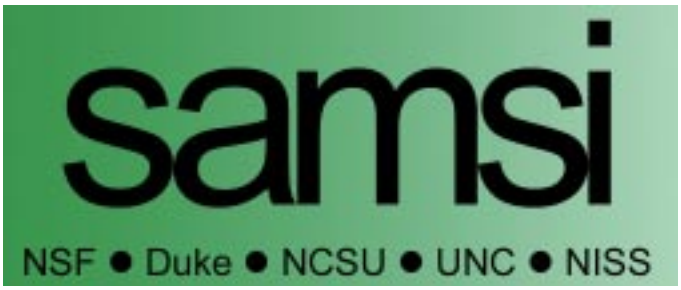


Parsimonious Function Representation and Optimal Predictive Model Selection

Ernest Fokoue

Technical Report #2004-19
August 4, 2004

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Parsimonious Function Representation and Optimal Predictive Model Selection

Ernest Fokoué*

Draft of first version

Abstract

This paper proposes an intuitively appealing approach to function approximation that yields both parsimonious functional representations and optimal predictive models. Along the lines of the *median* probability model, the concept of *prevalence* is introduced and defined in terms of the posterior model probabilities. The posterior distribution of model size is used as the main device to determine the most probable number of basis elements. The model of size equal to the optimal number of basis functions is formed using the most prevalent basis elements. The resulting model, called *prevalence* model, is shown to be optimal predictive. The *median* probability model turns out to be a specific instance the more general and more adaptive *prevalence* model. The proposed method uses the sample path obtained from the simulation of a birth-and-death process to estimate the *prevalence* of the basis elements appearing in the expansion of the approximating function. The *most probable* number of basis elements is also estimated from the same sample path. It is interesting to note that despite using simple non-sparsity inducing priors, the method yields representations that are substantially more parsimonious than the ones obtained from methods based on sparsity-inducing priors. The framework is fairly general, and naturally lends itself to both traditional basis function expansion and the now popular kernel expansion. Several examples are provided to illustrate the performance of the scheme, along with comparisons to existing methods such as the support vector machine and the relevance vector machine.

Keywords: *Parsimonious Function Representation, Prevalence, Prediction, Sparsity, Birth-and-death process, Prior specification, Bayesian basis selection*

*Ernest Fokoué is Assistant Professor, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA. (**eMail:** epf@stat.ohio-state.edu). He is Postdoctoral Research Fellow at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709. (epf@samsi.info)

1 Introduction

Consider an input space $\mathcal{X}$ and a response space $\mathcal{Y}$. Let f be the generic name used to denote any arbitrary function that maps an element of $\mathcal{X}$ to an element of $\mathcal{Y}$. Consider the problem of using a dataset $\mathbf{D} = \{(\mathbf{x}_i, y_i)_{i=1}^n, \quad \mathbf{x}_i \in \mathcal{X}, \quad y_i \in \mathcal{Y}\}$ to learn the functional dependencies between a response variable Y and a predictor (explanatory) variable X. The problem can either be one of classification or regression. For simplicity, this paper will use regression as the framework for describing the proposed method. Therefore, $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ will be the input space and the response space respectively. From a parametric perspective, one of the most popular approaches to this problem consists in choosing a set $\boldsymbol{\mathcal{B}} = \{B_1, B_2, \cdots, B_m\}$ of m basis functions and then expressing each function value $f(\mathbf{x}_i)$ as a weighted sum of the form

$$f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{m} \beta_j B_j(\mathbf{x}_i) \tag{1}$$

It will be assumed as usual that the response variable Y_i is corrupted by a noise term ϵ_i, ie

$$Y_i = f(\mathbf{x}_i) + \epsilon_i \tag{2}$$

And for notational convenience, the traditional linear model (3) will be used throughout.

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & B_1(\mathbf{x}_1) & B_2(\mathbf{x}_1) & \cdots & B_m(\mathbf{x}_1) \\ 1 & B_1(\mathbf{x}_2) & B_2(\mathbf{x}_2) & \cdots & B_m(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & B_1(\mathbf{x}_n) & B_2(\mathbf{x}_n) & \cdots & B_m(\mathbf{x}_n) \end{bmatrix} \tag{4}$$

$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_m)^\mathsf{T}$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\mathsf{T}$. It will be assumed throughout this paper that the ϵ_i's are independent, each with a normal distribution with mean 0 and common variance σ^2. It will be also be assumed that the design matrix $\mathbf{X}$ is full rank, with $m < n$. Following a convenient notation used in Barbieri and Berger (2004), this paper will refer to equation (3) as the *full* model, and the method proposed will consider selecting from among submodels of the form

$$M_{\boldsymbol{v}}: \quad \boldsymbol{y} = \mathbf{X}_{\boldsymbol{v}}\boldsymbol{\beta}_{\boldsymbol{v}} + \boldsymbol{\epsilon} \tag{5}$$

where $\boldsymbol{v} = (v_1, v_2, \cdots, v_m)$ is the model index, defined coordinate-wise as follows:

$$v_i = \begin{cases} 1 & \textit{if } B_i \textit{ is used by model } M_{\boldsymbol{v}} \\ 0 & \textit{otherwise} \end{cases} \tag{6}$$

The definition of model index provided by (6) is particularly appropriate when a parsimonious representation of f is sought through a sequential basis search (selection) procedure (method). However, when parsimony is sought by way of a sparsity-inducing prior, the index of the model is better defined through $\boldsymbol{\beta}$ as follows:

$$v_i = \begin{cases} 1 & \textit{if } \beta_i \textit{ is set equal to } 0 \\ 0 & \textit{otherwise} \end{cases} \tag{7}$$

In equation (5) above, $\mathbf{X}_{\boldsymbol{v}}$ contains the columns of $\mathbf{X}$ corresponding to the nonzero coordinates of $\boldsymbol{v}$, and $\boldsymbol{\beta}_{\boldsymbol{v}}$ is the corresponding vector of regression coefficients.

Definition 1.1 *For all submodels of the form* (5)*, the size of a model* $M_{\boldsymbol{v}}$ *is defined as*

$$k_{\boldsymbol{v}} = |M_{\boldsymbol{v}}| = \sum_{i=1}^{m} v_i$$

A total of $2^m - 2$ models can be formed from the *full* model in (3), with model sizes varying from 1 to $m-1$, so that together with the *full* model, they will form a model space $\mathbb{M}$ made up of $2^m - 1$ models.

Definition 1.2 *The model space* $\mathbb{M}$ *considered throughout this paper is defined as*

$$\mathbb{M} = \{M_{\boldsymbol{v}} : \quad \boldsymbol{v} \in \{0,1\}^m \text{ and } \boldsymbol{v} \neq (0,0,\cdots,0)\}$$

Under the *predictive optimality* criterion, the aim of model selection is to select from $\mathbb{M}$ the model $M_{\boldsymbol{v}^*}$ that minimizes the expected loss

$$R(M_v) = \mathbb{E}\left[\ell(y^{\mathsf{new}}, \hat{y}^{\mathsf{new}})\right]$$

where $\hat{y}^{\mathsf{new}}$ is the prediction of the response produced by model $M_{\boldsymbol{v}}$ on upon seeing $\mathbf{x}^{\mathsf{new}}$, and the loss function $\ell(y^{\mathsf{new}}, \hat{y}^{\mathsf{new}})$ will be taken to be the squared error loss

$$\ell(y^{\mathsf{new}}, \hat{y}^{\mathsf{new}}) = (y^{\mathsf{new}} - \hat{y}^{\mathsf{new}})^2 \tag{8}$$

Despite providing a detailed account against the common perception that the Bayesian optimal predictive model will be the model with the highest posterior probability, Barbieri and Berger (2004) set out to explore various scenarios in which it is possible to give the optimal predictive model solely in terms of the posterior model probabilities $\boldsymbol{\pi}(M_{\boldsymbol{v}}|\boldsymbol{y})$. Their first ingredient is the overall posterior probability that a basis element is in a model, which is defined as follows:

Definition 1.3 *The posterior inclusion probability for basis element* B_j *is*

$$p_i \equiv \sum_{\boldsymbol{v}: v_i = 1} \boldsymbol{\pi}(M_{\boldsymbol{v}}|\boldsymbol{y}) \tag{9}$$

Definition 1.4 *If it exists, the* **median probability model**, $M_{\boldsymbol{v}^*}$*, is defined to be the model consisting of those variables whose posterior inclusion probability is at least* $1/2$*, where the coordinates of* $\boldsymbol{v}^*$ *is defined as*

$$v_i^* = \begin{cases} 1 & \text{if} \quad p_i \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Besides citing an impressive literature that establishes the superiority of the median probability model over the maximum probability model, Barbieri and Berger (2004) provide ample details of the Bayesian optimality of the median probability model. It is also important to note that Barbieri and Berger (2004) stress the fact that the median probability model is computationally easier to obtain than the maximum probability model.

1.1 The Prevalence model

This paper introduces a model selection method that, like the median probability model of Barbieri and Berger (2004), expresses the optimal predictive model in terms of the posterior model probabilities. The method uses two ingredients, namely the *most probable model size* k^{opt} and

Definition 1.5 *The optimum model size k^{opt} is the most probable model size defined as*

$$k^{\mathsf{opt}} = \arg \max_{1 \leq k \leq m} \boldsymbol{\pi}(k|\boldsymbol{y}) \tag{11}$$

Subsequent sections of this paper will describe how to obtain accurate estimates of k^{opt}. More specifically, the method described in this paper uses a continuous-time birth-and-death process to search through a model space with many models of varying dimensions k, and then uses the resulting sample obtained to estimate the most probable dimension.

Definition 1.6 *The prevalent model M^{prev} is the model size k^{opt} whose basis elements are the k^{opt} basis elements with the highest posterior inclusion probabilities p_i. The coordinates of the index v^{prev} are defined as*

$$v_i^{\mathsf{prev}} = \begin{cases} 1 & \text{if} \quad p_i \in \mathcal{P}^{\mathsf{prev}} \\ 0 & \quad \text{otherwise} \end{cases} \tag{12}$$

where

$$\mathcal{P}^{\mathsf{prev}} = \{\text{set of } k^{\mathsf{opt}} \text{ largest values of } p_i\}$$

Remark: Unlike the median probability model that may not exist in some cases, the *prevalence model* always exists. In most cases, the *prevalence model* coincide with both the median probability model and the maximum probability model.

Lemma 1.1 *If $\boldsymbol{\mathcal{B}}$ is orthogonal, then the prevalence model coincide with median probability model, so that*

$$R(M^{\mathsf{prev}}) = R(M^{\mathsf{med}})$$

PROOF. The proof follows immediately from the definition of the *prevalence* model and its connection to the *median* probability model. For now, it suffices to know that when $\boldsymbol{\mathcal{B}}$ is orthogonal, the p_i's are either closer to 1 for the relevant axes or closer to 0 for the irrelevant axes. So, as the number of iterations in the simulation tends to infinity and convergence is achieved, p_i's are either 1 or 0. Now, this means that the distribution of model size will be sharply picked around the number of relevant axes which in this orthogonal case will coincide with the cardinality of the set of p_i's greater than 1/2. The simulation examples on toy problems bear this out very clearly. QED □

Theorem 1.1 *For every choice of $\boldsymbol{\mathcal{B}}$, the prevalence model is the best predictive model.*

$$R(M^{\mathsf{prev}}) \leq R(M^{\mathsf{med}})$$

PROOF. From the definitions of the two models, it is clear that the *median* probability model is just an instance of the *prevalence* model. While the *median* probability model selects its bases using a fixed threshold on the posterior *inclusion* probabilities, the *prevalence* model performs an adaptive selection driven by the *most probable* model size. When the posterior distribution of model size is unimodal as it is often the case, the *prevalence* model emerges as predictive superior. A detailed proof is not yet available, but simulations are provided later that can serve as first evidence. □

2 Priors and Posteriors for a fixed submodel

The full model described earlier is nothing but the traditional normal linear model, and has likelihood function

$$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \mathcal{N}(\boldsymbol{y}; \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \tag{13}$$

A *common* zero-mean Gaussian prior will be assumed for all the coefficients β_i, that is

$$\boldsymbol{\pi}(\beta_i|\lambda) = \mathcal{N}(\beta_i; 0, \lambda^{-1}) \tag{14}$$

where λ is the precision hyperparameter. For notational convenience, the prior for the whole vector $\boldsymbol{\beta}$ is written $\boldsymbol{\pi}(\boldsymbol{\beta}|\lambda) \equiv \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \boldsymbol{\Psi})$, where the isotropic covariance matrix is $\boldsymbol{\Psi} = \lambda^{-1}\mathbf{I}_{m+1}$. Clearly, this prior *does not put any sparsity pressure* on the parameter space. In fact, the use of such a prior corresponds to a Bayesian form of the *frequentist* method of *ridge* regression, with the *regularized* version of the objective function given in this case by

$$-\log \boldsymbol{p}(\boldsymbol{\beta}, \delta, \lambda|\boldsymbol{y}) \;\propto\; \frac{\delta}{2}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{15}$$

where $\delta \equiv \sigma^{-2}$. Although the prior in (14) is not *per se* a sparsity-inducing prior, estimates of λ still do provide some ingredients for the interpretation of the overall characteristics of the coefficients β_i: "Large" values of λ will at least convey the idea that many of the coefficients are of "negligible" magnitude, and might be set equal to 0. However, λ will *not* be used that way in the method described in this paper. Instead, a Gamma hyperprior will be used for λ,

$$\boldsymbol{\pi}(\lambda \,|\, c, d) = \mathsf{Gamma}(\lambda; c, d), \tag{16}$$

and the corresponding full posterior of λ when the current model is $M_{\boldsymbol{v}}$ is given by

$$[\lambda|\cdots] \sim \mathsf{Gamma}\left(c + k_{\boldsymbol{v}} + 1, d + \frac{1}{2}\|\beta_{\boldsymbol{v}}\|^2\right) \tag{17}$$

Since the interpretation of λ is more readily tied to the full model than to its submodels, it might be sensible to use an empirical Bayes method to estimate λ from the full model, and then use that estimate throughout the search among models of varying dimensions. A similar reasoning has been used by Barbieri and Berger (2004) in a similar context of model selection. However, this paper does not use such a strategy and instead adopts a fully Bayesian approach through sampling from the posterior.

A standard treatment of the noise variance σ^2 is adopted in this context. More specifically, an inverse Gamma prior is used, or equivalently $\boldsymbol{\pi}(\sigma^{-2} \,|\, a, b) = \mathsf{Gamma}(\sigma^{-2}; a, b)$, and the corresponding full posterior is given by

$$[\sigma^{-2}|\cdots] \sim \mathsf{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{v}}\boldsymbol{\beta}_{\boldsymbol{v}}\|^2\right) \tag{18}$$

Since accurate estimates, of σ^2 can be obtained for large samples, it is assumed through this paper that σ^2 is *known* and *fixed*. However, in cases where there are reasons not to have readily available values of σ^2 or accurate estimates therefore, the full posterior of σ^2 may be added to the overall sampling scheme, and that is precisely the approach adopted in this paper.

Finally, it is straightforward to show that the full posterior for $\boldsymbol{\beta}_{\boldsymbol{v}}$ is given by

$$[\boldsymbol{\beta}_{\boldsymbol{v}}|\cdots] \sim \mathcal{N}_{k_{\boldsymbol{v}}+1}\left(\left[\mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}\mathbf{X}_{\boldsymbol{v}} + \sigma^2\boldsymbol{\Psi}^{-1}\right]^{-1}\mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}\boldsymbol{y}, \left[\sigma^{-2}\mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}\mathbf{X}_{\boldsymbol{v}} + \boldsymbol{\Psi}^{-1}\right]^{-1}\right) \tag{19}$$

3 Construction of the prevalence model

So far, all the derivations made have implicitly assumed that the size $k_{\boldsymbol{v}}$ of model $M_{\boldsymbol{v}}$ was known and fixed. In other words, the prior $\boldsymbol{\pi}(\boldsymbol{\beta}_{\boldsymbol{v}}, \sigma^2, \lambda)$ should be rigorously written as $\boldsymbol{\pi}(\boldsymbol{\beta}_{\boldsymbol{v}}, \sigma^2, \lambda|k_{\boldsymbol{v}})$. Since the present method considers searching among models of varying sizes, the distribution of $k_{\boldsymbol{v}}$ needs to be specified. For notational convenience, the model index subscript $\boldsymbol{v}$ will be dropped from both $k_{\boldsymbol{v}}$ and $\boldsymbol{\beta}_{\boldsymbol{v}}$. The prior that reflects this variation in k is therefore

$$\boldsymbol{\pi}(\boldsymbol{\beta}, \sigma^2, \lambda, k) = \boldsymbol{\pi}(k)\boldsymbol{\pi}(\boldsymbol{\beta}, \sigma^2, \lambda \,|\, k) \tag{20}$$

In expression (20), the prior $\boldsymbol{\pi}(\boldsymbol{\beta}, \sigma^2, \lambda \,|\, k)$ has already been discussed extensively. As far as the prior distribution for k is concerned, a good candidate is a Poisson distribution truncated at the right end by the maximum number m of basis elements, namely

$$\boldsymbol{\pi}(k) \propto \frac{\omega^k}{k!}e^{-\omega} \quad for \quad k = 1, \cdots, m \tag{21}$$

where ω is the hyperparameter. In this context, a uniform prior for k over $\{1, \cdots, m\}$, ie $k \sim \mathsf{Uniform}(1, \cdots, m)$ turns out to be a reasonable alternative, but we do not use it here.

Finally, besides the important issue of model size addressed by the distribution of k, there is the equally important issue of selection of the most probable basis elements from $\boldsymbol{\mathcal{B}}$. In fact, for one given value of k, there are $m!/k!(m-k)!$ ways select the k basis elements that form a model of size k. The prior specified in expression (20) needs to be extended to reflect this variability in the atoms that form a model of certain size. This paper uses the Bayesian approach to address this model uncertainty. However, unlike the very widely used approach that attempts to a complete model at each iteration of the sequence, this paper will maintain a set of good candidates throughout the process and use a stochastic birth and death process to update it.

3.1 Building block of the proposed method

More specifically, the proposed method proceeds by defining the set $\boldsymbol{\mathcal{I}} = \{1, 2, \cdots, m\}$ of indices of the elements of $\boldsymbol{\mathcal{B}}$, and then splitting the set $\boldsymbol{\mathcal{I}}$ into two *disjoint* subsets $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{D}}$:

- $\boldsymbol{\mathcal{A}} \equiv \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_k\}$, where $\boldsymbol{a}_i \in \{1, 2, \cdots, m\}$. The elements of $\boldsymbol{\mathcal{A}}$ are called the *active* elements, because they are the ones selected by the current submodel.

- The complement of $\boldsymbol{\mathcal{A}}$ is $\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{I}} \backslash \boldsymbol{\mathcal{A}}$, and contains the so-called *dormant* elements, since they are unused by the current model. $\boldsymbol{\mathcal{D}} \equiv \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{m-k}\}$, where $\boldsymbol{d}_i \in \{1, 2, \cdots, m\}$.

The active set $\boldsymbol{\mathcal{A}}$ implies the model $M_{\boldsymbol{v}} = \{v_j = 1 \textit{ if } j \in \boldsymbol{\mathcal{A}} \textit{ and } v_j = 0 \textit{ otherwise}\}$. Since the dynamics of model uncertainty (variability) can be embedded in the active set $\boldsymbol{\mathcal{A}}$, a prior will me put on $\boldsymbol{\mathcal{A}}$ instead of M, and the proposed method will seek to form the posterior for $\boldsymbol{\mathcal{A}}$. Traditionally, it is assumed that all models in $\mathbb{M} = 2^{\boldsymbol{\mathcal{B}}} \backslash \{\emptyset\}$ are equally likely and as a result, a *uniform* prior over the set of all $2^m - 1$ possible models is commonly used.

$$\boldsymbol{\pi}(M) = \mathbf{const} \qquad \forall M \in \mathbb{M} \tag{22}$$

However, such uniform priors have been seen to lead to misleading conclusions in situations such as the existence of highly correlated explanatory variables in linear models. In this paper, the prior over the active set $\boldsymbol{\mathcal{A}}$ is chosen to be conditional on the size k, namely

$$\boldsymbol{\pi}(\boldsymbol{\mathcal{A}}|k) = \mathbf{const} \qquad \forall \boldsymbol{\mathcal{A}} \in 2^{\boldsymbol{\mathcal{I}}} \backslash \{\emptyset\} \tag{23}$$

By assuming *regional (local)* uniformity instead of *global* uniformity of the prior, one hopes to counter the effect of regions of high redundancy.

From all the above, the overall prior to be used for the derivation of the proposed method is

$$\boldsymbol{\pi}(k, \boldsymbol{\mathcal{A}}, \boldsymbol{\beta}, \sigma^2, \lambda) = \boldsymbol{\pi}(k)\boldsymbol{\pi}(\boldsymbol{\mathcal{A}}|k)\boldsymbol{\pi}(\boldsymbol{\beta}, \sigma^2, \lambda \,|\, k, \boldsymbol{\mathcal{A}}) \tag{24}$$

All the inferences are then based on samples from the following posterior

$$\boldsymbol{\pi}(k, \boldsymbol{\mathcal{A}}, \boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}) = \boldsymbol{\pi}(k \,|\, \boldsymbol{y})\boldsymbol{\pi}(\boldsymbol{\mathcal{A}}|k, \boldsymbol{y})\boldsymbol{\pi}(\boldsymbol{\beta}, \sigma^2, \lambda \,|\, k, \boldsymbol{\mathcal{A}}, \boldsymbol{y}) \tag{25}$$

With $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \sigma^2, \lambda)$, samples from the posterior in (25) can be obtained using two-step scheme that can be summarized as follows:

Step 1: Birth-and-death: $(k^{(t+1)}, \boldsymbol{\mathcal{A}}^{(t+1)}) \sim \boldsymbol{\pi}(k, \boldsymbol{\mathcal{A}}|\boldsymbol{\theta}^{(t)}, \boldsymbol{y})$

Step 2: Gibbs sampling: $\boldsymbol{\theta}^{(t+1)} \sim \boldsymbol{\pi}(\boldsymbol{\theta}|k^{(t+1)}, \boldsymbol{\mathcal{A}}^{(t+1)}, \boldsymbol{y})$

The second step of the above scheme is easy and straightforward as all the full conditional posteriors for the parameters have been derived earlier in the previous sections. The first step will be described at length in the next section. In much greater details, the overall proposed method can described as follows:

Algorithm: Prevalence

Initialize $t := 0$ and $k^{(t)} := \lfloor m/10 \rfloor$

$\boldsymbol{\mathcal{A}}^{(t)} := \{$random sample of $k^{(t)}$ elements from $\{1, 2, \cdots, m\}\}$

Initialize *Posterior inclusion probabilities* $p^{(0)} := (0, 0, \cdots, 0)$

Repeat

$t := t + 1$

$\boldsymbol{\mathcal{A}}^{(t)} :=$ Birth-and-Death$(\boldsymbol{\mathcal{D}}^{(t-1)}, \boldsymbol{\mathcal{A}}^{(t-1)}, \lambda^{(t-1)}, (\sigma^2)^{(t-1)}, \boldsymbol{y})$

for $j := 1$ to $|\boldsymbol{\mathcal{B}}|$ if $j \in \boldsymbol{\mathcal{A}}^{(t)}$ then $p_j := p_j + 1$ end

$p^{(t)} := (p_1, p_2, \cdots, p_m)$

$k^{(t)} := |\boldsymbol{\mathcal{A}}^{(t)}| :=$ length$(\boldsymbol{\mathcal{A}}^{(t)})$

$\boldsymbol{\theta}^{(t)} :=$ Gibbs-sampling$(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{y})$

Until $t = T$

3.2 Definition of the birth and death process

Let $\mathcal{H}^{(t)} \equiv \{k^{(t)}, \boldsymbol{\mathcal{A}}^{(t)}, \boldsymbol{\theta}^{(t)}\} \equiv \{k^{(t)}, \boldsymbol{a}_1^{(t)}, \boldsymbol{a}_2^{(t)}, \cdots, \boldsymbol{a}_i^{(t)}, \cdots, \boldsymbol{a}_{k^{(t)}}^{(t)}, \boldsymbol{\theta}^{(t)}\}$ be a *random configuration* at time t. Besides, let $\mathcal{H}_{-i}^{(t)} = \mathcal{H}^{(t)} \backslash \{\boldsymbol{a}_i^{(t)}\}$ be the random configuration $\mathcal{H}^{(t)}$ without $\boldsymbol{a}_i^{(t)}$. Clearly, $\boldsymbol{\pi}(\mathcal{H}|\boldsymbol{y})$ is invariant under permutations of the $\boldsymbol{a}_i$'s. As a result, the sequence $\{\mathcal{H}^{(t)} : t > 0\}$ defines a *point process*, where the points being *counted* are simply the $\boldsymbol{a}_i$'s.

In practice, random configurations are drawn from such a point process by constructing the simulation of a birth-and-death process. The simulation of the type of birth-and-death process used in this paper has been extensively studied and applied in recent years, and the reader is referred to references like Stoyan et al. (1995), Geyer (1999) and Barndorff-Nielsen et al. (1999) for comprehensive coverage of applications of such sampling schemes in stochastic geometry and spatial statistics. Baddeley (1994) and van Lieshout (1994) also provide very useful insights into other aspects of such sampling schemes. Stephens (2000) provides a detailed account of his use of this type of technique in the analysis of finite mixtures. The birth and death process of interest is defined as follows: Given a configuration $\mathcal{H}$, two actions are possible, namely:

- ☐ A *birth* occurs from a birth density $\boldsymbol{b}$ such that the evocation $\boldsymbol{b}(\mathcal{H}; \boldsymbol{a})$ corresponds to adding the new element $\boldsymbol{a}$ to the current configuration $\mathcal{H}$. Clearly, such an event adds exactly one element to $\mathcal{H}$ so that the new configuration is now $\mathcal{H} \cup \{\boldsymbol{a}\}$. For each configuration $\mathcal{H}$, the birth rate is denoted by $\nu(\mathcal{H})$.
- ☐ A *death* occurs through a death density $\boldsymbol{d}$ such that $\boldsymbol{d}(\mathcal{H}; \boldsymbol{a}_i)$ removes the element $\boldsymbol{a}_i$ from $\mathcal{H}$ yielding the new configuration $\mathcal{H} \backslash \{\boldsymbol{a}_i\}$. It is important to note here the assumption that each point $\{\boldsymbol{a}_i\}$ dies independently of others according to a Poisson process with

rate $\delta_i(\mathcal{H}) \equiv \boldsymbol{d}(\mathcal{H}, \boldsymbol{a}_i)$ such that the total death rate of the configuration $\mathcal{H}$ is

$$\delta(\mathcal{H}) = \sum_{i=1}^{k} \delta_i(\mathcal{H})$$

Theorem 3.1 *The birth and the death being independent Poisson processes, the time to the next event (birth or death) is exponentially distributed with mean* $1/(\nu(\mathcal{H}) + \delta(\mathcal{H}))$.

Fact 3.1 *Since the overall rate of the birth-and-death process is equal to* $\nu(\mathcal{H}) + \delta(\mathcal{H})$*, the next event will be a birth with probability* $\nu(\mathcal{H})/(\nu(\mathcal{H}) + \delta(\mathcal{H}))$*, while the death of* $\boldsymbol{a}_i$ *will occur with probability* $\delta_i(\mathcal{H})/(\nu(\mathcal{H}) + \delta(\mathcal{H}))$.

One is therefore in the presence of a continuous-time process since the time to the next event is a continuous random variable, and, by virtue of the *memorylessness* property of the exponential distribution, one has a continuous time Markov process. In order to simulate such a continuous time process, a fixed unit of time, ρ, is defined, and a discrete-time Markov chain $\{\mathcal{H}^{(\rho)}, \mathcal{H}^{(2\rho)}, \mathcal{H}^{(3\rho)}, \cdots\}$ is constructed, and used as an approximation to the continuous-time chain $\{\mathcal{H}^{(\rho+s)} : s > 0\}$. This simply means that, at each discrete iteration ($t = 1, \cdots, T$), the birth-and-death process is run for a duration of ρ.

Preston (1976) stated sufficient conditions that the above densities b and d must satisfy for the above birth-and-death process to define an ergodic Markov chain with the desired equilibrium distribution. Preston (1976)'s work was later extended and applied by Ripley (1977), and recently adapted to the analysis of finite mixtures by Stephens (2000). The following theorem, which states the sufficient conditions that b and d must satisfy, is from Preston (1976) and Ripley (1977). A proof of its extended version as applied to finite mixtures can be found in Stephens (2000).

The following detailed-balance-like theorem shows that a birth-and-death process like the one just described can be used to generate random configurations $\mathcal{H}$ from $\boldsymbol{\pi}(\mathcal{H}|\boldsymbol{y})$ as desired.

Theorem 3.2 *If the birth density* $\boldsymbol{b}$ *and the death density* $\boldsymbol{d}$ *satisfy*

$$(k+1)\boldsymbol{d}(\mathcal{H} \cup \{\boldsymbol{a}\}; \boldsymbol{a})\boldsymbol{q}(\mathcal{H} \cup \{\boldsymbol{a}\}) = \nu(\mathcal{H})\boldsymbol{b}(\mathcal{H}; \boldsymbol{a})\boldsymbol{q}(\mathcal{H}) \tag{26}$$

for all configurations $\mathcal{H}$ *and all points* $\boldsymbol{a}$*, then the birth-and-death process defined above has* $\boldsymbol{q}(\mathcal{H}) \equiv \boldsymbol{\pi}(\mathcal{H}|\boldsymbol{y}) \equiv \boldsymbol{\pi}(k, \boldsymbol{\mathcal{A}}, \boldsymbol{\beta}, \sigma^2, \lambda \,|\, \boldsymbol{y})$ *as its stationary distribution.*

Intuitively, equation (26) means that, under the equilibrium distribution $\boldsymbol{\pi}(\cdot|\boldsymbol{y})$, transitions from $\mathcal{H}$ into $\mathcal{H} \cup \{\boldsymbol{a}\}$ are exactly matched by transitions from $\mathcal{H} \cup \{\boldsymbol{a}\}$ into $\mathcal{H}$. In a sense, (26) is a sort of detailed balance equation.

3.3 Simulation of the continuous-time birth-and-death process

Throughout this work, the simulation will operate with an overall constant birth rate $\nu(\mathcal{H}) \equiv \nu$. The death rate is easily derived from equation (26) as follows:

$$\boldsymbol{d}(\mathcal{H}; \boldsymbol{a}) = \boldsymbol{b}(\mathcal{H}; \boldsymbol{a}) \left[\frac{\nu(\mathcal{H})}{k}\right] \left[\frac{\boldsymbol{q}(\mathcal{H} \backslash \{\boldsymbol{a}\})}{\boldsymbol{q}(\mathcal{H})}\right], \tag{27}$$

Using the fact that $\boldsymbol{q}(\mathcal{H}) \equiv \boldsymbol{\pi}(k, \boldsymbol{\mathcal{A}}, \boldsymbol{\theta} \,|\, \boldsymbol{y}) \propto \boldsymbol{p}(\boldsymbol{y} \,|\, k, \boldsymbol{\mathcal{A}}, \boldsymbol{\theta}) \boldsymbol{\pi}(\boldsymbol{\mathcal{A}}|k) \boldsymbol{\pi}(k) \boldsymbol{\pi}(\boldsymbol{\theta}|k, \boldsymbol{\mathcal{A}})$, it is easy to see that with $\boldsymbol{\mathcal{A}}_{-i} \equiv \boldsymbol{\mathcal{A}} \backslash \{\boldsymbol{a}_i\}$, the death rate for element $\boldsymbol{a}_i$ $(i = 1, \cdots, m)$ is given by

$$\delta_i(\mathcal{H}) = \left[\frac{\nu}{k}\right] \left[\frac{\boldsymbol{b}(\mathcal{H}; \boldsymbol{a}_i)}{\boldsymbol{\pi}(\boldsymbol{a}_i)}\right] \left[\frac{\boldsymbol{p}(\boldsymbol{y}|k-1, \boldsymbol{\mathcal{A}}_{-i}, \boldsymbol{\theta})}{\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \boldsymbol{\theta})}\right] \left[\frac{\boldsymbol{\pi}(\boldsymbol{\theta}|k-1, \boldsymbol{\mathcal{A}}_{-i})}{\boldsymbol{\pi}(\boldsymbol{\theta}|k, \boldsymbol{\mathcal{A}})}\right] \left[\frac{\boldsymbol{\pi}(\boldsymbol{\mathcal{A}}_{-i}|k-1)}{\boldsymbol{\pi}(\boldsymbol{\mathcal{A}}|k)}\right] \left[\frac{\boldsymbol{\pi}(k-1)}{\boldsymbol{\pi}(k)}\right] \quad (28)$$

It is common in practice to let births occur from the prior, so that $\boldsymbol{b}(\mathcal{H}; \boldsymbol{a}_i) \equiv \boldsymbol{\pi}(\boldsymbol{a}_i)$. Although this may appear overly simplistic, the only main drawback is that such births might be immediately followed by deaths if the prior has little to do with the underlying process, thereby causing slow convergence. It must be noted however, that this simple choice works rather well for a wide range of problems. Another interesting fact is the availability of the marginal over the regression coefficients, namely

$$\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda) = \int \boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \boldsymbol{\beta}, \sigma^2, \lambda) \boldsymbol{\pi}(\boldsymbol{\beta}|k, \boldsymbol{\mathcal{A}}) d\boldsymbol{\beta} \quad (29)$$

which turns out to be the following Gaussian density

$$\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda) = \mathcal{N}(\boldsymbol{y}; 0, \sigma^2 \mathbf{I}_n + \lambda \mathbf{X}_{\boldsymbol{v}} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}) \quad (30)$$

As one would expect, it turns out in practice that the use of the marginal $\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda)$ in place of $\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \boldsymbol{\beta}, \sigma^2, \lambda) \boldsymbol{\pi}(\boldsymbol{\beta}|k, \boldsymbol{\mathcal{A}})$ yields faster convergence of the chain. Putting together all the above with the uniformity of the prior $\boldsymbol{\pi}(\boldsymbol{\mathcal{A}} \,|\, k)$ and the Poisson nature of $\boldsymbol{\pi}(k)$, the expression for the death rate of element i simplifies into

$$\delta_i(\mathcal{H}) = \left[\frac{\nu}{\omega}\right] \left[\frac{\boldsymbol{p}(\boldsymbol{y}|k-1, \boldsymbol{\mathcal{A}}_{-i}, \sigma^2, \lambda)}{\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda)}\right] \quad (31)$$

Finally, it is also reasonable and quite common in practice, to assume that the overall constant birth rate ν is equal to the hyperparameter ω used in the prior for k, which results in

$$\delta_i(\mathcal{H}) = \frac{\boldsymbol{p}(\boldsymbol{y}|k-1, \boldsymbol{\mathcal{A}}_{-i}, \sigma^2, \lambda)}{\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda)} \quad (32)$$

From the normality of the likelihood function, it is easy to see that

$$\delta_i(\mathcal{H}) = \frac{|\sigma^2 \mathbf{I}_n + \mathbf{X}_{\boldsymbol{v}}[\lambda \mathbf{I}_k] \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}|^{\frac{1}{2}} \exp\left[-\frac{1}{2} \boldsymbol{y}^{\mathsf{T}} (\sigma^2 \mathbf{I}_n + \mathbf{X}_{\boldsymbol{v}}[\lambda \mathbf{I}_{k-1}] \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}})^{-1} \boldsymbol{y}\right]}{|\sigma^2 \mathbf{I}_n + \mathbf{X}_{\boldsymbol{v}}[\lambda \mathbf{I}_{k-1}] \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}|^{\frac{1}{2}} \exp\left[-\frac{1}{2} \boldsymbol{y}^{\mathsf{T}} (\sigma^2 \mathbf{I}_n + \mathbf{X}_{\boldsymbol{v}}[\lambda \mathbf{I}_k] \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}})^{-1} \boldsymbol{y}\right]} \quad (33)$$

The bulk of the computations lies in the calculation of both the determinant and the inverse of the $n \times n$ matrix $\Omega = \sigma^2 \mathbf{I}_n + \lambda \mathbf{X}_{\boldsymbol{v}} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}$. From the well-known matrix inversion lemma,

$$(\sigma^2 \mathbf{I}_n + \lambda \mathbf{X}_{\boldsymbol{v}} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}})^{-1} = \sigma^{-2} \mathbf{I}_n - \sigma^{-2} \mathbf{X}_{\boldsymbol{v}} (\lambda^{-1} \mathbf{I}_k + \sigma^{-2} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}} \mathbf{X}_{\boldsymbol{v}})^{-1} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}} \sigma^{-2}$$

From a well known matrix determinant identity, one can write

$$|\sigma^2 \mathbf{I}_n + \lambda \mathbf{X}_{\boldsymbol{v}} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}}| = |\lambda \mathbf{I}_k| |\sigma^2 \mathbf{I}_n| |\lambda^{-1} \mathbf{I}_k + \sigma^{-2} \mathbf{X}_{\boldsymbol{v}}^{\mathsf{T}} \mathbf{X}_{\boldsymbol{v}}|$$

The above matrix identities turn $\mathcal{O}(n^3)$ algorithms into $\mathcal{O}(k^3)$ algorithms. The computational gain turns out to be very substantial, since k is always much more smaller than n.

Based on all the above ingredients, a pseudocode of the birth-and-death process is given by:

Algorithm: Birth-and-Death process

Initialize $\mathbf{time} := 0$

Repeat

Compute δ_j for $j = 1, \cdots, k$; $\quad \delta := \sum_{j=1}^{k} \delta_j$

$\mathbf{time} := \mathbf{time} + \mathsf{Exponential}(1/(\nu + \delta))$

$\mathsf{birth} := \mathsf{Bernoulli}\left(\frac{\nu}{\nu + \delta}\right)$

If $\mathsf{birth} = 1$

$\mathsf{in} := \mathsf{Uniform}(\{\boldsymbol{d}_1^{(t)}, \boldsymbol{d}_2^{(t)}, \cdots, \boldsymbol{d}_{m-k}^{(t)}\})$; $\quad k := k + 1$

$\boldsymbol{v}_{\mathsf{in}}^{(t)} := 1$; $\quad \boldsymbol{\mathcal{A}}^{(t)} := \boldsymbol{\mathcal{A}}^{(t)} \cup \{\mathsf{in}\}$; $\quad \boldsymbol{\mathcal{D}}^{(t)} := \boldsymbol{\mathcal{D}}^{(t)} \backslash \{\mathsf{in}\}$

Else

$i := \mathsf{Multinomial}(\delta_1/\delta, \cdots, \delta_k/\delta)$; $\quad \mathsf{out} := \boldsymbol{a}_i^{(t)}$ $\quad k := k - 1$

$\boldsymbol{v}_{\mathsf{out}}^{(t)} := 0$; $\quad \boldsymbol{\mathcal{A}}^{(t)} := \boldsymbol{\mathcal{A}}^{(t)} \backslash \{\mathsf{out}\}$; $\quad \boldsymbol{\mathcal{D}}^{(t)} := \boldsymbol{\mathcal{D}}^{(t)} \cup \{\mathsf{out}\}$

End;

Until $(\mathbf{time} \geq \rho)$

3.4 Possible extensions to nonparametric regression

Since the key ingredient $\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda)$ of the proposed method, does not depend on the regression coefficients $\boldsymbol{\beta}$, extending the scheme to some nonparametric settings should be done without too much extra work. For instance, if one considers Bayesian nonparametric regression with a Gaussian process prior over the function values, then the conditional likelihood is

$$\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \boldsymbol{f}, \sigma^2) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma^2 \mathbf{I}_n)$$

where $\boldsymbol{f} = (f_1, f_2, \cdots, f_n)^{\mathsf{T}}$ is the vector of function values, with $f_i = f(\mathbf{x}_i), \quad i = 1, \cdots, n$. The corresponding marginal when $\boldsymbol{f}$ is averaged out is given by

$$\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2) = \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{C}_{\boldsymbol{v}} + \sigma^2 \mathbf{I}_n)$$

where $\mathbf{C}_{\boldsymbol{v}}$ is the variance-covariance matrix driving the Gaussian process prior, with the model index $\boldsymbol{v}$ used in the very same way as earlier on to indicate which of the variables are used in forming the covariance matrix $\mathbf{C}_{\boldsymbol{v}}$. More specifically, if one chooses the Gaussian RBF kernel as the covariance function for the Gaussian process, then

$$\mathbf{C}_{\boldsymbol{v}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\alpha=1}^{m} v_\alpha \left[\frac{\mathbf{x}_i^{(\alpha)} - \mathbf{x}_j^{(\alpha)}}{r}\right]^2$$

and the Gaussian process prior is defined as $\boldsymbol{\pi}(\boldsymbol{f}|\boldsymbol{\mathcal{A}}, k, \mathbf{C}_{\boldsymbol{v}}) = \mathcal{N}(\boldsymbol{f}; \mathbf{0}, \mathbf{C}_{\boldsymbol{v}})$. It is important to remember that the use of Gaussian process prior for nonparametric regression with normal noise

is rather straightforward. Among other things, it is easy to show that the posterior predictive distribution is Gaussian with

$$\begin{aligned} \mathsf{mean}(y^{\mathsf{new}}) &= \boldsymbol{h}^{\mathsf{T}}\mathbf{C}_{\boldsymbol{v}}^{-1}\boldsymbol{y} \\ \mathsf{variance}(y^{\mathsf{new}}) &= \xi - \boldsymbol{h}^{\mathsf{T}}\mathbf{C}_{\boldsymbol{v}}^{-1}\boldsymbol{h} \end{aligned}$$

where $\xi = \mathbf{C}_{\boldsymbol{v}}(\mathbf{x}^{\mathsf{new}}, \mathbf{x}^{\mathsf{new}})$, $\boldsymbol{h} = [\mathbf{C}_{\boldsymbol{v}}(\mathbf{x}^{\mathsf{new}}, \mathbf{x}_1), \cdots, \mathbf{C}_{\boldsymbol{v}}(\mathbf{x}^{\mathsf{new}}, \mathbf{x}_n)]^{\mathsf{T}}$, and $\mathbf{C}_{\boldsymbol{v}}^{-1}$ is the $n \times n$ inverse of the covariance matrix $\mathbf{C}_{\boldsymbol{v}}$. If one assumes σ^2 known, then the Gibbs sampling updating step of the proposed scheme becomes irrelevant, and the method reduces to the simulation of the birth-and-death process. Examples used to test this possible extension will be explored in future research.

4 Applications of the proposed method

4.1 Regression via kernel expansion

The first example considered in this paper concerns function approximation via kernel expansion. In this case, the number of bases considered is equal to the sample size, ie $m = n$, and the bases themselves are similarity measures or kernels, ie $B_j = K(\cdot, \mathbf{x}_j)$. This corresponds to kernel expansion as treated by the support vector machine and the relevance vector machine. In this case, parsimony happens in dual space as opposed to input-space parsimony, and has to do with choosing from the sample $\mathbf{D}$ those $\mathbf{x}_j$'s that are *most influential* in explaining the functional dependencies between $\mathbf{x}$ and y.

4.1.1 Analysis of the sinc function

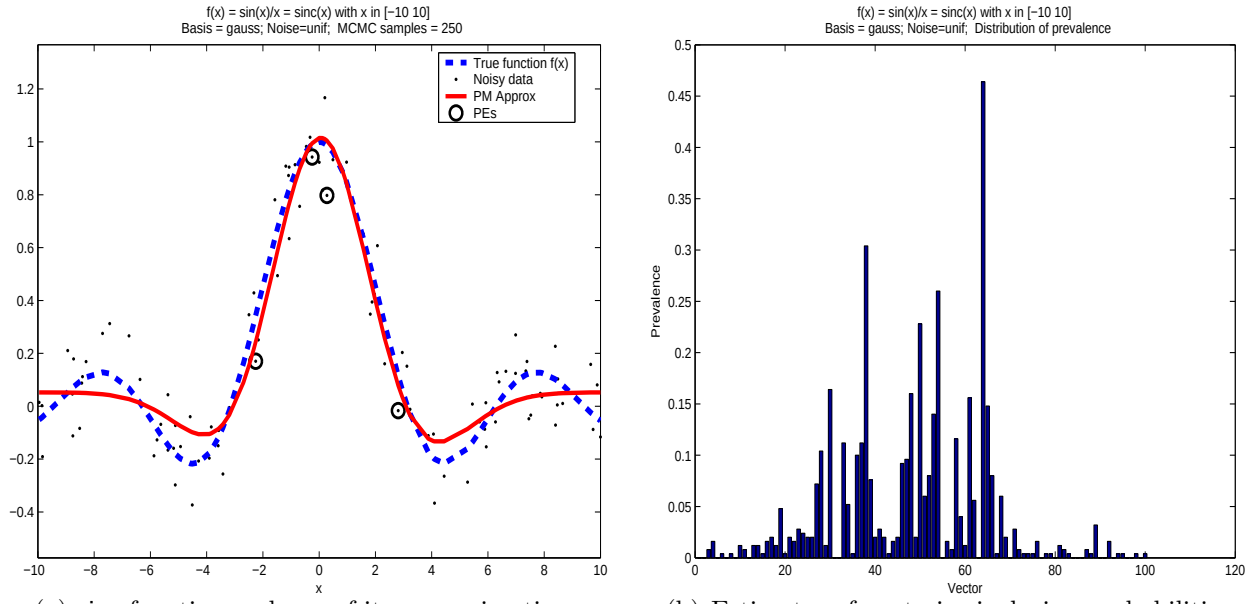
The sinc function is used here to illustrate the performance of the proposed method,

$$f(\mathbf{x}) = \mathsf{sinc}(\mathbf{x}) = \frac{\sin(\mathbf{x})}{\mathbf{x}} \quad \text{with} \quad \mathbf{x} \in [-10, 10]$$

The sinc function in and of itself is not particularly challenging, but Vapnik et al. (1997); Vapnik (1998) and Tipping (2001) have made it a popular choice now used by thousands as a benchmark example to illustrate the performance of support vector regression, relevance vector regression and related methods. The choice of the sinc function as an example allows a comparison of the proposed method to support vector machine and the relevance vector machine. For this example, 50 different sets $n = 100$ training points are generated from a uniform distribution in $[-10, 10]$, and the corresponding response values are formed as $y_i = f(\mathbf{x}_i) + \epsilon_i$ where the independent noise terms ϵ_i follows a zero mean Gaussian distribution with standard deviation 0.2. The corresponding root-mean squared errors are computed for test sets, and an estimate of the average prediction error is computed. The kernel used is the Gaussian RBF

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2r^2}\right) \tag{34}$$

In order to provide a full comparison with RVM and SVM for this sinc function example, both the uniform noise and the Gaussian noise are considered. As far as the stochastic simulation



(a) sinc function and one of its approximations (b) Estimates of posterior inclusion probabilities

Figure 1: Analysis of the sinc function via the prevalence method under Uniform noise

aspect of the method is concerned, it is important to note that the marginalization over the coefficients does allow the Markov chain to convergence very fast. The averages in the following table are based on 50 runs, and estimates at each run are on a MCMC sample path of length 250. Let **PBR** stand for *Prevalence Basis Regression*.

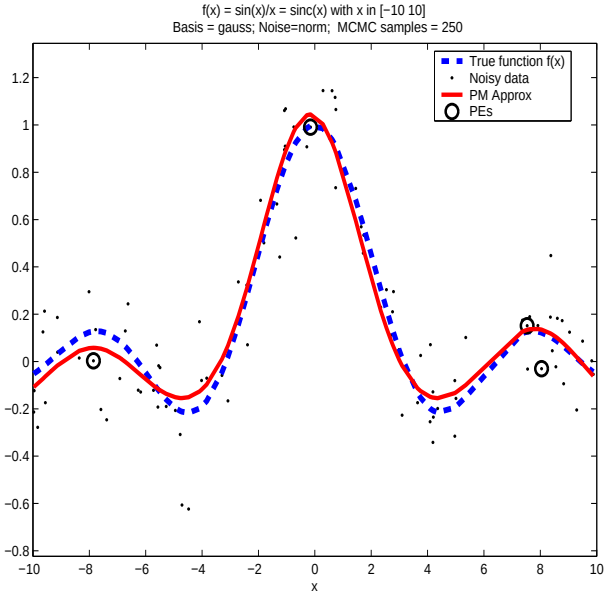
Dataset	**SVR**	**RVR**	**PBR**	**SVR**	**RVR**	**PBR**
Sinc(Gaussian noise)	0.378	0.326	0.232	45.2	6.7	3.5
Sinc(Uniform noise)	0.215	0.187	0.153	44.3	7.0	3.3

From the specification of the prior in the relevance vector machine (RVM), it is easy to anticipate that RVM regression is bound to produce a very sparse function representation. On the other hand, is both interesting, and rather intriguing to see how basis selection on a ridge prior does provide a representation that is more parsimonious that RVM. As far as the predictive superiority of the prevalence model is concerned, it is backed by theoretical results in Bayesian model selection.

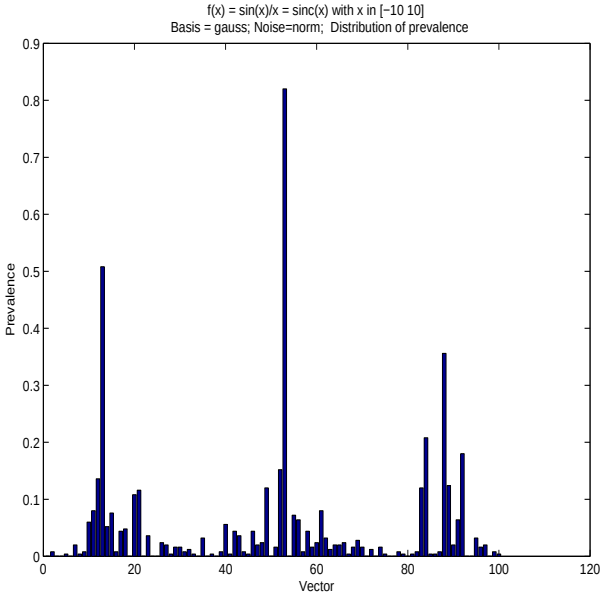
Regarding the comparison between the *prevalence* model and the *median* probability model in this context of kernel expansion, the *prevalence* model emerges as superior. Because of the nature of the problem, it is usually unlike to find bases with p_i at least equal to 1/2. This is precisely where the median probability model fails, while the adaptivity of the prevalence model allows it achieves optimal predictive model selection. The following table summarizes the comparison on the sinc function example using the RBF kernel

4.1.2 The Boston Housing Dataset

This is a very well known data set that is used for benchmarking regression methods. The whole dataset consists of 506 observations. Each $\mathbf{x}_i$ is a 13-dimensional vector of real numbers.

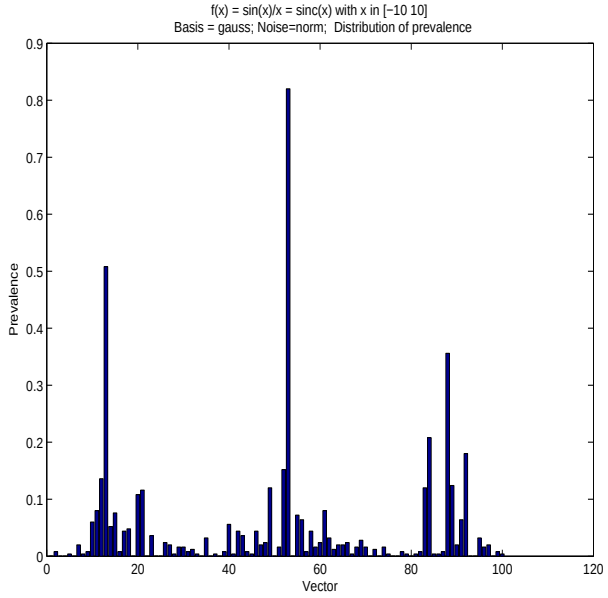


(a) sinc function and one of its approximations

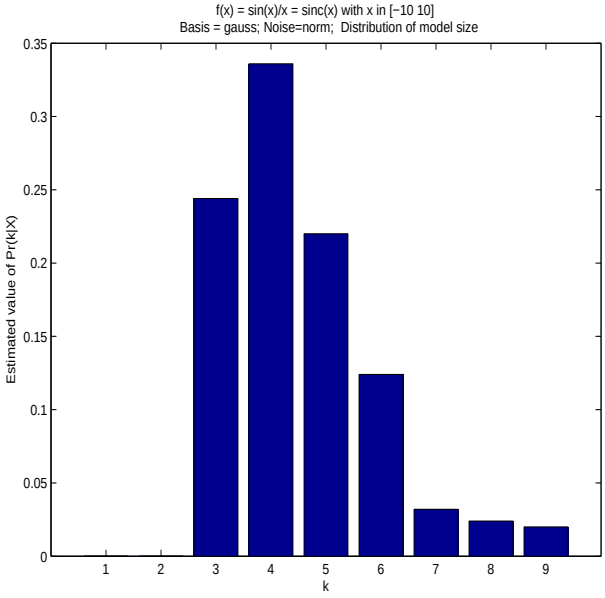


(b) Estimates of posterior inclusion probabilities

Figure 2: Analysis of the sinc function via the prevalence method under Gaussian noise



(a) Estimates of posterior inclusion probabilities



(b) The distribution of model size

Figure 3: Distribution of posterior model size in the prevalence method

Prevalence vs Median		
	Prevalence	Median
Average prediction error	0.227	0.298
Average model size	3.280	0.840

Table 1: Prevalence model vs median probability model on the sinc function

Both Support Vector Machine and Relevance Vector Machine split the dataset into $n = 481$ training samples and $m = 25$ test samples. The same thing was done in the present study on $T = 100$ random splits. One such split produced the following plots.

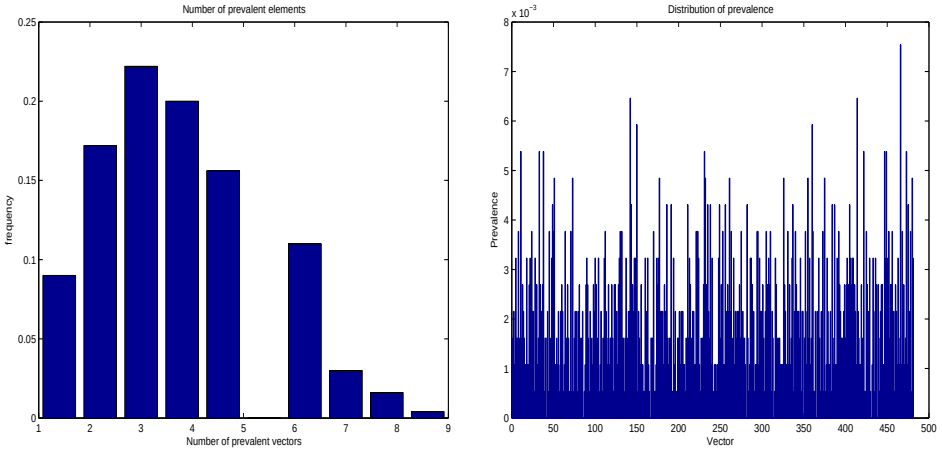


Figure 4: Performance of the Prevalence Method on the Boston Housing Data set

Dataset	SVR	RVR	PBR	SVR	RVR	PBR
Boston Housing	8.04	7.46	7.40	142.8	39	4.0

4.2 Traditional basis expansion

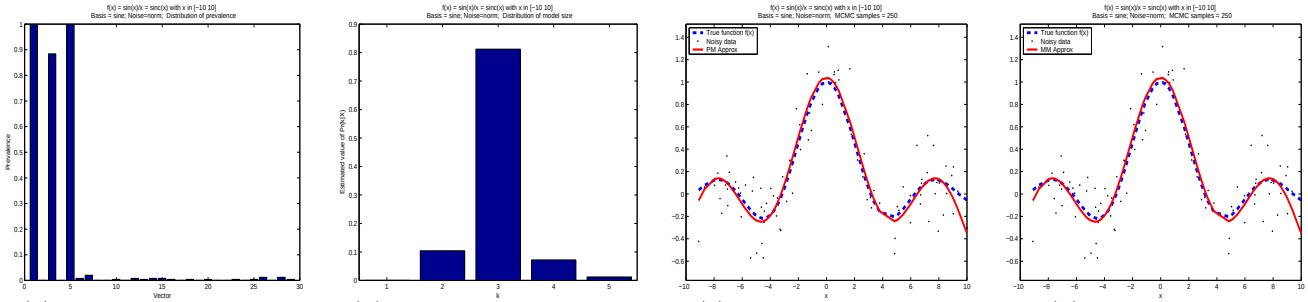
The problem here is one of basis selection, and two examples are provided to illustrate the performance of the proposed method in both basis selection and predictive accuracy. This sub-section can also be regarding as an application of the scheme to polynomial regression.

When $d = 1$ and B_j is a member of traditional basis function set like Fourier, Chebyshev, Haar wavelets, Legendre, Hermite, polynomial, etc ..., we are in the presence of a function approximation by basis expansion. Sparse representation in this case means choosing the fewest number of basis functions that achieve the best predictive accuracy.

4.2.1 What happens when the basis is orthogonal?

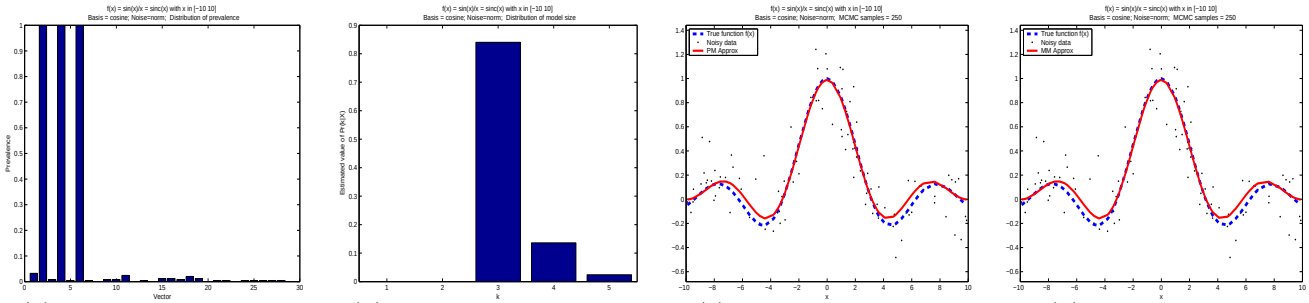
In this section, many different orthogonal basis sets are used, and it turns out that the prevalence model always produces the most optimal model.

Basis set	Prevalence size	Median Size	Prevalence Error	Median Error
Sine	2.4	2.4	0.220	0.220
Cosine	3.0	3.0	0.203	0.203
Legendre	4.2	4.0	0.197	0.196
Chebyshev	7.6	8.2	0.238	0.238



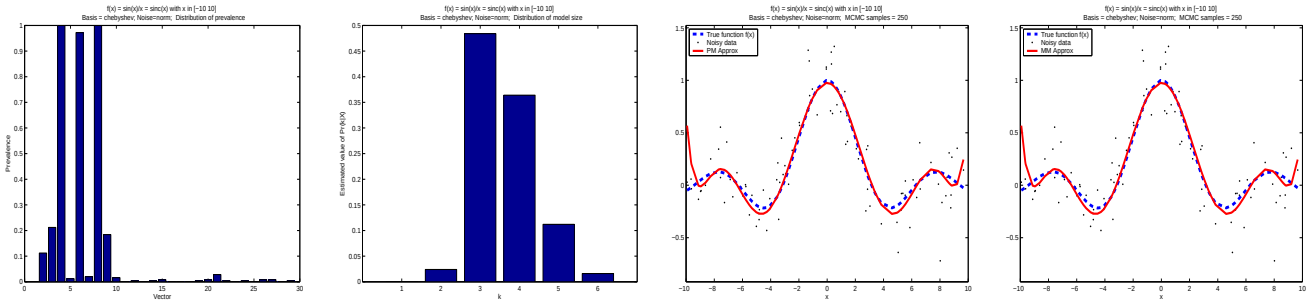
(a) Posterior inclusion (b) Distribution of size (c) Prevalence model (d) Median model

Figure 5: The sinc function, approximated in the Fourier sine basis



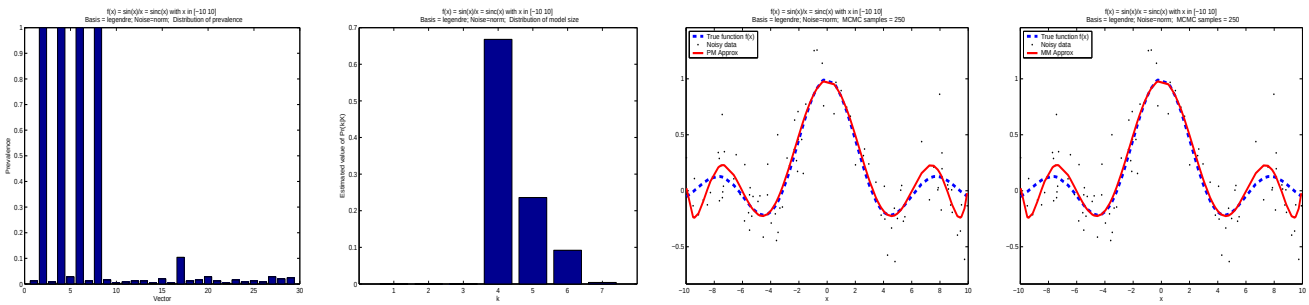
(a) Posterior inclusion (b) Distribution of size (c) Prevalence model (d) Median model

Figure 6: The sinc function, approximated in the Fourier cosine basis



(a) Posterior inclusion (b) Distribution of size (c) Prevalence model (d) Median model

Figure 7: The sinc function, approximated in the chebyshev basis



(a) Posterior inclusion (b) Distribution of size (c) Prevalence model (d) Median model

Figure 8: The sinc function, approximated in the legendre basis

4.2.2 Basis selection

Recall that the Legendre polynomials are defined on $[-1, 1]$, with $\ell_1(\mathbf{x}) = 1$, $\ell_2(\mathbf{x}) = \mathbf{x}$ and

$$\ell_i(\mathbf{x}) = \left(\frac{2i-1}{i}\right)\mathbf{x}\ell_{i-1}(\mathbf{x}) - \left(\frac{i-1}{i}\right)\ell_{i-2}(\mathbf{x}) \qquad i = 3, \cdots$$

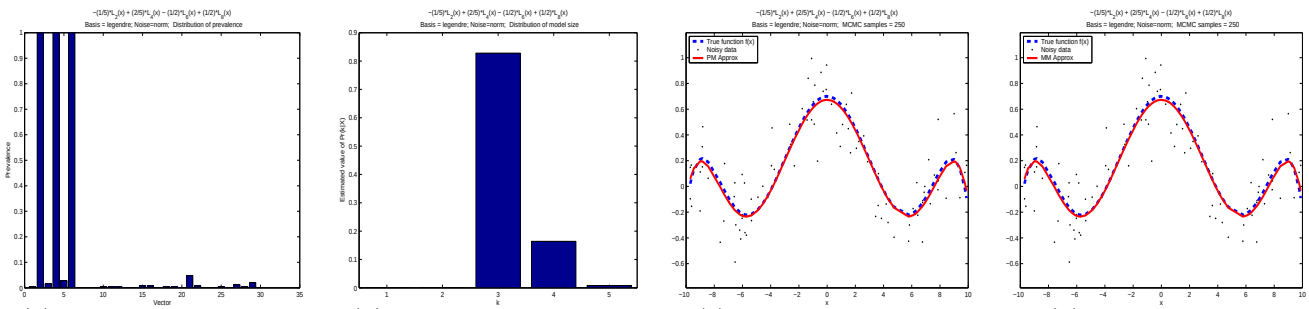
Now consider, on [-10, 10], the following function

$$f(\mathbf{x}) = -\frac{1}{5}\ell_2(\mathbf{x}) + \frac{2}{5}\ell_4(\mathbf{x}) - \frac{1}{2}\ell_6(\mathbf{x}) + \frac{1}{8}$$

For this example, the number of basis elements in the full model is $m = 30$, and the known value of k^{opt} is 3. The method is applied to the above function, using $n = 100$ for each run, and the average prediction error is estimated. As the table, and the graphs show, the method does a very job at determining the prevalence model. Besides, with the orthogonality of $\boldsymbol{\mathcal{B}}$, the prevalence model coincide with the median probability model.

Prevalence vs Median Probability		
	Prevalence	Median
Average prediction error	0.195	0.195
Average model size	3.00	3.00

Table 2: Comparison of the performances of model selection methods



(a) Posterior inclusion (b) Distribution of size (c) Prevalence model (d) Median model

Figure 9: The expanded function approximated in the legendre basis

5 Conclusion and discussion

It is important to note that traditional variable selection in multiple regression under normal noise corresponds in the proposed framework to $d = p$ and $B_j(\mathbf{x}) = x_j$. The application of the prevalence method to this problem is straightforward.

Although the inclusion probabilities p_i of the basis elements are defined in terms of the posterior weights $\boldsymbol{p}(M_{\boldsymbol{v}} \,|\, \boldsymbol{y})$ of the models, the scheme describe in this paper does not need to explicitly tabulate those posterior model weights in order to computer estimates of the p_i's. In fact, each

p_i turns out to be the fraction of time the basis element B_i appears in the active set $\boldsymbol{\mathcal{A}}$, and this is obtained directly from the sample path without any calculation.

This paper has indeed demonstrated that parsimony can be achieved even when non sparsity-inducing prior is used. The stochastic basis selection search presented in this paper is shown to produce models that are both sparse and predictively optimal. The framework presented is fairly general and is readily applicable to some of the most commonly used statistical models.

The method presented here is shown to produce results that compare favorably to many existing methods aimed at the type of problems addressed in this paper.

The method presented in this paper can be adapted to handled classification, and such an adaptation is straightforward with very minimal modifications.

In case of multimodality of $\boldsymbol{\pi}(k|\boldsymbol{y})$, one way out might be to form a model average. Such a case might be an indication of serious model uncertainty suggesting the need for model averaging or any other method that address model uncertainty.

Although $\boldsymbol{p}(\boldsymbol{y}|k, \boldsymbol{\mathcal{A}}, \sigma^2, \lambda)$ cannot be rigorously referred to as the marginal likelihood in the traditional Bayesian model uncertainty sense, it is very interesting to see from the expression of δ_i in equation (32) that the dynamics of the simulation of the continuous-time birth-and-death process uses the very same ingredients that drive the traditional methods of model comparison.

Unlike the prevalence approach, the Relevance Vector Machine (RVM) achieves a parsimonious function representation by specifying a separate independent Gaussian prior for each coefficient β_i, that is $\boldsymbol{p}(\beta_i|\lambda_i) = \mathcal{N}(0, \lambda_i^{-1})$ so that $\boldsymbol{p}(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = \mathsf{diag}(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1})$ and $\boldsymbol{\alpha} = (\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_n)^{\mathsf{T}}$. RVM is now a well established tool in machine learning. One of the greatest advantages of Prevalence over RVM is that Prevalence does not suffer from the so-called Neyman-Scott problem. In fact, since the number of hyperparameters in RVM grows linearly with the sample size, RVM cannot produce statistically consistent estimates of those hyperparameters since there is non exchangeability and therefore no dimension reduction.

One of the main weaknesses of RVM is that it relies on the estimates of the magnitude of the precision of β_j to decide which of the points to keep as relevant points. This paper claims that such magnitudes alone do provide a sufficient criterion for relevance. Examples can be constructed that reveal large magnitudes of precision with no possible sparse representation. Besides, RVM is developed only for kernel expansion and always uses all the points during the estimation procedure.

One of the main advantages of this method is that it almost never uses the totality of the m basis elements available. This is a very appealing feature from a computationally perspective as many methods are plagued by large sizes of $\boldsymbol{\mathcal{B}}$.

Finally, the prevalence model seems to always outperform the *median* probability model.

Appendix A.

References

Baddeley, A. (1994). Discussion Representation of Knowledge in Complex Systems *by Grenander & Miller. Journal of the Royal Statistical Society, Series B 56*, 584–585.

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*.

Barndorff-Nielsen, O., W. Kendall, and M. van Lieshout (1999). *Stochastic Geometry: Likelihood and Computation*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Geyer, C. (1999). Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W. Kendall, and M. van Lieshout (Eds.), *Stochastic Geometry: Likelihood and Computation*, Monographs on Statistics and Applied Probability, Chapter 3, pp. 79–140. Chapman & Hall.

Preston, C. (1976). Spatial birth-and-death process. *Bull. Inst. Internat. Statist. 46*, 371–391.

Ripley, B. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B 39*, 172–212.

Stephens, M. (2000). Bayesian Analysis of Mixtures models with an Unknown Number of Components - An alternative to Reversibe jump methods. *Annals of Statistics 28*, 40–74.

Stoyan, D., W. Kendall, and J. Mecke (1995). *Stochastic Geometry and its Applications* (second ed.). Wiley Series in Probability and Statistics. John Wiley & Sons.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research 1*, 211–244.

van Lieshout, M. (1994). Discussion on *representation of knowledge in complex systems by grenander & miller. Journal of the Royal Statistical Society, Series B 56*, 585.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Vapnik, V. N., S. E. Golowich, and A. J. Smola (1997). *Support Vector method for function approximation, regression estimation and signal processing*. In M. I. J. M. C. Mozer and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, Number 9. MIT Press.