M}_i) : i \in I)$ is the unique solution to the equation system

$$\sum_{j \preceq i} \{L'(\mathcal{M}_i) - L'_{ij}\} L'(\mathcal{M}_j) = 0, \qquad i \in I, \tag{3.16}$$

that has all components positive.

According to expression (3.9), $\mathrm{sBIC}(\mathcal{M}_i)$ is the logarithm of a weighted average of the approximations L'_{ij}, with the weights depending on the data. As discussed in Section 2, for priors with smooth and positive densities it holds that $\lambda_i(\pi_0) \leqslant \dim(\mathcal{M}_i)/2$ and $m_i(\pi_0) \geqslant 1$ for all $\pi_0 \in \mathcal{M}_i$. Assuming that $n \geqslant 3$, this implies that

$$\frac{n^{\lambda_i(\pi_0)}}{\log(n)^{m_i(\pi_0)-1}} \leqslant n^{\dim(\mathcal{M}_i)/2}.$$

Consequently, the singular BIC is of the form

$$\mathrm{sBIC}(\mathcal{M}_i) = \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \mathrm{penalty}(\mathcal{M}_i),$$

where $\mathrm{penalty}(\mathcal{M}_i)$ is a data-dependent penalty that satisfies

$$\mathrm{penalty}(\mathcal{M}_i) \leqslant \dim(\mathcal{M}_i)/2 \log(n)$$

and thus is milder than that in the usual BIC.

Remark 3. Although we envision that the use of a uniform prior on models in definition 1 is reasonable for many applications, deviations from this default can be of interest; compare, for instance, Nobile (2005) who discussed priors for the number of components in mixture models. Via equation system (3.11), a non-uniform prior on models can be readily incorporated in the definition of the singular BIC. Later large sample results would not be affected.

Remark 4. sBIC defined by equation (3.16) is a function of the approximations L'_{ij} from equation (3.8), which in turn depend only on the maxima of the likelihood functions and the numbers λ_{ij} and m_{ij}. In our treatment so far the λ_{ij} are learning coefficients and the m_{ij} their multiplicities; recall condition (3.7). However, as we shall see for applications discussed in Section 6, interesting versions of sBIC also arise when setting the λ_{ij} and m_{ij} equal to bounds on learning coefficients and multiplicities respectively.

4. Large sample properties

As mentioned in Section 1, Schwarz's BIC with its dimension-based penalty has been shown to be consistent in many settings, including many singular model selection problems. Theorem 1 in this section asserts similar consistency for the singular BIC from definition 1. We then proceed to show that sBIC has the properties that we set out to obtain. Indeed, by proposition 3, Section 4.2, the data-dependent penalty in sBIC successfully adapts to the data-generating distribution, meaning that in large samples the penalty that sBIC assigns to a true model $\mathcal{M}_i$ agrees with the penalty that is obtained from the (in practice unknown) learning coefficient $\lambda_i(\pi_0)$ and its multiplicity $m_i(\pi_0)$. As stated in theorem 2, it follows that sBIC is indeed Bayesian in the sense that it deviates from the log-marginal-likelihood by terms that are bounded in probability.

4.1. Set-up and assumptions

We consider a finite set of models $\{\mathcal{M}_i : i \in I\}$ that is closed under intersection. Fix a data-generating distribution $\pi_0 \in \cup_{i\in I} \mathcal{M}_i$. A model $\mathcal{M}_i$ is *true* if $\pi_0 \in \mathcal{M}_i$. Otherwise, $\mathcal{M}_i$ is *false*. Since the set of models is closed under intersection, there is a unique *smallest true* model, which we denote by $\mathcal{M}_{i_0}$ for index $i_0 \in I$.

Throughout this section, we assume that Watanabe's result from equation (2.6) holds with generic learning coefficients λ_{ij} and multiplicities m_{ij} as in condition (3.7). Then sBIC($\mathcal{M}_i$) is computed from the approximations in equation (3.8), where

$$\frac{n^{\lambda_{ij}}}{\log(n)^{m_{ij}-1}}, \tag{4.1}$$

acts as a measure of complexity of model $\mathcal{M}_i$. We refer to this measure of complexity as the *(generic) Bayes complexity* of $\mathcal{M}_i$ along its submodel $\mathcal{M}_j$. Let '$\leqslant$' denote the lexicographic order on $\mathbb{R}^2$, i.e. $(x_1, y_1) \leqslant (x_2, y_2)$ if $x_1 < x_2$ or if $x_1 = x_2$ and $y_1 \leqslant y_2$. Then two Bayes complexities are ordered as

$$\frac{n^{\lambda_1}}{\log(n)^{m_1-1}} \leqslant \frac{n^{\lambda_2}}{\log(n)^{m_2-1}}$$

for all large n if and only if $(\lambda_1, -m_1) \leqslant (\lambda_2, -m_2)$.

To present a general result, we make the following assumptions about the behaviour of likelihood ratios and the learning coefficients and their multiplicities, under a fixed data-generating distribution π_0.

Assumption 1. For any two true models $\mathcal{M}_i$ and $\mathcal{M}_k$, the sequence of likelihood ratios

$$\frac{P(\mathbf{Y}_n|\hat{\pi}_k, \mathcal{M}_k)}{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)}$$

is bounded in probability as $n \to \infty$.

Assumption 2. For any pair of a true model $\mathcal{M}_i$ and a false model $\mathcal{M}_k$, there is a constant $\delta_{ik} > 0$ such that, with probability tending to 1 as $n \to \infty$, we have that

$$\frac{P(\mathbf{Y}_n|\hat{\pi}_k, \mathcal{M}_k)}{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)} \leqslant \exp(-\delta_{ik} n).$$

Assumption 3. The generic Bayes complexities are increasing with model size in the sense that, for any model indices $i, k \in I$ and submodel indices $j, l \in I$, we have that

$$(\lambda_{ij}, -m_{ij}) < (\lambda_{kj}, -m_{kj}) \qquad \text{if } j \preceq i \prec k,$$

and

$$(\lambda_{il}, -m_{il}) < (\lambda_{ij}, -m_{ij}) \qquad \text{if } l \prec j \preceq i.$$

The reader is accustomed with assumptions 1 and 2 from any treatment of consistency of Schwarz's BIC. Assumption 1, which is the more subtle of the two conditions, holds in problems that involve possibly singular submodels of exponential families and other well-behaved models. In such problems, the likelihood ratios in assumption 1 typically converge to a limiting distribution (Drton, 2009). Examples are Gaussian models such as reduced rank regression and factor analysis, but also latent class and other models for categorical data. As also mentioned when discussing equation (2.7), the sequence of likelihood ratios for mixture models is typically bounded in probability when the parameter space is compact; without compactness the sequence need not be bounded. For Gaussian mixtures, for instance, the log-likelihood ratios could be of the same $\log\{\log(n)\}$ order that the multiplicities $m_i(\pi_0)$ have an effect on (Hartigan, 1985; Bickel and Chernoff, 1993).

The first set of inequalities in assumption 3 pertains to a fixed (generic) data-generating distribution in $\mathcal{M}_j$ and makes the natural requirement that, among any two true models $\mathcal{M}_i$ and $\mathcal{M}_k$, the larger model, which is taken to be $\mathcal{M}_k$, has the larger Bayes complexity. The second set of inequalities in assumption 3 requires that the Bayes complexity of a fixed model $\mathcal{M}_i$ decreases when the data-generating distribution is moved from a generic member of a submodel $\mathcal{M}_j$ to a generic member of $\mathcal{M}_l \subsetneq \mathcal{M}_j$. Indeed, the parameters of singular models are typically 'less identifiable' at special distributions that correspond to smaller submodels, and the second set of inequalities quantifies such a property. The inequalities from assumption 3 hold in all the aforementioned examples for which learning coefficients have been computed; in particular, the assumption holds for the applications that we shall treat later including reduced rank regression from example 2.

4.2. Consistency

Our first result clarifies that the singular BIC selects the smallest true model in the large sample limit. We emphasize that we fix a data-generating distribution π_0 and then consider large sample limits.

Theorem 1. Let $\mathcal{M}_{i_0}$ be the smallest true model, and let $\mathcal{M}_{\hat{\imath}}$ be the model selected by maximizing the singular BIC, i.e.

$$\hat{\imath} = \arg\max_{i \in I} \mathrm{sBIC}(\mathcal{M}_i).$$

Under assumptions 1–3, the probability that $\hat{\imath} = i_0$ tends to 1 as $n \to \infty$.

Remark 5. The consistency result in theorem 1 does not rely on the λ_{ij} being learning coefficients. Indeed, consistency holds for any version of sBIC that is based on numbers λ_{ij} and m_{ij} that satisfy assumption 3. We shall explore this in the applications in Section 6, where λ_{ij} and m_{ij} will be bounds on learning coefficients and their multiplicities respectively; recall also remark 4.

Since we are concerned with a finite set of models $\{\mathcal{M}_i : i \in I\}$, the consistency result in theorem 1 can be established by pairwise comparisons. More precisely, it suffices to show that

(a) the singular BIC of any true model is asymptotically larger than that of any false model, and
(b) the singular BIC of a true model can be asymptotically maximal only if the model is the smallest true model.

The comparisons (a) and (b) are addressed in propositions 2 and 3 respectively. Throughout, $(L'(\mathcal{M}_i) : i \in I)$ refers to the unique positive solution of equations (3.16), i.e. $\log\{L'(\mathcal{M}_i)\} = \mathrm{sBIC}(\mathcal{M}_i)$.

Proposition 2. Under assumption 2, if model $\mathcal{M}_i$ is true and model $\mathcal{M}_k$ is false, then the probability that $\mathrm{sBIC}(\mathcal{M}_i) > \mathrm{sBIC}(\mathcal{M}_k)$ tends to 1 as $n \to \infty$.

Proof. Fix an index $j \preceq i$ and a second index $l \preceq k$. Since $\mathcal{M}_k$ is false, assumption 2 implies that the ratio L'_{kl}/L'_{ij} converges to 0 in probability as $n \to \infty$ i.e. $L'_{kl} = o_p(L'_{ij})$. Since j was arbitrary, $L'_{kl} = o_p(L'_{i\,\min})$, where

$$L'_{i\,\min} = \min\{L'_{ij} : j \preceq i\};$$

note that for fixed i and varying j the approximations L'_{ij} share the likelihood term and differ only in the learning coefficients or their multiplicities.

According to expression (3.10), $L'(\mathcal{M}_k)$ is a weighted average of the terms L'_{kl} with $l \preceq k$. We obtain that

$$L'(\mathcal{M}_k) \leqslant \max\{L'_{kl} : l \preceq k\} = o_p(L'_{i\,\min}). \tag{4.2}$$

Similarly, $L'(\mathcal{M}_i)$ is a weighted average of the L'_{ij}, $j \preceq i$, and it thus holds that

$$L'(\mathcal{M}_i) \geqslant L'_{i\,\min} > 0. \tag{4.3}$$

We conclude that

$$L'(\mathcal{M}_k) = o_p\{L'(\mathcal{M}_i)\}. \tag{4.4}$$

It follows that $L'(\mathcal{M}_i) > L'(\mathcal{M}_k)$ with probability tending to 1 as $n \to \infty$, which yields the claim because $\mathrm{sBIC}(\mathcal{M}_i) = \log\{L'(\mathcal{M}_i)\}$.

Proposition 3. Let $\mathcal{M}_i$ be a true model. Then, under assumptions 1–3,

$$\mathrm{sBIC}(\mathcal{M}_i) = \log(L'_{ii_0}) + o_p(1),$$

and thus for all $i \succ i_0$, with probability tending to 1 as $n \to \infty$,

$$\mathrm{sBIC}(\mathcal{M}_i) < \mathrm{sBIC}(\mathcal{M}_{i_0}).$$

Proof. First note that under assumption 3 the second assertion is a straightforward consequence of the first; compare expression (4.8) below. By exponentiating, the first assertion is seen to be equivalent to

$$L'(\mathcal{M}_i) = L'_{ii_0}\{1 + o_p(1)\}, \qquad i \succeq i_0.$$

We shall argue by induction on i.

To establish the base for the induction, consider the smallest true model, i.e. $i = i_0$. Let $j \prec i_0$. Then we know from equation (4.4) that $L'(\mathcal{M}_j) = o_p\{L'(\mathcal{M}_{i_0})\}$. Using the exponentially fast decay of the ratio in assumption 2, the arguments in the proof of proposition 2 also yield that $L'(\mathcal{M}_j) f(n) = o_p\{L'(\mathcal{M}_{i_0})\}$ for any polynomial $f(n)$. Since $L'_{i_0 j}/L'_{i_0\,\min}$ is a deterministic function that grows at most polynomially with n, and since $L'_{i_0\,\min} \leqslant L'(\mathcal{M}_{i_0})$ according to inequality (4.3), we have

$$L'_{i_0 j} L'(\mathcal{M}_j) = o_p\{L'(\mathcal{M}_{i_0})^2\}. \tag{4.5}$$

Applying these observations to the coefficients b_{i_0} and c_{i_0} from expressions (3.13) and (3.14), we obtain that $c_{i_0} = o_p\{L'(\mathcal{M}_{i_0})^2\}$ and $b_{i_0} + L'_{i_0 i_0} = o_p\{L'(\mathcal{M}_{i_0})\}$. From the quadratic equation defining $L'(\mathcal{M}_{i_0})$, we deduce that

$$L'(\mathcal{M}_{i_0})^2 - L'_{i_0 i_0} L'(\mathcal{M}_{i_0}) = o_p\{L'(\mathcal{M}_{i_0})^2\}. \tag{4.6}$$

Hence, the equation's positive solution satisfies our claim, namely

$$L'(\mathcal{M}_{i_0}) = L'_{i_0 i_0}\{1 + o_p(1)\}. \tag{4.7}$$

For the induction step, assume that the claim is true for proper submodels of $\mathcal{M}_i$, i.e.

$$L'(\mathcal{M}_k) = L'_{ki_0}\{1 + o_p(1)\}, \qquad i_0 \preceq k \prec i.$$

Further note that arguing similarly as for $i = i_0$, the contributions of false models to the coeffi-

cients b_i and c_i from expressions (3.13) and (3.14) are seen to negligible. We thus have

$$b_i = -L'_{ii} + \left\{ \sum_{i_0 \preceq j \prec i} L'(\mathcal{M}_j) \right\} \{1 + o_p(1)\} = -L'_{ii} + \left(\sum_{i_0 \preceq j \prec i} L'_{ji_0} \right) \{1 + o_p(1)\}$$

and

$$c_i = \left\{ \sum_{i_0 \preceq j \prec i} L'_{ij} L'(\mathcal{M}_j) \right\} \{1 + o_p(1)\} = \left(\sum_{i_0 \preceq j \prec i} L'_{ij} L'_{ji_0} \right) \{1 + o_p(1)\}.$$

By assumptions 1 and 3,

$$L'_{ki_0} = o_p(L'_{i_0 i_0}), \qquad i_0 \preceq k \prec i, \tag{4.8}$$

and also

$$L'_{ij} = o_p(L'_{ii_0}), \qquad i_0 \preceq j \preceq i.$$

We obtain that

$$b_i = -L'_{ii} + L'_{i_0 i_0} \{1 + o_p(1)\} = L'_{i_0 i_0} \{1 + o_p(1)\}$$

and

$$c_i = L'_{ii_0} L'_{i_0 i_0} \{1 + o_p(1)\}.$$

Consequently,

$$\begin{aligned} L'(\mathcal{M}_i) &= \tfrac{1}{2}\{-b_i + \sqrt{(b_i^2 + 4c_i)}\} \\ &= \tfrac{1}{2}\{-L'_{i_0 i_0} + \sqrt{({L'_{i_0 i_0}}^2 + 4L'_{ii_0} L'_{i_0 i_0})}\}\{1 + o_p(1)\} \\ &= \tfrac{1}{2}[-L'_{i_0 i_0} + \sqrt{\{{L'_{i_0 i_0}}^2 + 4L'_{ii_0} L'_{i_0 i_0} + (2L'_{ii_0})^2\}}]\{1 + o_p(1)\}, \end{aligned}$$

where the last equality follows from $L'_{ii_0} = o_p(L'_{i_0 i_0})$. However, this is what was to be shown because

$$\tfrac{1}{2}[-L'_{i_0 i_0} + \sqrt{\{(L'_{i_0 i_0})^2 + 4L'_{ii_0} L'_{i_0 i_0} + (2L'_{ii_0})^2\}}] = L'_{ii_0}.$$

Remark 6. Although we do not pursue this here, it would be interesting to establish further consistency properties for sBIC. For instance, one could seek to adapt the results in Gassiat and van Handel (2013) to give strong consistency results for sBIC. Gassiat and van Handel (2013) considered general information criteria for order selection, i.e. for problems in which the set of models is totally ordered by inclusion (as in mixture modelling or factor analysis). No upper bound on the number of such models was assumed in their work.

4.3. Connection to marginal likelihood

Under assumption 2, the marginal likelihood of a false model is with high probability exponentially smaller than that of any true model. The frequentist large sample behaviour of Bayesian model selection procedures is thus primarily dictated by the asymptotics of the marginal likelihood integrals of true models.

As pointed out in Section 3, the usual BIC from equation (2.5) with penalty depending solely on model dimension generally does not reflect the asymptotic behaviour of the marginal

likelihood of a true model that is singular, which is given by equation (3.1). Consequently, as the sample size increases, the Bayes factor that is obtained by forming the ratio of the marginal likelihood integrals for two true models may increase or decrease at a rate that is different from the rate for an approximate Bayes factor formed by exponentiating the difference of the two respective BIC-scores. Hence, there is generally nothing Bayesian about the usual BIC in singular model selection problems. In contrast, the new singular BIC is connected to the large sample behaviour of the log-marginal likelihood.

Theorem 2. Let $\mathcal{M}_i$ be a true model, let $\mathcal{M}_{i_0}$ be the smallest true model and let π_0 be a generic distribution in $\mathcal{M}_{i_0}$. Then, under assumptions 1–3, the marginal likelihood of $\mathcal{M}_i$ satisfies

$$\log\{L(\mathcal{M}_i)\} = \mathrm{sBIC}(\mathcal{M}_i) + O_p(1).$$

Proof. By proposition 3 and equation (3.8),

$$\begin{aligned}\mathrm{sBIC}(\mathcal{M}_i) &= \log(L'_{ii_0}) + o_p(1)\\ &= \log\{P(\mathbf{Y}_n|\hat{\pi}_i, \mathcal{M}_i)\} - \lambda_{ii_0}\log(n) + (m_{ii_0} - 1)\log\{\log(n)\} + o_p(1).\end{aligned}$$

By condition (3.7),

$$\begin{aligned}\lambda_i(\pi_0) &= \lambda_{ii_0},\\ m_i(\pi_0) &= m_{ii_0}.\end{aligned}$$

The claim thus follows from equation (2.7), which in turn follows from Watanabe's result (2.6) and assumption 1.

5. Applications in multivariate analysis

We apply sBIC to two singular model selection problems arising in multivariate analysis. First, we consider the problem of selecting the rank of the matrix of regression coefficients in reduced rank regression and perform a simulation study that illustrates consistency properties. Second, we treat the problem of selecting the number of factors in factor analysis and work with a well-known data set to show how sBIC can lead to an improved assessment of model uncertainty. For a third application of sBIC in multivariate analysis, we point the reader to Drton *et al.* (2016) who treat Gaussian latent forest models with similar findings to those for the examples that we report on here.

5.1. Rank selection

We take up the setting of reduced rank regression from example 3 and Aoyagi and Watanabe (2005). We consider a scenario with $N = 10$ responses and $M = 15$ covariates. We randomly generate an $N \times M$ matrix of regression coefficients π of fixed rank 5. More precisely, we fix the signal strength by fixing the non-zero singular values of π to be 1.2, 1.0, 0.8, 0.6 and 0.4. The matrix π is then obtained by drawing the left and the right singular vectors according to the Haar measures on the two relevant Stiefel manifolds. Given π, we generate n IID normal random vectors according to the reduced rank regression model, as specified in example 3. Rank estimates are then obtained by maximizing Schwarz's BIC or the new sBIC. For each value of n, we run 200 simulations with varying π.

In our simulations, we also consider the 'widely applicable Bayesian information criterion' WBIC of Watanabe (2013). The point of departure in the derivation of this criterion is the fact

that the marginal likelihood can be computed by thermodynamic integration; see also Friel and Pettitt (2008). Watanabe then analyses the large sample properties of the mean value obtained by applying the mean value theorem to the thermodynamic integral. The analysis shows that, for many models and sufficiently large sample size n, the temperature at which the mean value arises can be approximated by $\log(n)$. We computed WBIC for reduced rank regression by using a Metropolis–Hastings sampler for which we adapt the computer code that is available on Sumio Watanabe's Web site `http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/wbic2012e.html`.

We would like to stress that WBIC is not a direct competitor to our sBIC. WBIC does not use or require knowledge of the learning coefficients, and its computation involves integration as opposed to the maximization in sBIC. Another important difference is that WBIC involves an explicit choice of a prior on model parameters, whereas sBIC depends on the prior only through learning coefficients. The prior distribution in the code that we use for WBIC has the entries of the two matrices ω_{i1} and ω_{i2} IID normal with mean 0 and standard deviation 10. We tuned the standard deviations for the normal distributions used for proposals in a random walk to 0.015. Running the sampler for 10 000 steps after 1 000 steps of burn-in gave average acceptance rates that remained in the range from 0.1 to 0.9.

The results of the simulations are shown in Fig. 2, in which the new sBIC is seen to have good rank selection properties in finite samples. For instance, for a sample size of $n = 300$, sBIC identifies the true rank 5 in the vast majority of cases whereas the usual BIC selects a rank of 3 or 4 in virtually all cases. At $n = 1000$, BIC and sBIC are perfect, with the exception of two cases in which sBIC selects rank 6 and two cases in which BIC selects rank 4. The behaviour of

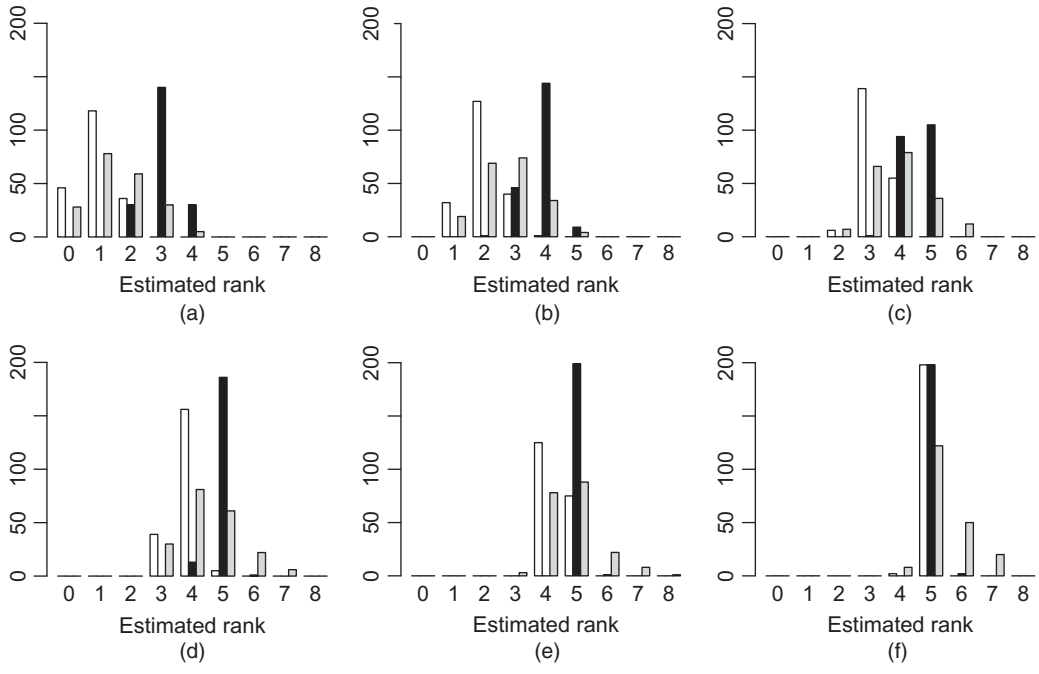

Fig. 2. Frequencies of rank estimates in reduced rank regression by using Schwarz's BIC (□), WBIC (▒) and sBIC (■) (results from 200 simulations with 10 × 15 matrices of true rank 5): (a) $n = 50$; (b) $n = 100$; (c) $n = 200$; (d) $n = 300$; (e) $n = 500$; (f) $n = 1000$

the implemented version of WBIC is somewhat different with the ranks selected having greater variance.

Our main conclusion is that sBIC yields an improvement over the standard dimension-based BIC in terms of frequentist rank selection properties. In this simulation study, sBIC also performs well compared with WBIC but the rank selection properties of WBIC could certainly be improved by tuning the prior distributions involved to the problem at hand, as opposed to employing the defaults from the computer code that we applied. Our conclusion from the comparison to WBIC is simply that sBIC can achieve state of the art performance in rank selection.

5.2. Factor analysis

Lopes and West (2004), section 6.3, fitted factor analysis models to data $\mathbf{Y}_n$ concerning changes in the exchange rates of six currencies relative to the British pound. The sample size is $n = 143$. We write $\mathcal{M}_i$ for the factor analysis model with i factors, which in this example comprises multivariate normal distributions for a random vector taking values in $\mathbb{R}^6$. The distributions in $\mathcal{M}_i$ have an arbitrary mean vector but their covariance matrix is constrained to be of the form $\boldsymbol{\Sigma} + \boldsymbol{\beta}\boldsymbol{\beta}'$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with positive entries and $\boldsymbol{\beta}$ is a real $6 \times i$ matrix. This particular covariance structure arises from conditional independence of the six observed random variables given i latent factors.

Lopes and West (2004) restricted the number of factors i to at most 3, so as not to overparameterize the 6×6 covariance matrix. Their Tables 3 and 5 report the following two sets of posterior model probabilities obtained from Markov chain Monte Carlo computation:

$$P(\mathcal{M}_1|\mathbf{Y}_n) = 0.00, \qquad P(\mathcal{M}_2|\mathbf{Y}_n) = 0.88, \qquad P(\mathcal{M}_3|\mathbf{Y}_n) = 0.12 \tag{5.1}$$

and

$$P(\mathcal{M}_1|\mathbf{Y}_n) = 0.00, \qquad P(\mathcal{M}_2|\mathbf{Y}_n) = 0.98, \qquad P(\mathcal{M}_3|\mathbf{Y}_n) = 0.02. \tag{5.2}$$

They are based on slightly different priors for the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\beta})$ of each model. Both types of prior have all parameters independent and use inverse gamma distributions for the diagonal entries of $\boldsymbol{\Sigma}$. The entries of $\boldsymbol{\beta}$ are IID normal, but in doing so different identifiability constraints are used for result (5.1) *versus* result (5.2). The detailed specification of the prior is given in sections 2.3 and 6.3 of Lopes and West (2004).

We consider these same data and compute Schwarz's BIC as well as our singular BIC. We find it natural to consider also the model $\mathcal{M}_0$ that postulates independence of the six changes in exchange rates considered. On the basis of on-going work of the first author and collaborators, we use the learning coefficients λ_{ij} for sBIC in Table 2, with all multiplicities $m_{ij} = 1$. These learning coefficients do not include the contribution of $6/2 = 3$ from the means of the six variables. Note that the 'top coefficient' λ_{ii} equals the dimension of the set of covariance matrices in model $\mathcal{M}_i$; for a computation of this dimension see, for example, theorem 2 in Drton *et al.* (2007).

Exponentiating and renormalizing either set of BIC-scores, we obtain the following approximate posterior model probabilities:

	$P(\mathcal{M}_0\|\mathbf{Y}_n)$	$P(\mathcal{M}_1\|\mathbf{Y}_n)$	$P(\mathcal{M}_2\|\mathbf{Y}_n)$	$P(\mathcal{M}_3\|\mathbf{Y}_n)$
BIC	0.0000	0.0000	0.9999	0.0001
sBIC	0.0000	0.0000	0.9797	0.0203.

(5.3)

Table 2. Learning coefficients λ_{ij} for sBIC

i	*Coefficients for the following values of j:*			
	$j=0$	$j=1$	$j=2$	$j=3$
0	3			
1	9/2	6		
2	6	29/4	17/2	
3	15/2	17/2	19/2	21/2

Comparing result (5.3) with results (5.1) and (5.2), we see that the approximation that is given by sBIC gives results that are closer to the Monte Carlo approximations than those from the standard BIC which leads to overconfidence in model $\mathcal{M}_2$. Of course, this assessment is necessarily subjective as it pertains to a comparison with two particular priors $P(\pi_i|\mathcal{M}_i)$ in each model.

To explore the connection between the information criteria and fully Bayesian procedures further, we subsampled the exchange rate data considered to create 10 data sets for each sample size $n \in \{25, 50, 75, 100\}$. For each data set we ran the Markov chain Monte Carlo algorithms of Lopes and West (2004), focusing on the prior underlying result (5.2). In Fig. 3 we present boxplots of the four posterior model probabilities. When comparing the spread in the approximate posterior probabilities, sBIC gives a far better agreement with the fully Bayesian procedure than does the standard BIC.

For the data considered, the model uncertainty mostly concerns the decision between two and three factors and can be summarized by the Bayes factor for this model comparison. In Fig. 4, we plot the log-Bayes-factors that were obtained from the Markov chain Monte Carlo procedure against those computed via the information criteria. The results from sBIC are seen to be quite close to Bayesian; the filled points in the scatter plot cluster around the 45° line. The plot also illustrates one more time that BIC is overly certain about the number of factors being 2.

6. Applications in mixture modelling

We now apply sBIC to select the number of mixture components for finite mixture models, which is a problem where the standard dimension-based BIC has a tendancy to underselect the number of components (Charnigo and Pilla (2007), section 4.2). Determining the learning coefficients for mixture models can be a complicated problem but it is possible to give simple and general bounds, and we demonstrate that these bounds yield useful versions of sBIC (recall remark 4). We begin with simulations for mixtures of binomial distributions. Next, we fit Gaussian mixture models to the familiar galaxies data (e.g. Roeder and Wasserman (1997)) to illustrate that sBIC allows for more posterior mass to be assigned to larger models, which seems more in line with fully Bayesian procedures for model determination. Finally, we present simulations for latent class analysis, which involves mixture models with multiparameter component distributions. In this setting, the values of the learning coefficients depend in important ways on the choice of prior distributions, which can have substantial influence on the model selection behaviour of sBIC.

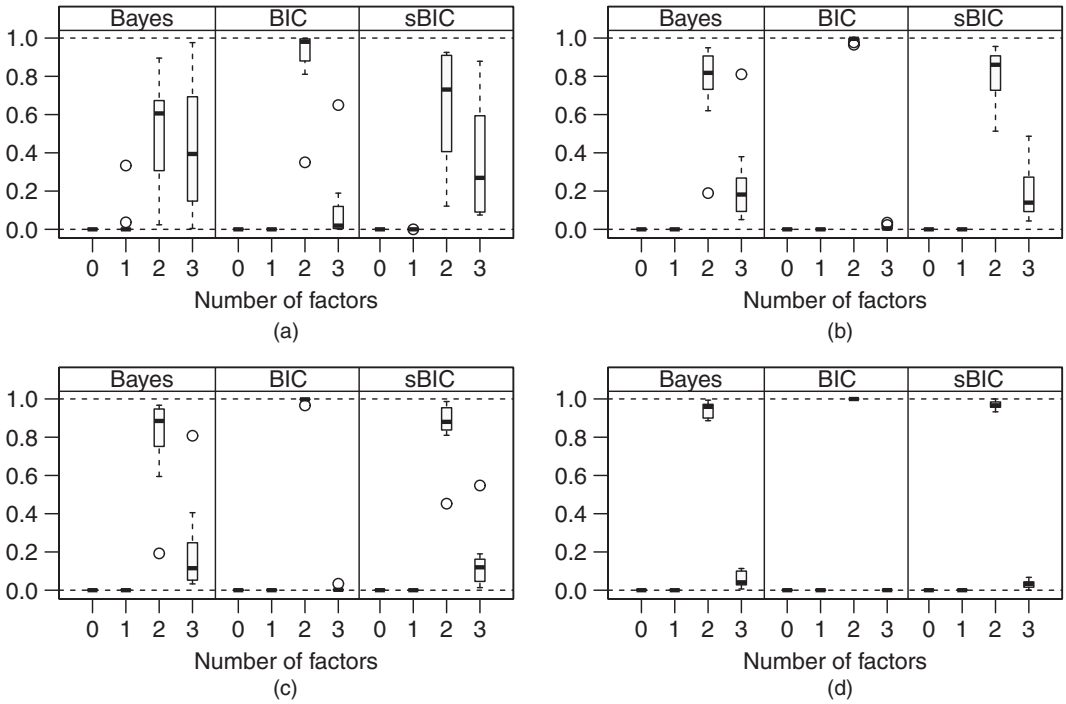

Fig. 3. Boxplots of posterior model probabilities in a factor analysis of exchange rate data under subsampling to size (a) $n = 25$, (b) $n = 50$, (c) $n = 75$ and (d) $n = 100$; results from a Markov chain Monte Carlo algorithm ('Bayes'), Schwarz's BIC and the new sBIC

6.1. Binomial mixtures

Suppose that $Y_{n1}, \ldots, Y_{nn}$ are IID counts whose distribution π is modelled as a mixture of binomial distributions. We write $\mathcal{B}(k, \theta)$ for the binomial distribution with sample size parameter k and success probability $\theta \in [0, 1]$. To match previously used notation, let i denote the number of mixture components, and let model $\mathcal{M}_i$ comprise the distributions

$$\pi_i(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{h=1}^{i} \alpha_h \mathcal{B}(k, \theta_h),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_i)$ is a vector of unknown non-negative mixture weights that sum to 1, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_i) \in [0, 1]^i$ is a vector of unknown success probabilities. We assume that the binomial sample size parameter k is known. Throughout this subsection, we assume that each prior distribution $P(\boldsymbol{\alpha}, \boldsymbol{\theta} | \mathcal{M}_i)$ has a density that is bounded away from zero on $\Delta_{i-1} \times [0, 1]^i$.

Consider now a data-generating distribution $\pi_0 \in \mathcal{M}_i$. The fibre of π_0 under the parameterization of $\mathcal{M}_i$ is the preimage

$$\mathcal{F}_i(\pi_0) = \{(\boldsymbol{\alpha}, \boldsymbol{\theta}) \in \Delta_{i-1} \times [0, 1]^i : \pi_i(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \pi_0\}, \tag{6.1}$$

containing all parameter vectors $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ that define the same distribution π_0. Here, Δ_{i-1} denotes the $(i-1)$-dimensional probability simplex. Clearly, if $(\boldsymbol{\alpha}, \boldsymbol{\theta}) \in \mathcal{F}_i(\pi_0)$ then $\mathcal{F}_i(\pi_0)$ also contains any vector that is obtained by permuting the entries of $\boldsymbol{\theta}$ and, accordingly, those of $\boldsymbol{\alpha}$. When i is not too large with respect to k, specifically, if $2i - 1 \leqslant k$, then the fibre of a distribution

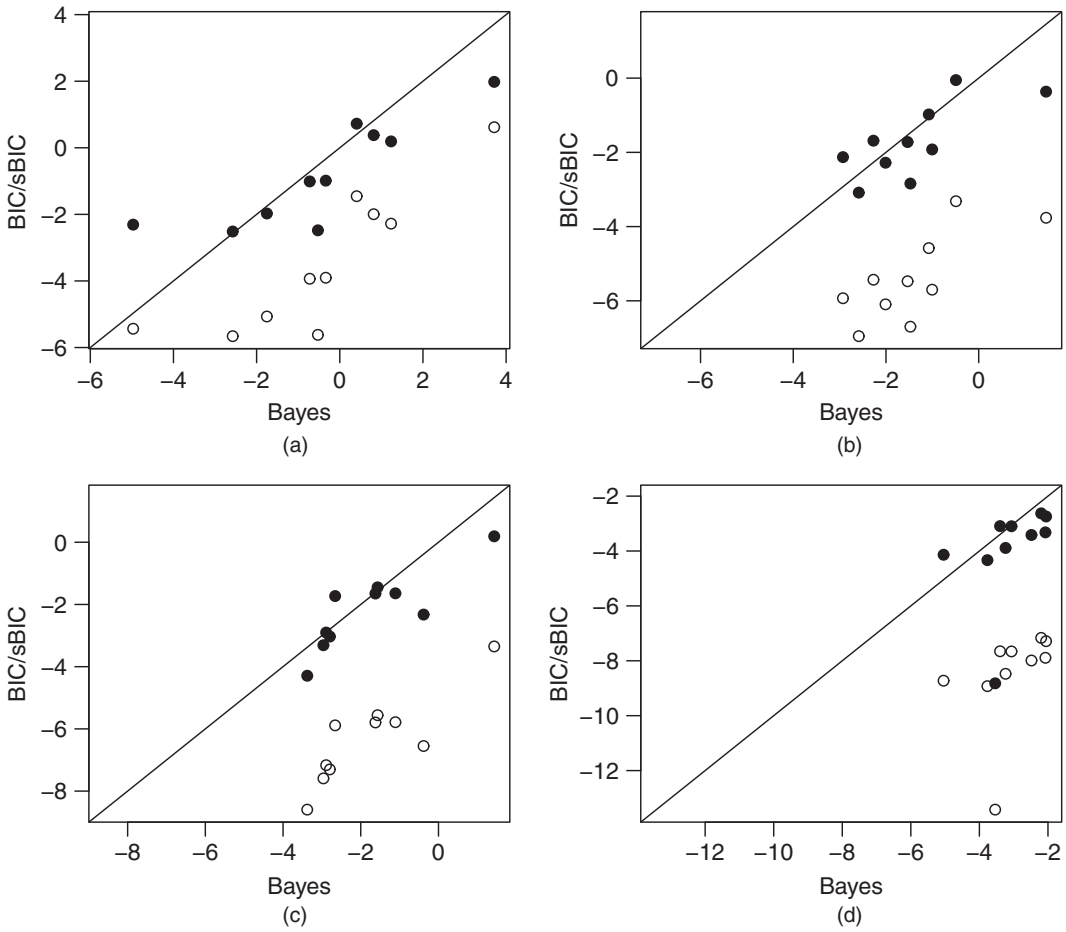

Fig. 4. Scatter plot of log-Bayes factors comparing the results of a Markov chain Monte Carlo algorithm with BIC (∘) and sBIC (•) in a factor analysis of exchange rate data under subsampling to size (a) $n = 25$, (b) $n = 50$, (c) $n = 75$ and (d) $n = 100$

$\pi_0 = \pi_i(\boldsymbol{\alpha}, \boldsymbol{\theta})$ in $\mathcal{M}_i \setminus \mathcal{M}_{i-1}$ contains only the $i!$ points in the orbit of $(\boldsymbol{\alpha}, \boldsymbol{\theta})$, and

$$\dim(\mathcal{M}_i) = 2i - 1; \tag{6.2}$$

see proposition 4 in Teicher (1963) or also section 3.1 in Titterington *et al*. (1985).

When the number of mixture components i is larger than needed, however, a severe non-identifiability problem arises. This is discussed, for instance, in section 1.3 of Frühwirth-Schnatter (2006). We provide some details that form the basis for bounds on learning coefficients that we shall consider for a definition of sBIC.

Proposition 4. Suppose that $2i - 1 \leqslant k$ and consider a binomial mixture $\pi_0 \in \mathcal{M}_j \setminus \mathcal{M}_{j-1}$ for $j < i$. Then the fibre $\mathcal{F}_i(\pi_0)$ from equation (6.1) is the intersection of $\Delta_{i-1} \times [0, 1]^i$ with a finite union of $(i - j)$-dimensional affine spaces. In particular, $\mathcal{F}_i(\pi_0)$ has dimension $i - j$.

Proof. Since $\pi_0 \in \mathcal{M}_j \setminus \mathcal{M}_{j-1}$, we have

$$\pi_0 = \sum_{h=1}^{j} \alpha_{0h}\, \mathcal{B}(k, \theta_{0h}), \tag{6.3}$$

where the success probabilities $\theta_{01}, \ldots, \theta_{0j}$ are pairwise disjoint and the mixture weights $\alpha_{01}, \ldots, \alpha_{0j}$ are positive. The probabilities θ_{0h} and α_{0h} in equation (6.3) are unique up to permutation.

We may represent π_0 as an element of $\mathcal{M}_i$ by setting $i-j$ of the mixture weights to 0, in which case $i-j$ of the success probabilities can be chosen arbitrarily. More precisely, if we take

$$\boldsymbol{\alpha} = (\alpha_{01}, \ldots, \alpha_{0j}, 0, \ldots, 0) \in \Delta_{i-1},$$

then $(\boldsymbol{\alpha}, \boldsymbol{\theta}) \in \mathcal{F}_i(\pi_0)$ for any vector $\boldsymbol{\theta}$ with $\theta_h = \theta_{0h}$ for $1 \leqslant h \leqslant j$. Hence, the fibre $\mathcal{F}_i(\pi_0)$ contains the $(i-j)$-dimensional set

$$\{(\alpha_{01}, \ldots, \alpha_{0j}, 0, \ldots, 0)\} \times \{(\theta_{01}, \ldots, \theta_{0j})\} \times [0,1]^{i-j} \tag{6.4}$$

and its orbit under the action of the symmetric group.

A second way to represent π_0 as an element of $\mathcal{M}_i$ is to choose a vector $\boldsymbol{\theta} \in [0,1]^i$ that has precisely j distinct entries, the distinct values being $\theta_{01}, \ldots, \theta_{0j}$. For each index $h \in \{1, \ldots, j\}$, let J_h be the set of indices $l \in \{1, \ldots, i\}$ such that $\theta_l = \theta_{0h}$. Then $J_1, \ldots, J_j$ form a partition of $\{1, \ldots, i\}$. For instance, if

$$\boldsymbol{\theta} = (\theta_{01}, \theta_{02}, \ldots, \theta_{0(j-1)}, \theta_{0j}, \ldots, \theta_{0j}) \in [0,1]^i, \tag{6.5}$$

then $J_h = \{h\}$ for all $h < j$ and $J_j = \{j, \ldots, i\}$. For $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ to be in the fibre $\mathcal{F}_i(\pi_0)$, it needs to hold that

$$\sum_{l \in J_h} \alpha_l = \alpha_{0h}, \qquad h = 1, \ldots, j.$$

Clearly, there are now $i-j$ degrees of freedom in the choice of the mixture weights. For instance, the fibre $\mathcal{F}_i(\pi_0)$ contains the $(i-j)$-dimensional set

$$\{(\alpha_{01}, \ldots, \alpha_{0,j-1})\} \times (\alpha_{0j} \Delta_{i-j}) \times \{(\theta_{01}, \ldots, \theta_{0j}, \theta_{0j}, \ldots, \theta_{0j})\} \tag{6.6}$$

and its orbit under the action of the symmetric group.

When $i = 2$ and $\theta_0 = \mathcal{B}(k, \frac{2}{3})$ with $k \geqslant 3$, then the fibre $\mathcal{F}_i(\pi_0)$ is a union of three line segments. This fibre is plotted in Fig. 5: the two grey lines intersect the boundary of the probability simplex Δ_1, i.e. they have $\alpha = \alpha_1 = 0$ or $1 - \alpha = \alpha_2 = 0$.

By proposition 4, the fibre $\mathcal{F}_i(\pi_0)$ of a generic distribution $\pi_0 \in \mathcal{M}_j$, $j \leqslant i$, is a set of dimension $i-j$. The learning coefficient $\lambda_i(\pi_0)$ for model $\mathcal{M}_i$ depends only on j and can be bounded by subtracting the dimension of the fibre from the model dimension; see section 7.3 in Watanabe (2009). Writing $\lambda_{ij} = \lambda_i(\pi_0)$, we find that

$$\lambda_{ij} \leqslant \bar{\lambda}_{ij}^{1} := \tfrac{1}{2}\{\dim(\mathcal{M}_i) - (i-j)\} = \tfrac{1}{2}\{2j - 1 + (i-j)\} = \tfrac{1}{2}(i+j-1). \tag{6.7}$$

Now it is known that the actual learning coefficient for binomial mixture models ($i \geqslant 2$) is smaller than the $\bar{\lambda}_{ij}^{1}$ from expression (6.7). Indeed, for a prior density that is bounded away from zero on $\Delta_{i-1} \times [0,1]^i$, Yamazaki and Watanabe (2004) have shown that

$$\lambda_{i,i-1} = i - \tfrac{5}{4},$$

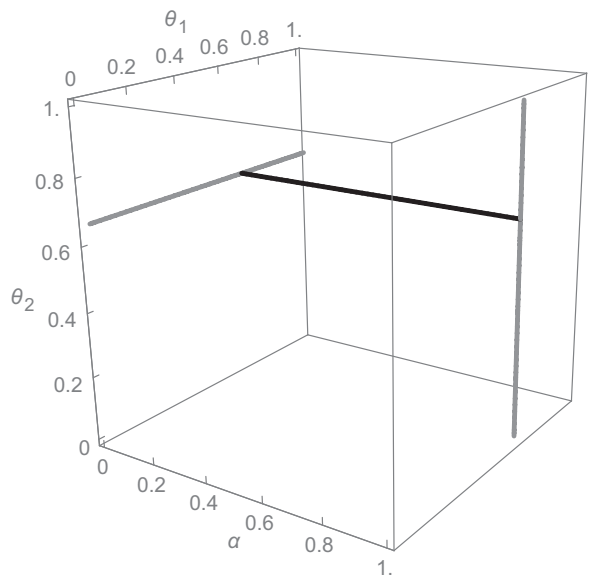

Fig. 5. Fibre of a binomial distribution in the model that mixes two binomial distributions

whereas $\bar{\lambda}^{1}_{i,i-1} = i - 1$. The model dimension from equation (6.2) yields the looser bound $i - \frac{1}{2}$. Although no general formulae for the coefficients λ_{ij} have been obtained thus far, the analysis of Rousseau and Mengersen (2011), equations (5) and (6), yields the tighter bound

$$\lambda_{ij} \leqslant \bar{\lambda}^{0.5}_{ij} := \tfrac{1}{2}\{2j - 1 + \frac{1}{2}(i - j)\} = \frac{i + 3j}{4} - \frac{1}{2}. \tag{6.8}$$

For $j = i - 1$, we have $\bar{\lambda}^{0.5}_{ij} = \lambda_{ij} = i - \frac{5}{4}$ but we do not expect this to be true in general.

In light of the above discussion, we argue that using the bounds $\bar{\lambda}^{0.5}_{ij}$ from expression (6.8) or even the very easily derived bound $\bar{\lambda}^{1}_{ij}$ from expression (6.7) is more appropriate for the definition of a BIC than merely working with the model dimension from expression (6.2). For a numerical experiment, we generate data from a distribution π_0 that is a mixture of four (but not fewer) binomial distributions that each have sample size parameter $k = 30$. Specifically, we consider the mixture weights

$$\alpha_{01} = \tfrac{1}{4}, \qquad \alpha_{02} = \tfrac{1}{4}, \qquad \alpha_{03} = \tfrac{1}{4}, \qquad \alpha_{04} = \tfrac{1}{4}$$

and the success probabilities

$$\theta_{01} = \tfrac{1}{5}, \qquad \theta_{02} = \tfrac{2}{5}, \qquad \theta_{03} = \tfrac{3}{5}, \qquad \theta_{04} = \tfrac{4}{5}.$$

For varying values n, we generate an IID sample of size n from π_0 and select the number of mixture components by maximizing

(a) Schwarz's BIC which uses the model dimension $2i - 1$,
(b) $\overline{\text{sBIC}}_{0.5}$, by which we mean the singular BIC computed by using the bounds $\bar{\lambda}^{0.5}_{ij}$ from expression (6.8), and
(c) $\overline{\text{sBIC}}_{1}$, which stands for the singular BIC computed by using the $\bar{\lambda}^{1}_{ij}$ from expression (6.7).

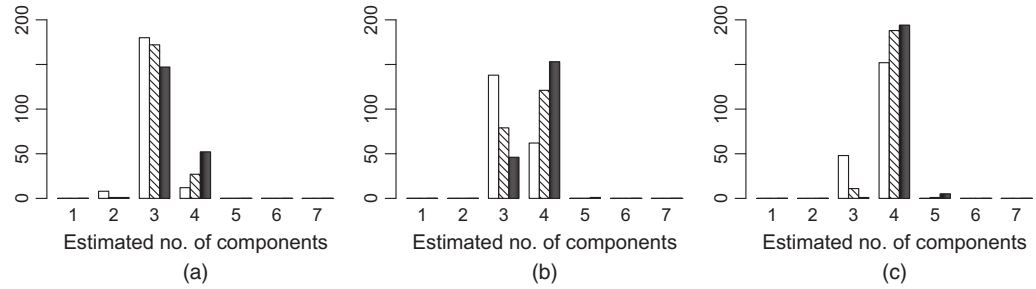

Fig. 6. Frequencies of estimated number of binomial mixture components using Schwarz's BIC (□), $\overline{\text{sBIC}}_{0.5}$ (■) and $\overline{\text{sBIC}}_{1}$ (▨) (results from 200 simulations with sample size parameter $k = 30$ and true number of components equal to 4): (a) $n = 50$, (b) $n = 200$, (c) $n = 500$

Both $\overline{\text{sBIC}}_{0.5}$ and $\overline{\text{sBIC}}_{1}$ have all multiplicities m_{ij} set to their lower bound 1. We repeat the model selection 200 times.

The frequencies of how often a particular number of components was selected by each method are depicted in Fig. 6, where we show plots for $n = 50, 200, 500$. The results are similar to those in the rank selection experiment from Section 5.1 in that our singular BIC allows us to identify the true number of components earlier than BIC. Both $\overline{\text{sBIC}}_{0.5}$ and $\overline{\text{sBIC}}_{1}$ alleviate some of the overpenalization that arises when using solely the model dimension, with $\overline{\text{sBIC}}_{0.5}$ performing the best.

6.2. Gaussian mixtures

Aoyagi (2010a) has found the learning coefficients of univariate Gaussian mixture models when the variances of the component distributions are known and equal to a common value. Using them in sBIC yields a criterion whose model selection properties are similar to what we have shown for reduced rank regression and binomial mixtures. In this section, we report instead on a data analysis with Gaussian mixtures where the variances are unknown and allowed to be unequal.

Let $\mathcal{M}_i$ be the (univariate) Gaussian mixture model with i mixture components, which comprises the distributions

$$\pi_i(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{h=1}^{i} \alpha_h \mathcal{N}(\mu_h, \sigma_h^2)$$

for a vector of mixture weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_i) \in \Delta_{i-1}$, choices of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_i) \in \mathbb{R}^i$ and variances $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_i^2) \in (\varepsilon, \infty)^i$. Here, we made explicit that the software that we shall use later, namely the R package `mclust` (Fraley *et al.*, 2012), uses a lower bound $\epsilon > 0$ to avoid the well-known singularities in the likelihood surfaces that are obtained by letting one or more variances tend to 0. Such a lower bound also appears in consistency theory for BIC (Keribin (2000), proposition 4.2).

In the Gaussian mixture model $\mathcal{M}_i$, the fibre of a distribution π_0 is the set

$$\mathcal{F}_i(\pi_0) = \{(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \in \Delta_{i-1} \times \mathbb{R}^i \times (\epsilon, \infty)^i : \pi_i(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \pi_0\}. \tag{6.9}$$

By proposition 1 in Teicher (1963), if $\pi_0 \in \mathcal{M}_i \setminus \mathcal{M}_{i-1}$ then $\mathcal{F}_i(\pi_0)$ is finite with $|\mathcal{F}_i(\pi_0)| = i!$, and we have

$$\dim(\mathcal{M}_i) = 3i - 1. \tag{6.10}$$

However, as discussed in Section 6.1, a distribution $\pi_0 \in \mathcal{M}_j \subset \mathcal{M}_i$, $j < i$, will have an infinite fibre $\mathcal{F}_i(\pi_0)$ due to the obvious non-identifiability problem arising from specifying the number of mixture components i larger than needed.

As described in the proof of proposition 4, a distribution $\pi_0 \in \mathcal{M}_j \setminus \mathcal{M}_{j-1}$ with $j < i$ can be represented as a member of $\mathcal{M}_i$ by setting $i - j$ of the mixture weights to 0, which here leaves $i - j$ of the mean parameters and $i - j$ of the variance parameters free. Hence, the fibre contains a set of dimension $2(i - j)$ that is made up of triples $(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ that have $\boldsymbol{\alpha}$ on the boundary of the probability simplex Δ_{i-1}. By this fact, the learning coefficient $\lambda_{ij} = \lambda_i(\pi_0)$ can be bounded as

$$\lambda_{ij} \leqslant \bar{\lambda}^1_{ij} := \tfrac{1}{2}\{\dim(\mathcal{M}_i) - 2(i-j)\} = \tfrac{1}{2}\{3j - 1 + (i-j)\} = \tfrac{1}{2}(i + 2j - 1); \tag{6.11}$$

see again section 7.3 in Watanabe (2009). For the bound to apply, however, the density of the prior distribution $P(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mathcal{M}_i)$ must be bounded away from zero in a neighbourhood of a $2(i - j)$-dimensional subset of $\mathcal{F}_i(\pi_0)$. This is so if the prior density for $\boldsymbol{\alpha}$ is bounded away from zero on and near the boundary of the probability simplex Δ_{i-1}; a uniform distribution on Δ_{i-1} would be an example.

Keeping with $\pi_0 \in \mathcal{M}_j \subset \mathcal{M}_i$, let Δ^o_{i-1} denote the interior of the probability simplex, and consider instead the fibre

$$\mathcal{F}^o_i(\pi_0) = \{(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \in \Delta^o_{i-1} \times \mathbb{R}^i \times (\epsilon, \infty)^i : \pi_i(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \pi_0\} \tag{6.12}$$

that has all mixture weights non-zero. This 'positive fibre' $\mathcal{F}^o_i(\pi_0)$ is of lower dimension than $\mathcal{F}_i(\pi_0)$. Indeed, equating means and variances between mixture components by analogy with expressions (6.5) and (6.6) shows that $\mathcal{F}^o_i(\pi_0)$ has dimension $i - j$. Hence, for a prior that is supported on a subset of Δ^o_{i-1}, subtraction of the fibre dimension leads to the bound

$$\lambda_{ij} \leqslant \bar{\lambda}^2_{ij} := \tfrac{1}{2}\{\dim(\mathcal{M}_i) - (i-j)\} = \tfrac{1}{2}\{3j - 1 + 2(i-j)\} = \tfrac{1}{2}(2i + j - 1). \tag{6.13}$$

Nevertheless, the more refined analysis from Rousseau and Mengersen (2011), equations (5) and (6), shows that the bound $\bar{\lambda}^1_{ij}$ from expression (6.11) remains valid when the prior density is bounded away from zero in a neighbourhood of a point in $\mathcal{F}^o_i(\pi_0)$.

To illustrate the above result in an example, take $\pi_0 = \mathcal{N}(0, 1)$: a standard normal distribution. Then the fibre $\mathcal{F}_2(\pi_0)$ in the two-component mixture model is the union of two planes and a line intersected with $\Delta_1 \times \mathbb{R}^2 \times (\epsilon, \infty)^2$. The structure of the fibre is as in Fig. 5, except that the two grey line segment now are two-dimensional rectangular strata. The black part with mixture weights $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$ remains a line segment. The set $\mathcal{F}^o_i(\pi_0)$ then comprises only this line segment but not the two-dimensional strata.

Using the bounds $\bar{\lambda}^1_{ij}$ from expression (6.11) and setting all multiplicities to 1 yields a version of sBIC, which we denote by $\overline{\text{sBIC}}_1$. (We shall briefly comment on the bounds $\bar{\lambda}^2_{ij}$ in our conclusion.) We apply $\overline{\text{sBIC}}_1$ to a familiar example, namely the galaxies data set that has been discussed in detail in Aitkin (2001) and also in example 4 in Marin *et al.* (2005). We use the EM algorithm implemented in the R package `mclust` (Fraley *et al.*, 2012) to fit the mixture models and base our results on the best local maxima of the likelihood function that were found in repeated EM runs. For each model, we ran the EM algorithm from 5000 random initializations that were created by drawing, independently for each data point, a vector of cluster membership probabilities from the uniform distribution on the relevant probability simplex. Fig. 7 depicts the resulting values of BIC and $\overline{\text{sBIC}}_1$. These are converted into posterior model probabilities in

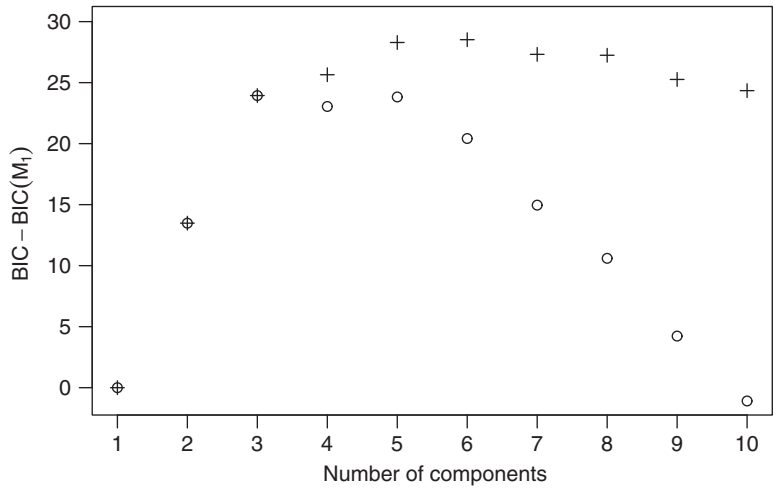

Fig. 7. Galaxies data (mixture of Gaussians with unequal variances): ∘, BIC; +, sBIC_1

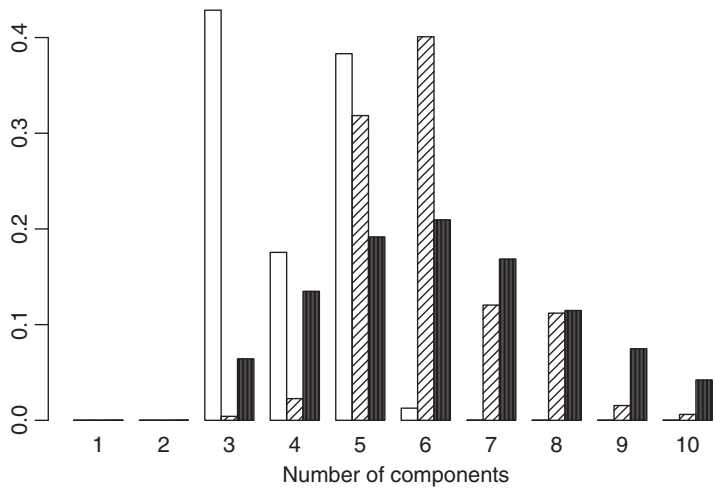

Fig. 8. Galaxies data (mixture of Gaussians with unequal variances): posterior model probabilities from BIC (□), $\overline{\text{sBIC}}_1$ (▨) and Markov chain Monte Carlo sampling per Richardson and Green (1997) (■)

Fig. 8, where we also show posterior probabilities from the fully Bayesian analysis of Richardson and Green (1997) who, in particular, adopted a uniform prior for the mixture weights.

Fig. 7 shows that the information criteria assign essentially the same value to the models $\mathcal{M}_i$ with $i \leqslant 3$. This is due to a poor model fit, i.e. very small maximal likelihood under $\mathcal{M}_1$ and $\mathcal{M}_2$. Starting with four components differences emerge. BIC attains high values for $i \in \{3, 4, 5\}$ and decreases very quickly for larger i. The decrease is nearly as quick as the increase through the models with $i \leqslant 3$ components, where those with $i \leqslant 2$ seem too simple. In contrast, $\overline{\text{sBIC}}_1$ is largest for $i = 6$ followed closely by $i = 5$, and its values remain rather large for $i \in \{7, 8\}$. The decay for larger i is far slower than for BIC. In Fig. 8, approximate posterior model probabilities from $\overline{\text{sBIC}}_1$ are closer to the Monte Carlo estimates that were reported by (Richardson and Green (1997); see also Lee and Robert (2013).

6.3. Latent class analysis

Our last experiments pertain to latent class analysis (LCA), in which the joint distribution of a collection of categorical variables, the *items*, is modelled to exhibit conditional independence given a categorical latent variable. The values of the latent variable are the *classes*. LCA models are also known as naive Bayes models (Geiger *et al*., 2001) and are related to secant varieties of Segre varieties studied in algebraic geometry (Drton *et al*. (2009), chapter 4.1). We consider them here because LCA models are mixture models in which the component distributions are taken from a family of larger dimension. As we shall see, this makes the choice of the priors on the mixture weights more important.

We shall treat the case of r binary items whose values we code to be in $\{0, 1\}$. The model $\mathcal{M}_i$ with i classes then postulates that the joint probabilities for the binary items $Y_1, \ldots, Y_r$ are of the form

$$\Pr(Y_1 = y_1, \ldots, Y_r = y_r) = \sum_{h=1}^{i} \alpha_h \prod_{l=1}^{r} p_{hl}^{y_l} (1 - p_{hl})^{1-y_l},$$

where α_h is the probability of being in class h, and p_{hl} is the conditional probability of $Y_l = 1$ given membership in class h. We emphasize that r is equal to the dimension of the family of distributions from which the mixture components are taken. Counting parameters we expect that the dimension of the LCA model $\mathcal{M}_i$ is

$$\min\{ir + (i-1), 2^r - 1\}. \tag{6.14}$$

There are exceptional cases where this is not the correct dimension; see for example example 4.1.8 in Drton *et al*. (2009). However, theorem 2.3 in Catalisano *et al*. (2005) guarantees that all models in our simulation study below have dimension given by expression (6.14). All these models are also generically identifiable up to label swapping by corollary 5 in Allman *et al*. (2009).

Let $\pi_0 \in \mathcal{M}_j \setminus \mathcal{M}_{j-1}$ for $j < i$, and assume that $\dim(\mathcal{M}_i) = i - 1 + ir \leqslant 2^r - 1$. Reasoning as in Section 6.2, dimension counting yields two simple bounds on the learning coefficients. For $\phi > 0$, define

$$\bar{\lambda}_{ij}^{\phi} := \tfrac{1}{2}\{jr + j - 1 + (i-j)\phi\}. \tag{6.15}$$

(Note that r is the dimension of the model for a single mixture component. The notation from expression (6.15) matches the earlier use in Section 6.1, where $r = 1$, and that in Section 6.2, where $r = 2$.) When allowing zero mixture weights α_h, the fibre of π_0 has dimension $r(i-j)$ and the analogue of expression (6.11) becomes

$$\lambda_{ij} \leqslant \tfrac{1}{2}\{\dim(\mathcal{M}_i) - r(i-j)\} = \tfrac{1}{2}rj + i - 1 = \bar{\lambda}_{ij}^{1}. \tag{6.16}$$

This bound is of relevance when the prior distribution of the mixture weights α_h is bounded away from zero in a neighbourhood of the boundary of the probability simplex Δ_{i-1}. Similarly, if the fibre is restricted to include only points with all $\alpha_h > 0$, then the dimension of this 'positive fibre' is only $i - j$ and the analogue of expression (6.13) is

$$\lambda_{ij} \leqslant \tfrac{1}{2}\{\dim(\mathcal{M}_i) - (i-j)\} = \tfrac{1}{2}(ri + j - 1) = \bar{\lambda}_{ij}^{r}. \tag{6.17}$$

This bound is of interest when the prior distribution of the mixture weights is zero along the boundary of Δ_{i-1} but bounded away from zero in a neighbourhood of a point in the positive

fibre. However, as in Sections 6.1 and 6.2, we may conclude from the work of Rousseau and Mengersen (2011) that for such priors it holds that

$$\lambda_{ij} \leqslant \tfrac{1}{2}\Big\{ rj + j - 1 + (i-j)\frac{r}{2} \Big\} = \bar{\lambda}_{ij}^{r/2}. \tag{6.18}$$

Contrasting the difference in dimension of the fibres $\mathcal{F}_i(\pi_0)$ and $\mathcal{F}_i^o(\pi_0)$ when $\pi_0 \in \mathcal{M}_j$ with $j < i$, it is clear that the choice of priors for the mixture weights $\boldsymbol{\alpha}$ may considerably impact posterior model probabilities. In particular, if the prior assigns non-negligible mass near the boundary of the probability simplex, then the likelihood function for a sample from $\pi_0 \in \mathcal{M}_j$ will be large near the high dimensional strata of $\mathcal{F}_i(\pi_0)$. Model $\mathcal{M}_i$ then behaves like a low dimensional model, and the Occam's razor effect from integrating the likelihood function in a Bayesian approach to model determination is weak. For our sBIC, this expresses itself via smaller values of (bounds on) learning coefficients, which leads to less penalization of the likelihood. In LCA and similar examples of mixtures of multiparameter distributions, it is thus useful to be more explicit about the effects of priors.

Suppose that the prior distribution $P(\boldsymbol{\alpha}|\mathcal{M}_i)$ is a Dirichlet distribution with all hyperparameters equal to $\phi > 0$, and that the remaining parameters p_{hl} are independent of $\boldsymbol{\alpha}$ *a priori* and have a positive joint density on $[0,1]^{ir}$. Then the learning coefficients $\lambda_{ij} = \lambda_i(\pi_0)$ depend on ϕ, and the result in Rousseau and Mengersen (2011), equations (5) and (6), shows that the bounds that were considered above may be refined to

$$\lambda_{ij} \leqslant \min\{\bar{\lambda}_{ij}^{\phi}, \bar{\lambda}_{ij}^{r/2}\}. \tag{6.19}$$

In light of this bound, we let $\overline{\mathrm{sBIC}}_\phi$ denote the version of our information criterion that is obtained when using the $\bar{\lambda}_{ij}^{\phi}$ from expression (6.15) as values of the learning coefficients and setting all multiplicities to 1. The behaviour of $\overline{\mathrm{sBIC}}_\phi$ may depend heavily on the choice of ϕ, with larger values of ϕ leading to stronger penalties and selection of a smaller number of mixture components.

When the goal is to stay close to Bayesian inference using Dirichlet priors for $\boldsymbol{\alpha}$, it may be clear which value of ϕ to use. It is less clear, however, what a default choice for ϕ should be when $\overline{\mathrm{sBIC}}_\phi$ is intended to be used as an information criterion with good frequentist model selection properties. Some guidance is provided by theorem 1 in Rousseau and Mengersen (2011), which shows that sufficiently small Dirichlet hyperparameters allow for detection of 0 components in an overfitting mixture model. According to their result, working with a single overfitting mixture model can be an alternative to the model selection set-up that is treated in this paper. When aiming to determine the number of mixture components in a model selection approach, however, larger Dirichlet hyperparameters have appeal in that they avoid large marginal likelihood for models for which one or more mixture components will remain empty when using the model for clustering. This point is also made in section 4.2 of Frühwirth-Schnatter (2006). More specifically, if we wish to avoid that overfitting mixture models act like models with fewer components, then theorem 1 of Rousseau and Mengersen (2011) suggests that ϕ should be chosen no less than $r/2$. Given the bound from inequality (6.19), we shall thus explore the properties of $\overline{\mathrm{sBIC}}_\phi$ with ϕ close to $r/2$ and compare it with the standard BIC, which is also equal to $\overline{\mathrm{sBIC}}_{r+1}$. This said, learning coefficients as large as $\bar{\lambda}_{ij}^{\phi}$ with $\phi > r/2$ cannot be realized when the prior density for the probabilities p_{hl} is everywhere positive but they could arise from priors whose densities are zero at the singularities with mixture weights $\alpha_h > 0$; see Petralia *et al.* (2012) for work that is related to this issue.

Our simulations apply BIC and $\overline{\mathrm{sBIC}}_\phi$ for recovery of the number of classes i in LCA. We adopt the following four settings from Nylund *et al.* (2007) that each have binary items:

Table 3. LCA: frequencies of selection of the number of classes by BIC and $\overline{\text{sBIC}}_\phi$ for various values of ϕ and four true classes

Model		*Frequencies for the following classes:*						*Frequencies for the following classes:*						*Frequencies for the following classes:*					
	n	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
8 item (equal)		*BIC*						$\overline{sBIC}_5$						$\overline{sBIC}_{4.5}$					
	50	32	39	26	*3*	0	0	2	9	21	*68*	0	0	2	4	14	*79*	1	0
	100	2	13	32	*53*	0	0	0	0	3	*97*	0	0	0	0	0	*100*	0	0
	150	0	1	6	*93*	0	0	0	0	0	*99*	1	0	0	0	0	*98*	2	0
	200	0	0	2	*98*	0	0	0	0	0	*99*	1	0	0	0	0	*97*	3	0
	500	0	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0
		$\overline{sBIC}_4$						$\overline{sBIC}_{3.5}$						$\overline{sBIC}_3$					
	50	0	2	9	*85*	4	0	0	0	4	*78*	18	0	0	0	1	*61*	29	9
	100	0	0	0	*96*	4	0	0	0	0	*80*	20	0	0	0	0	*62*	33	5
	150	0	0	0	*94*	5	1	0	0	0	*87*	12	1	0	0	0	*65*	30	5
	200	0	0	0	*96*	3	1	0	0	0	*86*	10	4	0	0	0	*78*	17	5
	500	0	0	0	*97*	3	0	0	0	0	*91*	9	0	0	0	0	*83*	14	3
8 item (unequal)		*BIC*						$\overline{sBIC}_5$						$\overline{sBIC}_{4.5}$					
	100	9	80	11	*0*	0	0	0	21	66	*13*	0	0	0	12	67	*20*	1	0
	200	0	46	50	*4*	0	0	0	3	61	*36*	0	0	0	1	51	*47*	1	0
	300	0	23	70	*7*	0	0	0	0	31	*68*	1	0	0	0	25	*73*	2	0
	500	0	0	53	*47*	0	0	0	0	5	*95*	0	0	0	0	3	*97*	0	0
	1000	0	0	2	*98*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0
		$\overline{sBIC}_4$						$\overline{sBIC}_{3.5}$						$\overline{sBIC}_3$					
	100	0	7	58	*32*	2	1	0	3	45	*43*	8	1	0	1	23	*53*	16	7
	200	0	1	43	*53*	3	0	0	0	35	*57*	8	0	0	0	19	*59*	22	0
	300	0	0	21	*76*	3	0	0	0	13	*76*	11	0	0	0	5	*72*	20	3
	500	0	0	1	*98*	1	0	0	0	0	*91*	9	0	0	0	0	*82*	18	0
	1000	0	0	0	*100*	0	0	0	0	0	*99*	1	0	0	0	0	*91*	9	0
10 item		*BIC*						$\overline{sBIC}_6$						$\overline{sBIC}_{5.5}$					
	100	0	18	82	*0*	0	0	0	0	55	*44*	1	0	0	0	46	*51*	3	0
	200	0	1	84	*15*	0	0	0	0	23	*77*	0	0	0	0	16	*83*	1	0
	300	0	0	52	*48*	0	0	0	0	5	*95*	0	0	0	0	3	*96*	1	0
	500	0	0	11	*89*	0	0	0	0	1	*99*	0	0	0	0	1	*99*	0	0
	1000	0	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0
		$\overline{sBIC}_5$						$\overline{sBIC}_{4.5}$						$\overline{sBIC}_4$					
	100	0	0	31	*65*	4	0	0	0	22	*63*	15	0	0	0	10	*50*	32	8
	200	0	0	12	*84*	4	0	0	0	9	*77*	13	1	0	0	6	*67*	22	5
	300	0	0	0	*99*	1	0	0	0	0	*94*	6	0	0	0	0	*85*	14	1
	500	0	0	0	*100*	0	0	0	0	0	*96*	4	0	0	0	0	*87*	12	1
	1000	0	0	0	*98*	2	0	0	0	0	*98*	2	0	0	0	0	*96*	3	1

Table 4. LCA: frequencies of selection of the number of classes by BIC and $\overline{\text{sBIC}}_\phi$ for various values of ϕ and three true classes (15-item model)

n	*Frequencies for the following classes:*						*Frequencies for the following classes:*						*Frequencies for the following classes:*					
	1	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
	BIC						$\overline{sBIC}_{8.5}$						$\overline{sBIC}_{8}$					
50	0	0	*100*	0	0	0	0	0	*92*	8	0	0	0	0	*88*	12	0	0
100	0	0	*100*	0	0	0	0	0	*98*	2	0	0	0	0	*93*	7	0	0
200	0	0	*100*	0	0	0	0	0	*99*	1	0	0	0	0	*97*	3	0	0
300	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*99*	1	0	0
400	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0	0
500	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*99*	1	0	0
1000	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0	0
	$\overline{sBIC}_{7.5}$						$\overline{sBIC}_{7}$						$\overline{sBIC}_{6.5}$					
50	0	0	*82*	18	0	0	0	0	*69*	27	4	0	0	0	*51*	36	10	3
100	0	0	*81*	19	0	0	0	0	*68*	32	0	0	0	0	*53*	43	4	0
200	0	0	*94*	5	1	0	0	0	*86*	13	1	0	0	0	*68*	25	7	0
300	0	0	*98*	2	0	0	0	0	*96*	4	0	0	0	0	*88*	12	0	0
400	0	0	*100*	0	0	0	0	0	*94*	6	0	0	0	0	*87*	13	0	0
500	0	0	*99*	1	0	0	0	0	*96*	4	0	0	0	0	*91*	9	0	0
1000	0	0	*100*	0	0	0	0	0	*100*	0	0	0	0	0	*100*	0	0	0

(a) $r=8$ items, $i_0=4$ true classes and equal class sizes ($\alpha_1=\alpha_2=\alpha_3=\alpha_4$);
(b) $r=8$ items, $i_0=4$ true classes and unequal class sizes;
(c) $r=10$ items, $i_0=4$ true classes and unequal class sizes;
(d) $r=15$ items, $i_0=3$ true classes and equal class sizes ($\alpha_1=\alpha_2=\alpha_3$).

Settings (b) and (c) have their unequal class sizes given by $\alpha_1=1/21$, $\alpha_2=2/21$, $\alpha_3=3/21$ and $\alpha_4=15/21$. We refer the reader to Table 2 in Nylund *et al.* (2007) for the precise description of the distributions that we simulate from. (Our weights α_i are proportional to the values 0.05, 0.1, 0.15 and 0.75, with sum 1.05, that are stated in Table 2 of Nylund *et al.* (2007).) Settings (a) and (b) differ only in the values of the mixture weights or class probabilities α_h. They are both 'simple' in the sense that for each item l only one of the class conditional probabilities p_{hl} is large. Setting (d) is simple in the same sense. In setting (c), each item has two large and equal class conditional probabilities p_{hl} and the other two probabilities p_{hl} are small and equal. For each setting, we draw 100 samples of various sizes n and select the number of classes $i \in \{1,\dots,6\}$ by maximizing the information criteria. Specifically, we optimized the standard dimension-based BIC as well as $\overline{\text{sBIC}}_\phi$ with $\phi=0.5,1,1.5,\dots,r$. Maximum likelihood estimates were computed by using the R package `poLCA` (Linzer and Lewis, 2011).

For settings (a)–(c), Table 3 reports the frequencies of how often a particular number of classes i was selected by BIC and $\overline{\text{sBIC}}_\phi$ with $2\phi \in \{r-2, r-1, r, r+1, r+2\}$. We see a tendency for the standard BIC to select overly simple models, especially at small sample size. This underselection is alleviated when using the criterion $\overline{\text{sBIC}}_\phi$ but some overselection arises for $\phi < r/2$. The choice $\phi=r/2$ performs quite well, and so does $\phi=(r+1)/2$.

We do not list any results for small values of ϕ, such as $\phi=1$, which corresponds to a uniform distribution as prior for the mixture weights. In all except a handful of cases, $\overline{\text{sBIC}}_1$ selected

the largest allowed number of classes, i.e. $i = 6$. When the sample size is $n = 500$ in setting (a), the relative frequency of $\overline{\text{sBIC}}_\phi$ selecting the truth of $i_0 = 4$ classes is 0.01, 0.18 and 0.57 for $\phi = 1.5$, 2, 2.5 respectively. For $n = 1000$ in setting (b), these numbers are 0.02, 0.31 and 0.70. For $n = 1000$ in setting (c), they are 0.00, 0.00 and 0.14.

Table 4 lists the model selection frequencies for the problem with $r = 15$ items and $i_0 = 3$ true classes. This is a problem in which models with $i \leqslant 2$ classes fit so poorly that they are never selected. All methods using a heavy penalty thus select the true number of classes in all cases. This happens for BIC and $\overline{\text{sBIC}}_\phi$ with $\phi \geqslant 11$. As earlier, we report details for $2\phi \in \{r-2, r-1, r, r+1, r+2\}$. We see overselection for $\phi < r/2 = 7.5$, which decreases as ϕ is increased to $\phi = r/2 = 7.5$, $\phi = (r+1)/2 = 8$ and $\phi = r/2 + 1 = 8.5$. We note that $\overline{\text{sBIC}}_\phi$ with $\phi \leqslant 3$ always selected the maximum allowed number of classes ($i = 6$), and the true number of classes ($i_0 = 3$) is never selected when $\phi = 3.5$. When $n = 1000$, the true number of classes ($i_0 = 3$) is selected with relative frequency 0.03, 0.33, 0.69, 0.88 and 0.96 when $\phi = 4$, 4.5, 5, 5.5, 6 respectively.

In summary, $\overline{\text{sBIC}}_\phi$ can provide considerable improvements over the standard BIC in terms of frequentist model selection properties. To avoid drastic overselection, ϕ should not be chosen too small, compared with $r/2$. Our above simulations suggest that taking $\phi = r/2$ or possibly a little larger, e.g. as $\phi = (r+1)/2$, could be a good default beyond the specific settings of LCA that we treated.

7. Conclusion

In this paper we introduced a new BIC for singular statistical models. The new criterion, sBIC, is free of Monte Carlo computation and coincides with the widely used BIC of Schwarz when the model is regular. Moreover, the criterion is consistent and maintains a rigorous connection to Bayesian approaches even in singular settings. This latter behaviour is made possible by exploiting theoretical knowledge about the learning coefficients that capture the large sample behaviour of the marginal likelihood integrals concerned. In simulations and data analysis, we showed that sBIC indeed leads to a 'more Bayesian' assessment of model uncertainty and that it may also lead to improved frequentist model selection when compared with the standard BIC.

7.1. Priors matter for sBIC

The marginal likelihood of a singular model may depend quite heavily on the prior distribution. In fact, the choice of prior may also have a strong influence on the learning coefficients that quantify the Occam's razor effect resulting from the integration over parameters. Therefore, different versions of sBIC, motivated by different choices of priors, can be of interest for a given singular model selection problem.

An example where prior distributions play an important role is mixture modelling with component distributions from a multiparameter family; recall our discussion in Section 6. Using LCA for illustration (Section 6.3), we showed how the learning coefficients and thus also sBIC depend in particular on whether and how quickly the prior density for the mixture weights decays to zero or diverges as the weight vector approaches the boundary of the probability simplex. We explored this in the context of Dirichlet prior distributions. (Strictly speaking, we considered general bounds for the learning coefficients of mixture models.) For good frequentist model selection of sBIC we suggest that Dirichlet hyperparameters are not chosen too small. In particular, the sBIC based on a uniform distribution on the mixture weights cannot be recommended as a default for analysing mixtures of multiparameter distributions. Similar recommendations for fully Bayesian approaches to mixture model selection can be found in Frühwirth-Schnatter (2006).

7.2. Dependence of sBIC on the 'universe of models'

When computing the sBIC of model $\mathcal{M}_i$, we average asymptotic proxies for the marginal likelihood that are based on Schwarz's idea of retaining terms from an asymptotic expansion. The fact that there is not just a single quantity to contemplate is a feature that distinguishes singular from regular models. The terms that are being averaged correspond to submodels $\mathcal{M}_j \subseteq \mathcal{M}_i$ that are deemed competitors in the model selection problem. As a result, the sBIC of a singular model $\mathcal{M}_i$ will generally depend on which set of models we wish to select from.

In most model selection problems there is a canonical set of models to be considered. For instance, in mixture modelling one typically considers all models with up to a certain number of components. We envision that sBIC will generally be applied with respect to such a natural collection of models, even if the primary focus was on two specific models.

It is also clear from its definition that the sBIC of model $\mathcal{M}_i$ can change only when omitting from consideration a model $\mathcal{M}_j \subset \mathcal{M}_i$. We would expect this to be done only if it is certain that these simpler models are fitting the data poorly, which would then have little effect on sBIC-scores. Consider as an example the version of sBIC for the galaxies data from Section 6.2 (denoted there as $\overline{\text{sBIC}}_1$). We might wonder how the sBIC-score for $\mathcal{M}_3$ would change if we no longer considered the too simplistic $\mathcal{M}_1$ and $\mathcal{M}_2$, which have only one and two mixture components respectively. In the new context, $\mathcal{M}_3$ would be the minimal model and its sBIC-score would coincide with the ordinary BIC of $\mathcal{M}_3$. In Fig. 7, the points depicting the BIC- and the sBIC-score for $\mathcal{M}_3$ cannot be distinguished. There is virtually no change in the sBIC-scores when omitting models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Nevertheless, it would be only more appealing if we could define the sBIC of a model without reference to the fit of other models. The mathematical reason for our consideration of other models is the fact that our criterion leverages large sample asymptotics that are based on fixing a data-generating distribution and letting the sample size grow. As in related distribution theory for hypothesis tests, the limits that we obtain will in general not change in a continuous fashion as we vary the data-generating distribution. (Of course, finite sample behaviour of the marginal likelihood will depend on the data-generating distribution in a continuous fashion.) Hence, if we want to avoid consideration of other models in the definition of a 'singular BIC', then more refined mathematical insights would be necessary. Specifically, we would need to find uniform asymptotic expansions to the marginal likelihood, in the sense of Wong (2001), chapter VII. This, however, is a task that would be significantly more difficult to accomplish than finding the already non-trivial to obtain learning coefficients. Indeed, we are not aware of any discussion of uniform expansions in the statistical literature, let alone any results on their form for specific examples. In light of these difficulties, we consider our proposed sBIC a promising approach of averaging pointwise expansions to mimic how uniform expansions would have to behave.

7.3. Large numbers of models

For problems that involve a moderate number of models and are amenable to an exhaustive model search, the computational effort in the calculation of sBIC-scores is comparable with that for the ordinary BIC as the effort is typically dominated by the process of fitting all models considered to the available data. However, the fact that our definition of sBIC requires fitting all models considered has a clear computational disadvantage when an exhaustive search is not possible. Indeed, it is not immediately clear how to implement strategies such as greedy search with sBIC. One possible approach would be to define sBIC by averaging only over 'neighbouring' submodels but the merit of such strategies still needs to be explored. This said, the work of Drton *et al.* (2016) shows promising results for selection of Gaussian latent forest models.

We note that when treating problems with a large number of models it can be beneficial to adopt non-uniform prior model probabilities; compare for example the work on regression models by Chen and Chen (2008) and Scott and Berger (2010), and the work on graphical models by Foygel and Drton (2010) and Gao *et al*. (2012). As mentioned in remark 3, it would be straightforward to incorporate prior model probabilities in the definition of sBIC.

7.4. Use of maximum likelihood estimates

Our aim was to generalize Schwarz's BIC in a way that recovers his familiar criterion when the models considered are regular (recall remark 2). For this, we estimate the true likelihood by evaluating the likelihood function at the maximum likelihood estimator. However, other estimators could be used instead. For instance, Roeder and Wasserman (1997) used posterior means. Similarly, one could consider posterior modes or penalized likelihood methods to stay closer to a fully Bayesian analysis or simply for regularization; see Fraley and Raftery (2007) and Baudry and Celeux (2015) for work on Gaussian mixtures. We note that penalization of the likelihood function would provide a way to address the failure of assumption 1 from Section 4 that may occur in mixture models with unbounded parameter space (Hartigan, 1985).

7.5. When learning coefficients are not known

To our knowledge, sBIC is the first statistical method to make use of mathematical information about the values of learning coefficients of singular models. The theoretical insights allow us to obtain (crude) approximations to posterior model probabilities without Monte Carlo integration. At the same time, the reliance on theory also presents a limitation as the learning coefficients may not always be known. Previous studies have shown that, when exact values of learning coefficients are difficult to find, it may still be possible to obtain bounds. For priors that are bounded from above, a learning coefficient can be trivially bounded by the model dimension and using dimensions in sBIC recovers the standard BIC (recall remark 2). However, more interesting bounds can often be found by arguments that are only slightly more complicated than parameter counting. The usefulness of such bounds was demonstrated in Section 6.

Finally, our sBIC provides strong positive motivation for theoretical studies of learning coefficients. From a statistical perspective, past work had a negative flavour; knowing the values one could stress just how much smaller they can be than a parameter count. In contrast, new theoretical insights now yield new statistical methodology. We expect that this positive motivation will lead to further work and results on learning coefficients.

Acknowledgements

This collaboration started at a workshop at the American Institute of Mathematics, and we thank the participants of the workshop for helpful discussions. Particular thanks go to Vishesh Karwa, Dennis Leung and Luca Weihs for help with some of the numerical work. Mathias Drton was supported by grants from the National Science Foundation (DMS-1305154) and the RRF at the University of Washington as well as an Alfred P. Sloan Fellowship.

References

Aitkin, M. (2001) Likelihood and Bayesian analysis of mixtures. *Statist. Modllng*, **1**, 287–304.
Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.
Allman, E. S., Matias, C. and Rhodes, J. A. (2009) Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, **37**, 3099–3132.

Allman, E. S., Rhodes, J. A., Sturmfels, B. and Zwiernik, P. (2015) Tensors of nonnegative rank two. *Lin. Alg. Appl.*, **473**, 37–53.
Aoyagi, M. (2009) Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three-layered neural network in Bayesian estimation. *Int. J. Pure Appl. Math.*, **52**, 177–204.
Aoyagi, M. (2010a) A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Communs Statist. Theor. Meth.*, **39**, 2667–2687.
Aoyagi, M. (2010b) Stochastic complexity and generalization error of a restricted Boltzmann machine in Bayesian estimation. *J. Mach. Learn. Res.*, **11**, 1243–1272.
Aoyagi, M. and Watanabe, S. (2005) Stochastic complexities of reduced rank regression in Bayesian estimation. *Neurl Netwrks*, **18**, 924–933.
Arnold, V. I., Guseĭn-Zade, S. M. and Varchenko, A. N. (1988) *Singularities of Differentiable Maps*, Vol. II. Boston: Birkhäuser.
Azaïs, J.-M., Gassiat, É. and Mercadier, C. (2006) Asymptotic distribution and local power of the log-likelihood ratio test for mixtures: bounded and unbounded cases. *Bernoulli*, **12**, 775–799.
Azaïs, J.-M., Gassiat, É. and Mercadier, C. (2009) The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Statist.*, **13**, 301–327.
Baudry, J.-P. and Celeux, G. (2015) EM for mixtures. *Statist. Comput.*, **25**, 713–726.
Bickel, P. J. and Chernoff, H. (1993) Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: a Raghu Raj Bahadur Festschrift* (eds K. P. J. K. Ghosh, S. K. Mitra and B. P. Rao), pp. 83–96. New Delhi: Wiley Eastern.
Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference*, 2nd edn. New York: Springer.
Catalisano, M. V., Geramita, A. V. and Gimigliano, A. (2005) Higher secant varieties of the Segre varieties $\mathbb{P}^1 \times \ldots \times \mathbb{P}^1$. *J. Pure Appl. Alg.*, **201**, 367–380.
Charnigo, R. and Pilla, R. S. (2007) Semiparametric mixtures of generalized exponential families. *Scand. J. Statist.*, **34**, 535–551.
Chen, J. and Chen, Z. (2008) Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759–771.
Cheng, X. and Phillips, P. C. (2012) Cointegrating rank selection in models with time-varying variance. *J. Econmetr.*, **169**, 155–165.
Claeskens, G. and Hjort, N. L. (2008) *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.
DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Statist. Ass.*, **92**, 903–915.
Drton, M. (2009) Likelihood ratio tests and singularities. *Ann. Statist.*, **37**, 979–1012.
Drton, M., Lin, S., Weihs, L. and Zwiernik, P. (2016) Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli*, to be published.
Drton, M., Sturmfels, B. and Sullivant, S. (2007) Algebraic factor analysis: tetrads, pentads and beyond. *Probab. Theor. Reltd Flds*, **138**, 463–493.
Drton, M., Sturmfels, B. and Sullivant, S. (2009) *Lectures on Algebraic Statistics*, vol. 39, Basel: Birkhäuser.
van Erven, T., Grünwald, P. and de Rooij, S. (2012) Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma (with discussion). *J. R. Statist. Soc.* B, **74**, 361–417.
Foygel, R. and Drton, M. (2010) Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neurl Inf. Process. Syst.*, **23**, 2020–2028.
Fraley, C. and Raftery, A. E. (2007) Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classificn*, **24**, 155–181.
Fraley, C., Raftery, A. E., Murphy, T. B. and Scrucca, L. (2012) MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report 597*. Department of Statistics, University of Washington, Seattle.
Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc.* B, **70**, 589–607.
Friel, N. and Wyse, J. (2012) Estimating the evidence—a review. *Statist. Neerland.*, **66**, 288–308.
Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. New York: Springer.
Gao, X., Pu, D. Q., Wu, Y. and Xu, H. (2012) Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sin.*, **22**, 1123–1146.
Gassiat, E. and van Handel, R. (2013) Consistent order estimation and minimal penalties. *IEEE Trans. Inform. Theor.*, **59**, 1115–1128.
Gassiat, E. and van Handel, R. (2014) The local geometry of finite mixtures. *Trans. Am. Math. Soc.*, **366**, 1047–1072.
Geiger, D., Heckerman, D., King, H. and Meek, C. (2001) Stratified exponential families: graphical models and model selection. *Ann. Statist.*, **29**, 505–529.

Hartigan, J. A. (1985) A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conf. Honor of Jerzy Neyman and Jack Kiefer*, vol. II, pp. 807–810. Belmont: Wadsworth.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) Data mining, inference, and prediction. In *The Elements of Statistical Learning*, 2nd edn. New York: Springer.
Haughton, D. (1989) Size of the error in the choice of a model to fit data from an exponential family. *Sankhya* A, **51**, 45–58.
Haughton, D. M. A. (1988) On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.
Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (with comments). *Statist. Sci.*, **14**, 382–417.
[Kass, R. E. Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, **90**, 928–934.
Keribin, C. (2000) Consistent estimation of the order of mixture models. *Sankhya* A, **62**, 49–66.
Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*. New York: Springer.
Lee, J. E. and Robert, C. P. (2013) Importance sampling schemes for evidence approximation in mixture models. *Preprint arXiv:1311.6000*.
Lin, S. (2011) Asymptotic approximation of marginal likelihood integrals. *Preprint arXiv:1003.5338v2*.
Linzer, D. A. and Lewis, J. B. (2011) poLCA: an R package for polytomous variable latent class analysis. *J. Statist. Softwr.*, **42**, 1–29.
Liu, X. and Shao, Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.*, **31**, 807–832.
Lopes, H. F. and West, M. (2004) Bayesian model assessment in factor analysis. *Statist. Sin.*, **14**, 41–67.
Marin, J.-M., Mengersen, K. and Robert, C. P. (2005) Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking: Modeling and Computation*, pp. 459–507. Amsterdam: Elsevier.
McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley-Interscience.
Neal, R. (1999) Erroneous results in 'Marginal likelihood from the Gibbs output'. *Letter*. Unpublished. (Available from `http://www.cs.toronto.edu/~radford`.)
Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.
Nobile, A. (2005) Bayesian finite mixtures: a note on prior specification and posterior computation. arXiv:0711.0458. Department of Statistics, University of Glasgow, Glasgow.
Nylund, K. L., Asparouhov, T. and Muthén, B. O. (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equ. Model.*, **14**, 535–569.
Okamoto, M. (1973) Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, **1**, 763–765.
Petralia, F., Rao, V. and Dunson, D. B. (2012) Repulsive mixtures. In *Advances in Neural Information Processing Systems*, vol. 25 (eds F. Pereira, C. Burges, L. Bottou and K. Weinberger), pp. 1889–1897. Red Hook: Curran Associates.
Raftery, A. E. (1995) Bayesian model selection in social research. *Sociol. Methodol.*, **25**, 111–163.
Reinsel, G. C. and Velu, R. P. (1998) *Multivariate Reduced-rank Regression*. New York: Springer.
Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792.
Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *J. Am. Statist. Ass.*, **92**, 894–902.
Rotnitzky, A., Cox, D. R., Bottai, M. and Robins, J. (2000) Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243–284.
Rousseau, J. and Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc.* B, **73**, 689–710.
Rusakov, D. and Geiger, D. (2005) Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.*, **6**, 1–35.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Scott, J. G. and Berger, J. O. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.*, **38**, 2587–2619.
Steele, R. J. and Raftery, A. E. (2010) Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, ch. 4.1, pp. 113–130. New York: Springer.
Teicher, H. (1963) Identifiability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.
Vasil'ev, V. A. (1979) Asymptotic behavior of exponential integrals in the complex domain. *Funkt. Anal. Prilzhn.*, **13**, 1–12.

Wasserman, L. (2000) Bayesian model selection and model averaging. *J. Math. Psychol.*, **44**, 92–107.
Watanabe, S. (2001) Algebraic analysis for nonidentifiable learning machines. *Neurl Comput.*, **13**, 899–933.
Watanabe, S. (2009) *Algebraic Geometry and Statistical Learning Theory*. Cambridge: Cambridge University Press.
Watanabe, S. (2013) A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.*, **14**, 867–897.
Watanabe, S. and Amari, S.-I. (2003) Learning coefficients of layered models when the true distribution mismatches the singularities. *Neurl Comput.*, **15**, 1013–1033.
Watanabe, K. and Watanabe, S. (2007) Stochastic complexity for mixture of exponential families in generalized variational Bayes. *Theoret. Comput. Sci.*, **387**, 4–17.
Wit, E., van den Heuvel, E. and Romeijn, J.-W. (2012) 'All models are wrong ...': an introduction to model uncertainty. *Statist. Neerland.*, **66**, 217–236.
Wong, R. (2001) *Asymptotic Approximations of Integrals*. Philadelphia: Society for Industrial and Applied Mathematics.
Yamazaki, K. and Watanabe, S. (2003) Singularities in mixture models and upper bounds of stochastic complexity. *Neurl Netwrks*, **16**, 1029–1038.
Yamazaki, K. and Watanabe, S. (2004) Newton diagram and stochastic complexity in mixture of binomial distributions. In *Algorithmic Learning Theory*, pp. 350–364. Berlin: Springer.
Yamazaki, K. and Watanabe, S. (2005) Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, **69**, 62–84.
Yang, Y. (2005) Can the strengths of AIC and BIC be shared?: a conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950.
Zwiernik, P. (2011) An asymptotic behaviour of the marginal likelihood for general Markov models. *J. Mach. Learn. Res.*, **12**, 3283–3310.
Zwiernik, P. and Smith, J. Q. (2012) Tree cumulants and the geometry of binary tree models. *Bernoulli*, **18**, 290–321.