SPARSE MODEL FITTING IN NESTED FAMILIES: BAYESIAN APPROACH VS PENALIZED LIKELIHOOD

Laure Amate and Maria João Rendas

Laboratoire I3S, UNSA/CNRS
2000 route des lucioles BP 121, 06903, Sophia Antipolis, France
phone: + (33) 4 92 94 27 71, fax: + (33) 4 92 94 28 98, email: amate,rendas@i3s.unice.fr
web: www.i3s.unice.fr

ABSTRACT

We study the problem of model fitting in the framework of nested probabilistic families. Our criteria are: (*i*) sparsity of the identified representation, (*ii*) its ability to fit the (finite length) data set available. As we show in this paper, current methodologies, often taking the form of penalized versions of the data likelihood, cannot simultaneously satisfy these requirements, as the examples presented clearly demonstrate. On the contrary, maximization of the Bayesian model posterior, even without assumption of a complexity penalizing prior, is able to select models with appropriate complexity, enabling sound determination of its parameters in a second step.

1. INTRODUCTION

1.1 Problem formulation

In many situations of interest the ultimate goal is to identify **one** simple model that "correctly" describes the structure of the observed data Z. One can then discard the original data using the identified model as a proxy for it. This is different from denoising-like problems, where we are not interested in *learning* the data structure, but rather to "clean it".
Generally speaking, the complexity of a model is a measure of the information needed to specify it. A convenient way of adjusting the model complexity to the data complexity is to consider a set $\mathscr{M}$ of candidate models that is the union of nested families of parametric models $\mathscr{M}_k$:

$$\mathscr{M} = \bigcup_{k\in\mathscr{K}} \mathscr{M}_k \quad ; \quad \mathscr{K} = \{k_{\min} \cdots k_{\max}\} \tag{1}$$

$$\mathscr{M}_k = \{p(\cdot|\theta), \theta \in \Theta_k\} \subset \mathscr{M}_{k+1} \ . \tag{2}$$

The integer k indexing each family of models $\mathscr{M}_k$ is directly related to their complexity: if $k' > k$, the complexity of models in $\mathscr{M}_{k'}$ is higher than the complexity of those in $\mathscr{M}_k$. The overall parameter space Θ of $\mathscr{M}$ is simply

$$\Theta = \bigcup_{k\in\mathscr{K}} \Theta_k \ . \tag{3}$$

To make the problem stated above tractable, we need to specify what "correctly" means, i.e., to define the criterion that the selected model $\hat{p} = p(\cdot|\hat{\theta}_{\hat{k}})$ must optimize. Two common choices for parameter estimation are Maximum Likelihood (ML) and Bayesian. It is well known that ML is inconsistent for models with this structure, systematically preferring the most complex models. Bayesian approaches regularize the identification problem through the definition of a prior over Θ, and optimize the expected value of some functional of the estimation error under the posterior distribution over Θ. Set (3) no longer has the structure of a vector space, even when each Θ_k is a vector space. Definition of distributions which are the basic entities manipulated by Bayesian techniques (e.g. prior or posterior distributions) must be done with care. When parametric probabilistic families over each Θ_k are known, as we assume her, an intuitive way is to use a mixture-like approach, writing densities over Θ as

$$p(\theta) = \sum_{k\in\mathscr{K}} p_k(\theta), \qquad \theta \in \Theta \ , \tag{4}$$

where each $p_k(\theta)$ is the Radon-Nikodym derivative of an un-normalized measure with respect to (w.r.t) the invariant measure over Θ_k: if $\theta \notin \Theta_k$ then $p_k(\theta) = 0$, and

$$1 \geq \int_{\Theta_k} p_k(\theta) d\theta \equiv Pr(\mathscr{M}_k) \ .$$

A proper density $\nu_k(\theta)$ over Θ_k is obtained by normalization:

$$\nu_k(\theta) \triangleq \frac{p_k(\theta)}{\int_{\Theta_k} p_k(\theta) d\theta} \equiv p(\theta|\mathscr{M}_k) \quad , \theta \in \Theta_k \ ,$$

where we stressed the meaning of ν_k as resulting from conditioning on $\theta \in \Theta_k$. Note that these "local" densities (as their un-normalized versions) are defined w.r.t distinct reference measures, the invariant measures, over each Θ_k and are thus not directly comparable.

1.2 Background

To correct the overfitting behavior of ML, many authors proposed the addition of corrective terms that favor selection of models using less parameters:

$$\hat{\theta}_{PL} = \arg\max_{k,\theta} p(Z|\theta) + \mathscr{P}(k) \ , \tag{5}$$

where $\mathscr{P}(k)$ is a decreasing function of k. These techniques are generally known by the name of "penalized likelihood" methods, and have received three distinct justifications. *(i)* They are sometimes dictated by the desire to impose some regularity characteristics on the solution, requiring in this case prior knowledge on its characteristics which is not necessarily available. *(ii)* They have also been justified as particular cases of MAP (Maximum a Posteriori) estimation for special selections of the prior $p(\theta)$, e.g. in [2, 4]. We will see below that there is a fundamental flaw in this approach. *(iii)* More generically, they are derived from

asymptotic arguments which relate them to the models' Bayesian marginal posterior [9] $Pr(\mathscr{M}_k|Z)$ – this is the case for AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and MDL (Minimum Description Length) – [13, 12], and thus cannot offer any guarantee for finite data sets.

As we see, generic justifications of penalized likelihood link it to Bayesian estimates: either to the model posterior $p(\theta|Z)$ or to the marginal model posteriors $Pr(\mathscr{M}_k|Z)$. Let us concentrate first on Bayesian estimates for θ. The two most popular Bayesian criteria are the MMSE (Minimum Mean Square Error) and the MAP (Maximum A Posteriori). For the union-type models like (1)-(2)-(3), the MMSE criterion is meaningless, the associated cost function $(\theta - \hat{\theta}(Z))^2$ being undefined when $\theta \in \Theta_k$ and $\hat{\theta}(Z) \in \Theta_{k'}$ with $k \neq k'$. On the contrary, the 0/1 cost function of the MAP criterion is well defined, and should enable determination of a unique model amongst the set of candidate models. When Θ has the simple structure of a vector space, this criterion leads to $\hat{\theta}(Z) = \arg\max_\theta p(\theta|Z)$. Surprisingly, several authors have proposed estimators based on a direct transposition of this equation to the model structure considered herein [1, 4]. We designate them by "naive MAP" estimators:

$$\begin{aligned} \hat{\theta}_{nMAP} = \arg\max_{\theta\in\Theta} p(\theta|Z) &= \arg\max_{k\in\mathscr{K}} \max_{\theta\in\Theta_k} p_k(\theta|Z) \quad (6) \\ &= \arg\max_{k\in\mathscr{K}} p_k(\hat{\theta}_k|Z) \ , \\ \hat{\theta}_k &= \arg\max_{\theta\in\Theta_k} p_k(\theta|Z) \end{aligned}$$

As we pointed out before, the un-normalized densities $p_k(\theta|Z)$ defined over each Θ_k are not defined with respect to the same measures. This criterion, that abusively compares them directly, may lead to estimates with pathological behavior as the examples given below demonstrate.

We address now the second justification of penalized likelihood, that relates it to Bayesian Model Selection (BMS) which does not attempt at directly selecting the model set *and* the parameter value, and instead starts by selecting the model $\mathscr{M}_k$ using the posterior probabilities $Pr(\mathscr{M}_k|Z)$, $k \in \mathscr{K}$ [14]:

$$\hat{k} = \arg\max_k \int_{\Theta_k} p(\theta|Z)\, d\theta \ . \quad (7)$$

Determination of these posteriors requires specification of a priori distribution π defined over Θ, that if π is of the form (4) implies $\forall k \in \mathscr{K}$,

$$Pr(\mathscr{M}_k|Z) = \frac{p(Z|\mathscr{M}_k)\pi(\mathscr{M}_k)}{\sum_{j\in\mathscr{K}} p(Z|\mathscr{M}_j)\pi(\mathscr{M}_j)} \ , \quad (8)$$

$$Pr(Z|\mathscr{M}_k) = \int_{\Theta_k} p(Z|\mathscr{M}_k,\theta_k)\,\pi(\theta_k|\mathscr{M}_k) d\theta_k \ . \quad (9)$$

We refer to [9] for some analysis about how one can specify the prior. Once the model is determined, the parameter $\theta \in \Theta_{\hat{k}}$ can be estimated using one of the standard statistical estimation criteria (ML, MMSE, MAP,...). This (sound) estimation approach selects k by comparing the total posterior probability mass accumulated over $\mathscr{M}_k$. One may question whether this marginal approach can guarantee the aptitude of the elements of $\mathscr{M}_{\hat{k}}$, alone, to fit the data well. We will see below that in all the examples considered the models selected by BMS have fitting properties close to those obtained by penalized likelihood criteria. More surprisingly, numerical studies not reported here show that their fitting performance in "signal-in-noise" problems is similar to direct MMSE "signal" estimation, which belongs to a set much richer than $\mathscr{M}$.

1.3 Numerical issues

For most problems of practical interest Bayesian approaches must resort to numerical (Monte Carlo) methods [3, 7]. A tool that has now become a standard to draw from posteriors $p(\theta|Z)$ with the structure (1)-(2) is RJMCMC (Reversible Jump Markov Chain Monte Carlo) [8]. It builds a Markov Chain over Θ that asymptotically converges to $p(\theta|Z)$, enabling numerical determination of expected values with respect to it.

Models with the structure (1)-(2)-(3) often occur in the search for a parsimonious parametric model $f(t;\theta)$ for a signal $s(t)$ of which we observe noisy samples:

$$Z(t_i) = s(t_i) + \varepsilon(t_i) \simeq f(t_i;\theta), \quad i = 1\cdots N \ . \quad (10)$$

where the distribution of the noise $\varepsilon(t)$ is known. For this problem BARS (Bayesian Adaptive Regression Splines) [6] uses RJMCMC to find an estimate of $s(t)$ as $E[f(t;\theta)|Z]$, giving up identification of a single parametric model for $s(t)$. In [2, 1], Reversible Jump Simulated Annealing (RJSA, a Simulated Annealing (SA) algorithm with a proposal distribution controlled by a RJMCMC kernel) is used to find the maximum of the "posterior density" in (6). We can also mention the trans-dimensional simulated annealing (TDSA) of [4], that maximizes a penalized likelihood interpreted by the authors as a "posterior over Θ".

1.4 Outline

In our (large) introduction, we concluded that only BMS can serve as a sound method for model identification using nested model families. In section 2 we describe a numerical implementation of BMS for problem (10), that first computes the marginal MAP criterion and then identifies a model within the family of models chosen previously using a standard MAP criterion. Section 3 illustrates the possible pathological behavior of the "naive MAP" estimator (6) – directly relating it to the varying dimension of the parameter sets Θ_k – in a simple case-study where analytical determination of the posterior is possible. The last section compares use of the BIC penalty to the actual computation of the BMS criterion within the context of free-knots splines curve modeling, revealing its biased behavior for small data sets.

2. TWO-STEP MAP ESTIMATION

We present now a numerical implementation of the "Two-step MAP estimator",

$$(i) \quad \hat{k} = \arg\max_{k\in\mathscr{K}} Pr(\mathscr{M}_k|Z) \ , \quad (11)$$

$$(ii) \quad \hat{\theta} = \arg\max_{\theta\in\Theta_{\hat{k}}} p(\theta|Z,\mathscr{M}_{\hat{k}}) \ . \quad (12)$$

that uses BMS to select $\mathscr{M}_k$ and MAP to identify a $\theta \in \Theta_{\hat{k}}$. In this manner optimization is done in each step using commensurable score functions: the (discrete) posterior distribution over the families of models $\mathscr{M}_k$ in *(i)*, and a regular posterior density over $\Theta_{\hat{k}}$, with respect to a selected base measure, in *(ii)*. Note that [10] proposes a similar idea for ML estimation, selecting first the $\mathscr{M}_k$ using the BMS criterion and computing the ML estimates of the $\theta \in \Theta_k$ in a second step.

We consider a prior π of the form (4),

$$\pi_\theta(\theta) = \sum_{k=k_{\min}}^{k_{\max}} \pi(\theta|\mathscr{M}_k)\, Pr(\mathscr{M}_k), \qquad \theta \in \Theta \ .$$

2.1 Identification of $\mathscr{M}_k$ (family selection)

For most problems the posterior probabilities $Pr(\mathscr{M}_k|Z)$ have no closed-form and their maximum must be determined numerically. We obtain estimates $\hat{Pr}(\mathscr{M}_k|Z)$ by sampling from $p(\theta|Z), \theta \in \Theta$ using RJMCMC [8] and computing the total mass of each $\mathscr{M}_k$ as the corresponding marginals. RJMCMC uses a proposal distribution $q(\theta'|\theta)$, $\theta' \in \Theta_{k'}$, $\theta \in \Theta_k$, where θ (resp. θ') is the current (resp. candidate) state of the chain, that is a mixture of basic transition distributions (birth, death or change) moving across neighboring families of models $\mathscr{M}_k$ and $\mathscr{M}_{k+1}$. To ensure chain reversibility (and thus convergence in distribution to the target distribution $p(\theta|Z)$) the acceptance function of the chain is [8] $\alpha_{RJ}(\theta,\theta') = \min\{1, r_{RJ}\}$, with

$$r_{RJ} = \frac{p(\theta'|Z)}{p(\theta|Z)} \frac{q(\theta'|\theta)}{q(\theta|\theta')} J(\theta',\theta) \ , \qquad (13)$$

where $J(\theta',\theta)$ is the Jacobian of the mapping from θ to θ'. Finally, the family $\mathscr{M}_{\hat{k}}$ is chosen using the RJMCMC samples $\left(\theta^{(i)}_{k^{(i)}}\right)_{i=1}^{M} \propto p(\theta|Z)$ in the following manner

$$\hat{k} = \arg\max_k \hat{Pr}(\mathscr{M}_k|Z), \qquad \hat{Pr}(\mathscr{M}_k|Z) = \frac{M_k}{M}, \qquad (14)$$

where $M_k = \#\left\{\left(\theta^{(i)}_{k^{(i)}}\right)_{i=1}^{M} : k^{(i)} = k\right\}$ ($\# A$ is the cardinality of set A).

2.2 Estimation of θ (parameter estimation)

Once the family $\mathscr{M}_{\hat{k}}$ has been determined, parameter estimation is done for $\theta \in \Theta_{\hat{k}}$. Again, there is, in general, no analytical solution, and we must resort to a numerical method to find the model with the maximal posterior density $p_{\hat{k}}(\theta|Z)$. A common choice for approximating the solution of this optimization problem is Simulated Annealing (SA), with an acceptance probability α_{SA}:

$$\alpha_{SA}(\theta'_{\hat{k}}, \theta_{\hat{k}}, T_i) = \min\left\{1; \left(\frac{p(\theta'_{\hat{k}}|Z,\mathscr{M}_{\hat{k}})}{p(\theta_{\hat{k}}|Z,\mathscr{M}_{\hat{k}})}\right)^{\frac{1}{T_i}}\right\}, \qquad (15)$$

where $\theta_{\hat{k}}$ (resp. $\theta'_{\hat{k}}$) is the current (resp. candidate) state, and T_i is the chain temperature that must decrease according to a convenient cooling scheme. We refer the interested reader to [11] for details on SA.

σ^2	0.6	1	1.2	1.6	2	4
$Pr(\tilde{\theta} \in \mathscr{M}_1 \mid \mathscr{M}_2)$	0	0.25	0.61	0.945	0.985	1

Table 1: $Pr(\tilde{\theta} \in \mathscr{M}_1|\mathscr{M}_2)$, $\sigma^2 = (0.6, 1, 1.2, 1.6, 2, 4)$.

3. NAIVE MAP: PATHOLOGICAL BEHAVIOR

In this section we expose using a simple example the possible biased behaviour of the "naive MAP" criterion.

3.1 Case-study

Let $Z = [z_1, \ldots, z_n] \in \mathbb{R}^n$ be the observation vector, and consider a model with just two families: $\mathscr{M} = \mathscr{M}_1 \cup \mathscr{M}_2$, where[1]

$$\mathscr{M}_k = \left\{\mathscr{N}\left(f_k(\cdot|\theta), \sigma^2 \mathbf{I}_n\right), \theta \in \Theta_k\right\}, k = 1, 2 \ .$$

Above, $f_k(\cdot|\theta)$ is a k-piecewise linear signal, see Figure 1*(a)*, such that the parameters of the models are $\theta_1 = [P^0, \sigma^2]$ and $\theta_2 = [P^1, P^2, \sigma^2]$, where $\{P^i = (P^i_x, P^i_z) \in \mathbb{R}^2\}_{i=0}^2$ are the break point coordinates.

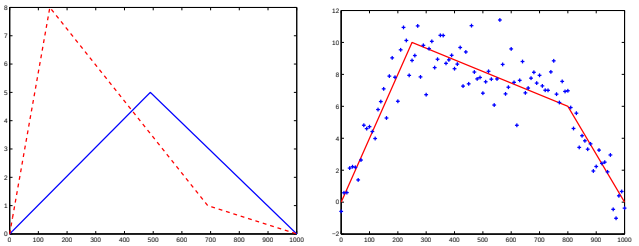
We use an uninformative factored *prior* distribution over both parameter spaces: $P^i_x \propto \mathscr{U}([0N])$, $P^i_z \propto \mathscr{N}(0, \Sigma^2), i = 0, 1, 2$, and $\sigma^2 \propto \mathscr{U}(\mathbb{R}^+)^2$. To minimize the impact of the prior we set $\Sigma^2 \gg 1$. Models are equiprobable: $\pi(\mathscr{M}_1) = \pi(\mathscr{M}_2)$.

3.2 Naive MAP estimator

Apparently, the prior chosen does not express preference for $\mathscr{M}_1$, and one would expect that the biased ML behaviour (always choose $\mathscr{M}_2$) would not be corrected. We will see that this is not the case, and that a bias of opposite sense is induced, (6) exhibiting a strong preference for the simpler model $\mathscr{M}_1$.

Table 1 displays estimates of the error probability $\Pr\left(\hat{\theta} \in \Theta_1 | \theta \in \mathscr{M}_2\right)$ for several values of the noise variance σ^2, obtained over 200 Monte Carlo runs. Parameters were set at: $P^1 = (250, 10), P^2 = (800, 6)$, $\Sigma = 10^{16}$ and $X = \{10m\}_{m=0}^{100}$, see figure 1*(b)*. We see that even for small values of the noise variance the error probability is very high: the "naive MAP" estimator is biased toward the **simplest** model $\mathscr{M}_1$, even if the data clearly shows the existence of 3 different slopes.

We will now show that this preference for simple models



(a) *(b)*

Figure 1: *(a)* $f(\cdot|\theta_1)$ ($-$ blue), $f(\cdot|\theta_2)$ ($--$ red) ; *(b)* $Z \propto p(\cdot) \in \mathscr{M}_2$, $\sigma^2 = 1.2$.

[1] $\mathscr{N}(\mu, \Sigma)$ denotes the normal density with mean μ and covariance matrix Σ.

[2] $X \propto p$ indicates that random variable X is drawn according to p and $\mathscr{U}(A)$ denotes the uniform distribution over set A.

that seems to be hidden in the prior chosen is actually an artifact caused by the comparison of densities defined with respect to distinct measures. Let $\rho_i^2 = ||Z - f_i(X;\theta_i)||^2$ be the residuals for model $\mathscr{M}_i$, $i = (1,2)$. Criterion (6) leads to a model selection rule $r \underset{\mathscr{M}_2}{\overset{\mathscr{M}_1}{\lessgtr}} 1$ where r is the ratio

$$\begin{aligned} r &= \frac{\pi(\mathscr{M}_2)\, p_2(\theta_2|Z)}{\pi(\mathscr{M}_1)\, p_1(\theta_1|Z)} \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{\Sigma N} \left(\frac{\rho_1}{\rho_2}\right)^n \exp\left(-\frac{(P_z^1)^2 + (P_z^2)^2 - (P_z^0)^2}{2\Sigma^2}\right) . \end{aligned}$$

If $\Sigma^2 \gg 1$, such that $\exp\left(-\frac{(P_z^1)^2+(P_z^2)^2-(P_z^0)^2}{2\Sigma^2}\right) \simeq 1$, then

$$\log\left(\frac{\rho_1}{\rho_2}\right) \underset{\mathscr{M}_2}{\overset{\mathscr{M}_1}{\lessgtr}} \frac{1}{n} \log\left(\Sigma N \sqrt{\frac{\pi}{2}}\right) = \gamma.$$

Note that ΣN is the ratio of the normalizing constants of the prior densities over Θ_1 and Θ_2, that increases with the number of breakpoints of $\mathscr{M}_2$. As the previous equation shows, the decision region for $\mathscr{M}_1$ increases monotonically with ΣN, explaining why the apparently uninformative prior leads to a strong bias in favor of $\mathscr{M}_1$. This biased behaviour is entirely due to the fact that we are comparing the "densities" $p_k(\cdot|Z)$ defined with respect to measures μ_k (the Lebesgue measure in both cases) over spaces of distinct dimensions ($d_1 = 3$, $d_2 = 5$). Depending on the priors chosen, this may bias the decision, in an unclear manner, in favor of simpler or more complex models. We stress that these remarks do not concern penalized likelihood methodologies globally, but only their interpretation as Bayesian MAP estimators.

4. BIC AND BAYESIAN MODEL SELECTION

In this section, we compare BIC to the two-step BMS/MAP semi-parametric identification described in section 2 for curve modeling with free-knot (cubic) splines. We begin with a brief description of the model and of numerical issues related to its optimization, presenting our comparative study in a second step.

4.1 Free-knots spline model

We assume that the observations follow a normal model

$$p(Z|\theta,\mathscr{M}_k) = \mathscr{N}\left(f(t;\theta), \sigma^2 I\right), \quad f(t;\theta) = \sum_{i=1}^{k} \beta^i b_i(t,\xi_k) \ . \tag{16}$$

where k is the number of knots, $b_i(t,\xi_k)$ is the i^{th} B-Spline function, $\xi_k \in [0\,1]^k$ is the (ordered) knots vector, and $\beta_k \in \mathbb{R}^{2k}$ is the vector of control points. We refer to [5] for details about splines. The parameter vector of $\mathscr{M}_k$ is $\theta_k = (\xi_k, \beta_k, \sigma^2)$.

4.2 Implementation

Maximization of the likelihood allows analytic determination of the estimates of β_k and σ^2 for a fixed model order k. Using the reduced likelihood at this estimated values as the target distribution of the SA algorithm, we can find the ML estimate of the knot vector ξ_k. Temperature T_i is initialized at $T_0 = 50$ and decreases every 500 iterations by a factor of 0.2. The maximum number of iterations is fixed to 10000. We thus obtain the ML estimate of the parameter vector (with k fixed). Adding the BIC penalty term to the likelihood and maximizing the sum with respect to k allows determination of the BIC-penalized estimate of k.

This scheme is adapted to find the MAP estimates of the parameters $\theta \in \Theta_k$ for fixed k. In this case, the target distribution of SA algorithm is the posterior $p(\theta|Z,\mathscr{M}_k)$. We use the factored prior already proposed for this problem in [6], except for the prior over k which we consider uniform, establishing thus no prior preference for simpler models:

$$\pi(\theta_k) = \pi(\beta_k|\mathscr{M}_k,\xi_k,\sigma^2)\,\pi(\xi_k|\mathscr{M}_k)\,\pi(\mathscr{M}_k)\,\pi(\sigma^2).$$

$\pi(\mathscr{M}_k) = \mathscr{U}(k \in [k_{\min}, k_{\max}])$; $\pi(\xi_k|\mathscr{M}_k) \sim \mathscr{U}([0\ 1]^k)$; $\pi(\beta_k|\mathscr{M}_k,\xi_k,\sigma^2) = \mathscr{N}\left(0, \sigma^2 N (B^T B)^{-1}\right)$ where $B = B_{\hat{k},\xi}$ is the spline design matrix with entries $b_i(t,\xi_k)$; $\pi(\sigma^2) = 1/\sigma^2$.

As described in section 2, we first identify $\mathscr{M}_{\hat{k}}$, using RJMCMC to estimate $Pr(\mathscr{M}_k|Z)$. The proposal distribution $q(\theta'|\theta)$ is classically taken as a mixture of basic transition laws that allow "jumps" between families of models: birth (b), death (d) and change (c) of a knot point in the knot vector ξ_k. With these priors, holding ξ_k fixed, we can analytically find the MAP estimates of the linear coefficients β_k and of the noise variance σ^2. We obtain one sample of the parameter vector at each iteration of the RJMCMC procedure, allowing thus the numerical determination of $\hat{k}$ for the criterion (14).

Fixing $k = \hat{k}$, the second step identifies the parameters of $\mathscr{M}_{\hat{k}}$ by maximizing the "local posterior density" $v_k(\theta) = p(\theta|Z,\mathscr{M}_{\hat{k}}), \theta \in \Theta_{\hat{k}}$. Maximization with respect to $(\beta_{\hat{k}}, \sigma^2)$ can again be found analytically, see [6], enabling the definition of the "reduced posterior" $\mathscr{P}(\xi_{\hat{k}}|Z,\mathscr{M}_{\hat{k}}) = p(\xi_{\hat{k}}|Z_1^N,\mathscr{M}_{\hat{k}})p(\hat{\beta}_{\hat{k}},\hat{\sigma}^2|Z,\mathscr{M}_{\hat{k}},\xi_{\hat{k}})$.

A SA algorithm is run with $\mathscr{P}(\xi_{\hat{k}}|Z,\mathscr{M}_{\hat{k}})$ as the score function, producing a sequence of values of $\xi_{\hat{k}}$ that converge in distribution to its maximum, completely identifying a single model amongst $\mathscr{M}$. We performed $M = 10000$ iterations of the RJMCMC algorithm followed by $L = 2000$ SA iterations. Temperature T_i is initialized at $T_0 = 0.02$, and is halved every 500 iterations.

4.3 Numerical results

We first compare the two methods BIC and BMS/MAP on simulated data, with the goal of exposing the asymptotic nature of the BIC criterion, which only for very large data sets is an approximation of BMS, inheriting the problems of "naive Bayes" under its interpretation as a posterior computed for a particular "prior."

We simulated data from a spline model with $k = 13$ knots and $\sigma^2 = 0.04$ (see Fig. 2), and considered two observation sets: $\mathscr{D}_1$ with $N = 51$ data points and $\mathscr{D}_2$ with $N = 401$ data points, see Figure 2. Figure 3 summarizes the comparison of the two methods. For the shorter data set $\mathscr{D}_1$ BIC systematically chooses a 'wrong' model order $k = 12$ (Fig. 3(left)), underestimating the data complexity, while BMS/MAP correctly identifies the true value $k = 13$ (Fig. 3(right)). For the larger data set $\mathscr{D}_2$ both criteria choose the same (and correct) model order, confirming that only asymptotically BIC yields an unbiased estimate of the model complexity, while the two-step

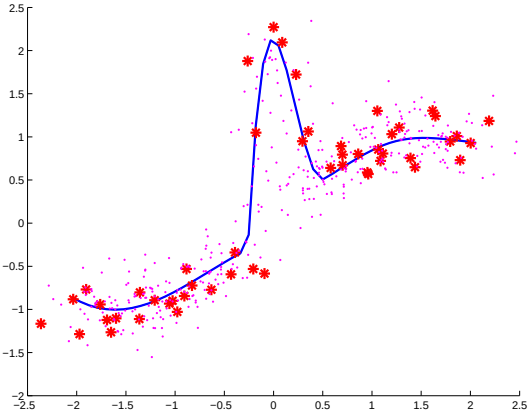


Figure 2: Spline model (-) and data sets $\mathscr{D}_1$ (*) and $\mathscr{D}_2$ (·).

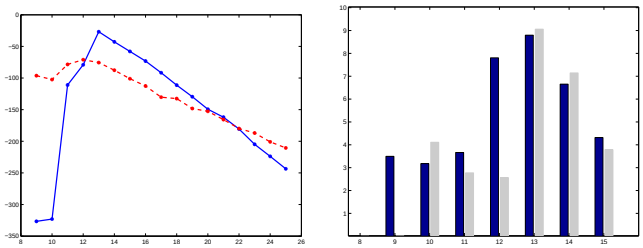


Figure 3: (left) BIC criterion for $\mathscr{D}_1$ $(--)$ and $\mathscr{D}_2$ $(-)$ depending on the model order in abscissa. (right) Log-posterior probability of models for $\mathscr{D}_1$ (black) and $\mathscr{D}_2$ (gray) with model order in abscissa.

numerical estimator BMS/MAP has an unbiased behavior for all values of N. Table 2 shows the mean square error of the models identified by the two criteria, averaged over the 50 MC runs. We note here that while producing overall similar error figures, for the larger data set $\mathscr{D}_2$ the two models have indeed virtually identical residual error, while for $\mathscr{D}_1$ BIC yields a slightly smaller error, revealing its close relation to ML.

5. CONCLUSION

This paper draws a comparative analysis of several approaches proposed in the literature for semi-parametric estimation in the context of model fitting using nested families of models. We concentrate in statistically based methodologies related to the use of penalized versions of the likelihood score and their frequent interpretations either as maximizing the model parameter "posterior distribution" or as asymptotic approximations of the marginal posterior probabilities of each model family. We stress a basic flaw of direct transposition of the MAP criterion to this complex setting as it has sometimes been done in the literature – and thus of the former interpretation of penalized likelihood – showing in a simple problem that it can lead to arbitrarily biased estimates. We then discuss the potential problems associated with asymptotic penalties, showing that the same negatively biased behavior can occur when the data set is of finite length, while computation of the true model posterior leads to unbiased estimates of the model complexity. Underestimating the model complexity for smaller data sets, BIC may fail to capture important structure of the data. As our example shows, even under a uniform prior over the models, use of the correct Bayesian criterion is not subject to the overfitting behavior typical of likelihood scores, relieving the

	$\mathscr{D}_1$	$\mathscr{D}_2$
BIC	0.0597	0.0848
BMS/MAP	0.0620	0.0845

Table 2: Average (over 50 runs) of the mean square error between data and models.

user from the need to choose a specific penalty on the model dimension.

The discussion in this paper shows that only a BMS two-step approach provides a safe methodology for identification with the type of models considered. The price payed for this better (unbiased) performance is an increase in the computational complexity of the estimation task. For the particular type of models used in the examples presented here (free knot splines), we think that a number of useful heuristics can be defined by exploiting the locality of the basis functions. We will address this issue in future publications.

REFERENCES

[1] C. Andrieu, N. de Freitas, and A. Doucet. Reversible jump mcmc simulated annealing for neural networks. In *UAI '00: Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 11–18, San Francisco, CA, USA, 2000.

[2] C. Andrieu, N. D. Freitas, and A. Doucet. Robust full bayesian learning for radial basis networks. *Neural Comput.*, 13(10):2359–2407, 2001.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[4] S. P. Brooks, N. Friel, and R. King. Classical model selection via simulated annealing. *Journal Of The Royal Statistical Society Series B*, 65(2):503–520, 2003.

[5] C. DeBoor. *A Practical Guide to Splines*. Springer, 1978.

[6] I. DiMatteo, C. Genovese, and R. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2001.

[7] W. R. Gilks, S. Richardson, and S. D. J. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1995.

[8] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[9] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[10] M. Kliger and J. M. Francos. Map model order selection rule for 2-d sinusoids in white noise. *IEEE Transactions on Signal Processing*, 53(7):2563–2575, 2005.

[11] R. H. J. M. Otten and L. P. P. P. van Ginneken. *The Annealing algorithm*. Kluwer Academic, 1989.

[12] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[13] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[14] L. Wasserman. Bayesian model selection and model averaging. *J Math Psychol*, 44(1):92–107, March 2000.