Statistical Methodology 8 (2011) 83–96

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Fully Bayesian analysis of the relevance vector machine with an extended hierarchical prior structure

Ernest Fokoué [a,*,1], Dongchu Sun [b,2], Prem Goel [c]

[a] Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, NY 14623, USA
[b] Department of Statistics, University of Missouri, 134C Middlebush Hall, Columbia, MO 65211, USA
[c] Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 21 February 2009
Received in revised form
19 May 2010
Accepted 20 May 2010

Keywords:
Neyman–Scott problem
Improper prior
Infinite dimensional
Ill-posedness
Reference prior
Regularization
Consistency
Sparsity

ABSTRACT

This paper proposes an extended hierarchical hyperprior structure for kernel regression with the goal of solving the so-called Neyman–Scott problem inherent in the now very popular relevance vector machine (RVM). We conjecture that the proposed prior helps achieve consistent estimates of the quantities of interest, thereby overcoming a limitation of the original RVM for which the estimates of the quantities of interest are shown to be inconsistent. Unlike the majority of other authors in this area who typically use an empirical Bayes approach for RVM, we adopt a fully Bayesian approach. Our consistency claim at this stage remains only a conjecture, to be proved theoretically in a subsequent paper. However, we use a computational argument to demonstrate the merits of the proposed solution.

1. Introduction

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$ be a dataset of observed values, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Kernel regression assumes that there exists a kernel function $K(\cdot, \mathbf{x})$ such that each response random variable

* Corresponding author. Tel.: +1 8104764697.
E-mail addresses: epfeqa@rit.edu, ernest.fokoue@gmail.com (E. Fokoué), sund@missouri.edu (D. Sun), goel@stat.osu.edu (P. Goel).
1 Ernest Fokoué is Assistant Professor with the Center for Quality and Applied Statistics at Rochester Institute of Technology.
2 Dongchu Sun is Professor, with Department of Statistics at University of Missouri.

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2010.05.005

y_i can be expressed as a weighted sum of the form

$$y_i = \mathrm{v} + \sum_{j=1}^{n} \mathrm{w}_i K(\mathbf{x}_i, \mathbf{x}_j) + \epsilon_i. \tag{1}$$

The kernel used throughout this paper is the RBF (radial basis function) Gaussian kernel

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

which has been one of the most used kernels in statistical machine learning literature. Of course, there are a variety of other kernels such as the polynomial kernel, the Laplace kernel, and the spline kernel, just to name a few. For notational convenience, Eq. (1) is often rewritten as

$$\mathbf{y} = \mathrm{v}\mathbf{1}_n + \mathbf{H}\mathbf{w} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$, $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ and

$$\mathbf{H} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}. \tag{4}$$

Perhaps it is fair to say that Eq. (1) is ubiquitous in the Machine Learning community, since it has been derived from some of the fundamental methodology developed by researchers in the field. Wherever convenient, $\tilde{\mathbf{w}}^\top = (\mathrm{v}, \mathbf{w}^\top)$, and $\tilde{\mathbf{H}} = [\mathbf{1}_n\ \mathbf{H}]$ are used to write (3) as

$$\mathbf{y} = \tilde{\mathbf{H}}\tilde{\mathbf{w}} + \boldsymbol{\epsilon}. \tag{5}$$

Interestingly, although [25,26] ended up with this representation having started from a binary classification and a regression analysis formulation, respectively, it has since been used as the starting point of the analysis, not the end of it. Indeed, many think of it as an instance of a nonparametric model, although it is qualitatively different from kernel smoothing models like Naradaya–Watson to which it bears a striking resemblance.

One of the obvious things to note with this class of "models", in that their corresponding "parameter spaces" are essentially infinite dimensional, an aspect that has posed a variety of statistical and computational problems to researchers dealing with such situations. In fact, the very motivation of this paper lies in the fact that the prior used by [23] falls short in the way of achieving consistent estimation, precisely because a healthy dose of appropriate regularization is needed to achieve the convergence of the estimators to the desired values.

Infinite dimensional spaces of this sort have studied before by both Bayesians and freqentists, with some researchers exploring the theoretical properties of posterior distributions for their particular prior specifications, while others concentrated on faster computational strategies to isolate unique solutions.

On the theoretical side, recent work by [33] provides a fairly thorough theoretical account of posterior consistency in the case of the Silverman g-prior for Bayesian model selection in the context of model (1). They essentially provided a unified framework in the Bayesian setting for addressing the model selection problem under the regularized objective

$$\min_{\mathrm{v},\mathbf{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \frac{g}{2} \mathbf{w}^\top \mathbf{H} \mathbf{w} \right\}.$$

Specifically, in a typical Bayesian tradition, their theoretical analysis was based on the submodels of the so-called full model. In fact, it turns out the details of their specifications coincide with that of [12,13] who provided a fully Bayesian computational exploration of kernel regression and related models.

From a theoretical standpoint, Bayesians have been interested in the analysis of models with infinite dimensional spaces for quite a long time [1,5,18,7,8,15,14], although it must be stressed here in the words of [17], that infinite dimensional refers in this case to more than just *"simple" subsets of the coordinate space* $\mathbb{R}^{\infty}$ *but are often defined in terms of sequences of real numbers associated in some way or another to a function, such as the Fourier coefficients of the function.* In this larger sense of infinite dimensional, it is fair to recognize that there has been a lot of serious research work in the Bayesian community on the consistency of posterior distributions for a variety of prior choice. However, it is not until recently – at least to the best of our knowledge – that such research effort has been dedicated specifically to ill-posed inverse problems, of which the above kernel regression of Eq. (1) is an instance.

With the surge of interest in kernel methods and an ever increasing number of applications of reproducing Kernel Hilbert spaces techniques to nonparametric regression problems, more and more Bayesians are dedicating some effort to both theoretical and computational studies of this type of model. The most recent work we are aware of comes [11] who provide a very rigorous analysis, along with applications to stochastic volatility. In a subsequent paper, we shall explore the theoretical properties of our posterior. From a standpoint of applications, it is important to note that although the RVM has has not benefited from the type of thorough rigorous theoretical study as the ones just mentioned, it quickly established itself as one of the most widely used models in the machine learning community and far beyond.

Immediately following the publication of [23], the number of applications of the RVM approach grew steadily. Signal processing applications were some of the earliest uses of RVM with notable papers by [4], and later [28] to name a few. Environmentalists, remote sensing engineers and agricultural scientists have also extensively applied RVM in various contexts as can be seen in [32,29,2,3]. Interestingly, there have been many applications of RVM to image processing, with notable papers by [22,27,30]. Two areas of applications have recently seen a surge of interest in RVM, namely text classification with papers by [19–21,6], and microArray data analysis in the emerging field of genomics, with papers like [31].

There has also been an increase of interest in the development of extensions of RVM and its connection to other techniques [9,24]. This wave of interest prompted the research in this paper, which is organized as follows. Section 2 provides an overview of the prior structure that turns kernel regression into the relevance vector machine, and constructs a fully Bayesian treatment of RVM and shows a computational demonstration of the inherent inconsistency of the estimates obtained via RVM. Section 3 presents the details of our main results along with the derivation of all the posterior distribution of interest for use in the computation of our estimates. Section 4 explores some artificial examples. Section 5 concludes with a discussion and elements of our future work.

2. The relevance vector machine

2.1. Genesis and intuitive appeal

The relevance vector machine (RVM) introduced in [23] as a Bayesian counterpart to the popular Support Vector Machine has had tremendous success in the Machine Learning community thanks to its simplicity and applicability. Initially promoted on the strength of its counter-intuitive yet effective way of achieving a sparse representation in data space, RVM turned out to also provide very competitive performances in prediction, specifically outperforming the generalization abilities of Support Vector Regression. The original RVM paper [23] was entirely motivated by the search for a sparse functional representation of the prediction mechanism in the Bayesian learning framework, with an emphasis on the derivation of accurate yet fast predictions. The relevance vector machine (RVM) provides an empirical Bayes treatment of function approximation by kernel basis expansion.

In its original form [23], RVM achieves a sparse representation of the approximating function by structuring a Gaussian prior distribution in a way that implicitly puts a sparsity pressure on the coefficients appearing in the expansion. RVM aims at retaining the tractability of the Gaussian prior while simultaneously achieving the assumed (and desired) sparse representation. This is achieved by specifying independent Gaussian priors for each of the coefficients.

In his introductory paper, [23] shows that for such a prior structure, the use of independent Gamma hyperpriors yields a product of independent Student-t marginal prior for the coefficients, thereby achieving the desired sparsity. However, such a prior structure gives complete freedom to the coefficients, making it impossible to isolate a unique solution to the function estimation task. At the other extreme, one could think of using a single hyperparameter for all the coefficients in the spirit of traditional regularized function estimation. With such a choice, a Gaussian prior distribution over the coefficients does not yield a sparse representation. This paper aims at providing a prior structure that achieves a trade-off between the two extremes. The key idea here is to reduce the dimensionality of the hyperparameter space by specifying a prior structure that reflects the possibility of correlation between the hyperparameters of the coefficients distribution. With this, it is possible to isolate a unique solution.

2.2. Relevance vector regression

Consider the kernel regression model introduced in Eq. (1). Assume that the noise terms are independent zero-mean Gaussian random variables with the same variance σ^2, i.e. $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The likelihood function based on model (5) is then Gaussian, namely

$$p(\mathbf{y} \mid \mathrm{v}, \mathbf{w}, \sigma^2) = \mathcal{N}_n(\mathbf{y} \mid \tilde{\mathbf{H}}\tilde{\mathbf{w}}, \sigma^2 \mathbf{I}_n). \tag{6}$$

In fact, since the expansion is over all the vectors in the sample, there will be a substantial amount overfitting. Besides, such an expansion over all the vectors in the sample makes it hard to control the complexity of the derived model. An added difficulty even comes from the fact that in the above formulation, the parameter vector $\boldsymbol{\theta} = (\mathrm{v}, \mathbf{w}, \sigma^2)$ is $(n+2)$-dimensional while we only have n data points to be used for its estimation. All these difficulties put together are precisely the reason why the problem at hand belongs to the class of so-called ill-posed inverse problems for which no worthy solution can be obtained without some form of regularization. In fact, a least squares estimate of the vector $\tilde{\mathbf{w}}$ can be "written" as

$$\hat{\tilde{\mathbf{w}}}_{\mathsf{OLS}} = \left(\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}}\right)^{-1} \tilde{\mathbf{H}}^\top \mathbf{y}, \tag{7}$$

which is bound to be unstable and unreliable for all the reasons mentioned earlier. On the other hand, under a Gaussian prior for $\tilde{\mathbf{w}}$ with mean $\mathbf{0}$ and variance-matrix $\widetilde{\boldsymbol{\Psi}}$, the MAP (Maximum A Posteriori) estimate for $\tilde{\mathbf{w}}$ given $(\sigma; \widetilde{\boldsymbol{\Psi}}, \mathbf{y})$ in this case would then be

$$\hat{\tilde{\mathbf{w}}}_{\mathsf{MAP}} = \left(\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}} + \sigma^2 \widetilde{\boldsymbol{\Psi}}\right)^{-1} \tilde{\mathbf{H}}^\top \mathbf{y}. \tag{8}$$

The choice of the variance-matrix $\widetilde{\boldsymbol{\Psi}}$ will be discussed in the subsequent sections.

Now, a common and indeed reasonable assumption in kernel regression is that many of the coefficients w_i's will turn out to be zero or at best of very negligible magnitude. This is the key justification for seeking a parsimonious version of Eq. (1). The desired sparse representation is usually obtained by a careful choice of prior for $\mathbf{w}$ that reflects our belief in the fact that many of the w_i's will have negligible magnitude or even zero values with a consequence that only very few of the coefficients w_i will be needed in the expansion. It is a well known fact in the literature that the double exponential (Laplace) prior for w_i's does indeed yield such a sparse representation. However, it is also well known that the non-differentiability of the Laplace density at zero poses many computational difficulties that render its use less attractive. [10] provides an insightful account of the use of the Laplace prior in a regression analysis of the type we are considering in this paper. A natural candidate for the prior over $\mathbf{w}$ is the Gaussian distribution. However, such a prior in and of itself does not naturally yield a sparse representation. For such a prior, something else must be done to achieve sparsity. [23] specifies a separate independent zero-mean Gaussian prior for each of the w_i's, i.e.,

$$(\mathrm{w}_i \mid \alpha_i) \sim \mathcal{N}(\mathrm{w}_i \mid 0, \alpha_i^{-1}). \tag{9}$$

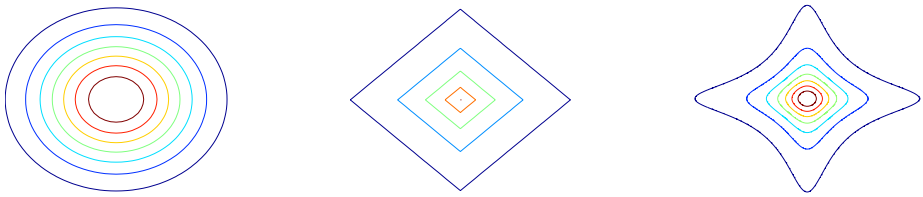

Fig. 1. Contours: (left) *Gaussian*, (center) *Lasso*, and (right) *RVM with* $(a, b) = (1, 0.25)$.

Although one would not expect a Gaussian prior to achieve sparsity, it turns out that using a Gamma hyperprior for each α_i yields a Student-t marginal prior for w_i when α_i is integrated out. In other words, with

$$(\alpha_i \mid a, b) \sim \mathsf{Ga}(\alpha_i \mid a, b), \tag{10}$$

the marginal prior for w_i is

$$p(\mathrm{w}_i) = \int p(\mathrm{w}_i \mid \alpha_i) p(\alpha_i) \mathrm{d}\alpha_i = \frac{b^a \Gamma(a + \frac{1}{2})}{(2\pi)^{\frac{1}{2}} \Gamma(a)} \left(b + \mathrm{w}_i^{2/2}\right)^{-(a+\frac{1}{2})}. \tag{11}$$

With such a marginal prior for each w_i, the prior for the vector $\mathbf{w}$ is a product of independent Student-t distributions whose density has contours that surprisingly induce sparsity. [23] uses a two-dimensional case $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2)^{\mathsf{T}}$ to show that with such a prior, the contours of the joint density of $(\mathrm{w}_1, \mathrm{w}_2)$ turn out to induce even more sparsity pressure (Fig. 1 (right)) than the amount of sparsity pressure yielded by the LASSO prior (Fig. 1 (center)).

2.3. Empirical Bayes computation of RVM

[23] provides all the details of the empirical Bayes computation of RVM for both regression and classification. In the interest of illustration, consider generating n input points from a uniform distribution in $[-1, 1]$, and then forming the corresponding response values as $y_i = f(x_i) + \epsilon_i$, where the independent noise terms ϵ_i follow a zero mean Gaussian distribution with standard deviation 0.2, i.e., $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathbf{N}(0, 0.2^2)$. Assume that the underlying true function is

$$f(x) = -x + \sqrt{2} \sin\left(\pi^{3/2} x^2\right). \tag{12}$$

The generated data pairs $\{(x_i, y_i), i = 1, \ldots, n\}$ are used as input to the relevance vector machine. The kernel used throughout this paper is the RBF Gaussian kernel

$$K(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2r^2}\right), \tag{13}$$

for which the cross-validated estimate of the length scale (bandwidth) r is found to be 0.2 for this dataset. Fig. 2 shows a fit obtained using the RVM on $n = 200$ data pairs.

2.4. Computational glimpses on RVM inconsistency

Although the above fit is as good as it gets, it is important to note that what matters here is really not the fit itself, but the consistency of the RVM as the estimator of the weights w_i used the kernel expansion. For clarity, we do not concern ourselves with the bias in this case, because we know the true function and we know that our kernel representation is just an approximation. We are therefore interested in the behavior of the variance of the RVM estimates as the same size n goes to infinity. More precisely, for the $j = 1, \ldots, k$ relevant vectors retained by the relevance vector machine, we want to know if

$$\lim_{n \to \infty} \mathsf{var}(\hat{\mathrm{w}}_j^{(n)}) \overset{?}{\to} 0.$$

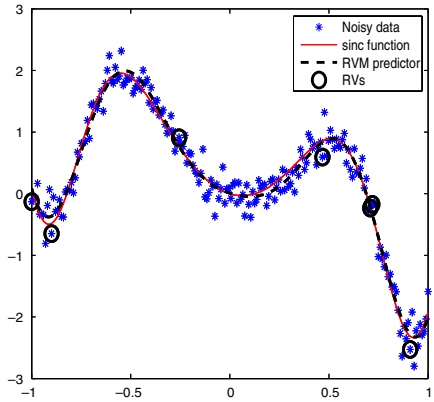

Fig. 2. Relevance vector machine fit for the above function using the RBF kernel. Fit obtained courtesy of the MATLAB code provided by [23].

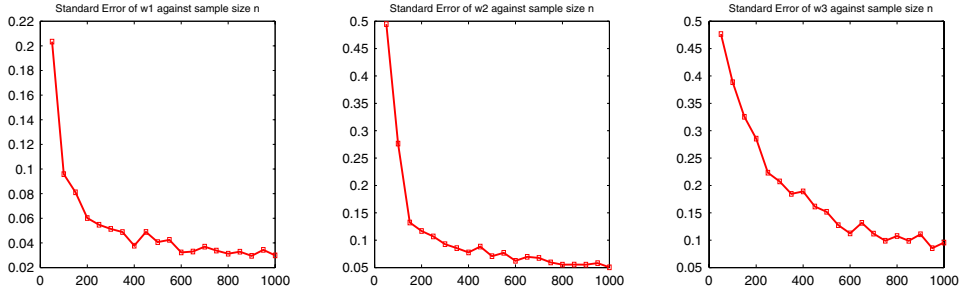

Fig. 3. Plots of the variances of the estimates of the weights: [left] $\mathrm{var}(\hat{w}_1^{(n)})$; [center] $\mathrm{var}(\hat{w}_2^{(n)})$ and [right] $\mathrm{var}(\hat{w}_3^{(n)})$, i.e. from the most relevant to the least relevant, with the x-axis representing the sample size n. As the plots show, consistency becomes more difficult to attain as the weights become less and less significant.

The plots in Fig. 3 are generated by letting the sample size n vary from $n = 100$ to $n = 1000$. For each value of n, we generate $m = 30$ different samples of size n each, which allows us to compute an estimate of the variance of each $\hat{w}_j^{(n)}$ for all the k relevant vectors retained by the relevance vector machine. The elbow for both $\mathrm{var}(\hat{w}_1^{(n)})$ and $\mathrm{var}(\hat{w}_2^{(n)})$ is attained rather quickly, and the remainder of the plot seems to flatten after that elbow. The plot for $\mathrm{var}(\hat{w}_3^{(n)})$ however remains steep throughout, indicating that there is no convergence of the variance to zero as n gets larger. As Fig. 4 shows, the least relevant the vector gets the harder it becomes for the estimators of its corresponding weights to be consistent. It might be interesting to investigate what happens if the sample size is taken to be many thousands, although such would be impractical in this context.

2.5. Fully Bayesian computation for the generic RVM

To complete our analysis, we assume an inverse gamma prior for σ^2,

$$\frac{1}{\sigma^2} \sim \mathrm{Gamma}(c_0, d_0). \tag{14}$$

Interestingly, because of the Gaussianity of the prior structure chosen for the RVM, the implementation of the fully Bayesian version of the above empirical Bayes solution is rather straightforward.

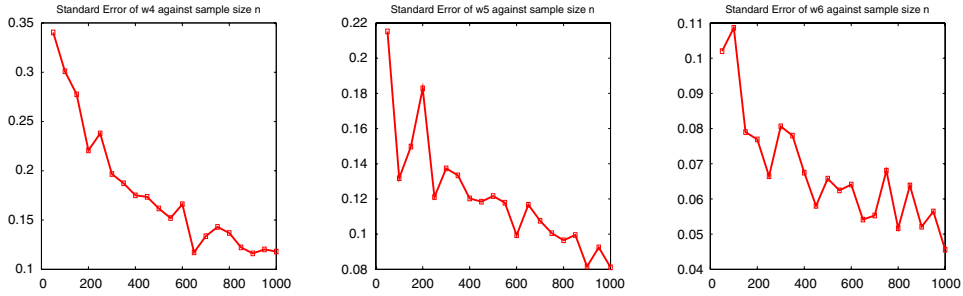

Fig. 4. Plots of the variances of the estimates of the weights: [left] $\mathrm{var}(\hat{\mathrm{w}}_4^{(n)})$; [center] $\mathrm{var}(\hat{\mathrm{w}}_5^{(n)})$ and [right] $\mathrm{var}(\hat{\mathrm{w}}_6^{(n)})$. Here, none of the variances ever stabilizes, and each remains rather very erratic, unstable, and very steep indicating that there is little hope for convergence to zero.

As a simple illustration, consider generating n input points from a uniform distribution in $[-1, +1]$, and then forming the corresponding response values as $y_i = f(x_i) + \epsilon_i$, where the independent noise terms ϵ_i follow a zero mean Gaussian distribution with standard deviation 0.2, i.e., $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 0.2^2)$. Assume that the same underlying true function f as in (12). i.e., the cross-validated estimate of the bandwidth r of the RBF Gaussian kernel is found to be 0.20 for this dataset. The following simple pseudo-code is implemented to perform the fully Bayesian analysis of the RVM on our artificial example.

Algorithm 1: Fully Bayesian Generic RVM

1. Choose (a, b) and (c_0, d_0).
2. Repeat

(a) Draw a new $\tilde{\mathbf{w}}$ according to

$$(\tilde{\mathbf{w}} \mid \text{others}) \sim \mathcal{N}_{n+1}\Big(\tilde{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{H}} + (\sigma^2\widetilde{\Psi})^{-1}\tilde{\mathbf{H}}^{\mathrm{T}}\mathbf{y},\ (\sigma^{-2}\tilde{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{H}} + \widetilde{\Psi})^{-1}\Big).$$

(b) Draw a new σ^{-2} according to

$$(\sigma^{-2} \mid \text{others}) \sim \text{Gamma}\Big(c_0 + \frac{n}{2},\ d_0 + \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{H}}\tilde{\mathbf{w}}\|^2\Big).$$

(c) Draw each new α_i according to

$$(\alpha_j \mid \text{others}) \sim \text{Gamma}\Big(a + \frac{1}{2},\ b + \frac{1}{2}\mathrm{w}_j^2\Big).$$

3. Until number of desired samples attained.

Fig. 5 shows a fit obtained using the RVM on $n = 99$ data pairs, with 5000 MCMC iterations.

In practice, it turns out that having to tune both a and b leads to a variety of problems of statistical and computational instability. Some values of b for instance lead to solutions that are almost entirely concentrated at one single point (observation). Other values require hundreds of thousands of iterations to explore a reasonable amount of the support of the function space. A slightly different prior from (9) and (10) on the weights w_j leads to the interpretation of the α_j's as measures of signal to noise ratio. Specifically, this means that the prior distribution for w_j is now

$$(\mathrm{w}_j \mid \sigma^2, \alpha_j) \sim \mathcal{N}(\mathrm{w}_j \mid 0, \alpha_j^{-1}\sigma^2), \tag{15}$$

for which the corresponding prior density is

$$p(\mathrm{w}_j \mid \sigma^2, \alpha_j) = \sqrt{\frac{\alpha_j}{2\pi\sigma^2}}\exp\Big\{-\frac{\alpha_j}{2\sigma^2}\mathrm{w}_j^2\Big\}. \tag{16}$$

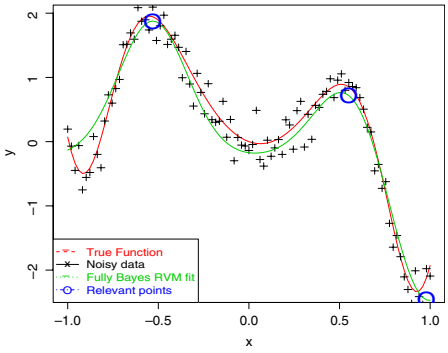

Fig. 5. RVM fit using a basic implementation of the Gibbs sampler with hyperparameters $a = 1$ and $b = 1/999$ for the Gamma distribution of the α_j's.

An alternative prior other than the usual gamma can be used for α_j. In particular, a Pareto distribution of α_j is used with the density

$$p(\alpha_j \mid c) = \frac{c}{(\alpha_j + c)^2}, \quad \alpha_j > 0. \tag{17}$$

Here $c > 0$ is a fixed constant. This prior is controlled by a single parameter c. Interestingly, it can be shown that is derived from the previously used gamma prior, simply by integrating out the rate hyperparameter b with b assumed to be gamma distributed. Indeed, if $b \sim \text{Gamma}(1, c)$, then

$$\int_0^\infty p(\alpha_j \mid a, b)p(b \mid c)\mathrm{d}b = \frac{\alpha_j^{a-1}}{\Gamma(a)} \frac{\Gamma(a+1)c}{(\alpha_j + c)^{a+1}} \overset{a=1}{=} \frac{c}{(\alpha_j + c)^2}.$$

Combining this with the prior over the weights w_j, the conditional posterior of the α_j is then of the form

$$p(\alpha_j \mid \textit{others}) \propto \frac{\sqrt{\alpha_j}}{(\alpha_j + c)^2} \mathrm{e}^{-\frac{\mathrm{w}_j^2}{2\sigma^2}\alpha_j}. \tag{18}$$

Let $\eta_j = \log \alpha_j$, so that $\alpha_j = \exp(\eta_j)$. It is easy to see that the conditional posterior density of η_j is then given by

$$p(\eta_j \mid \textit{others}) \propto \frac{\mathrm{e}^{\eta_j/2}}{(\mathrm{e}^{\eta_j} + c)^2} \exp\left(-\frac{\mathrm{w}_j^2}{2\sigma^2}\mathrm{e}^{\eta_j}\right)\mathrm{e}^{\eta_j}. \tag{19}$$

Crucially, from a computational perspective, it turns out that $p(\eta_j \mid \textit{others})$ has the desirable property of log-concavity.

Fact 1. *The conditional posterior $p(\eta_j \mid \textit{others})$ is log-concave.*

Proof. First of all, it is clear that

$$\log p(\eta_j \mid \textit{others}) = \frac{\eta_j}{2} - 2\log(\mathrm{e}^{\eta_j} + c) - \frac{\mathrm{w}_j^2}{2\sigma^2}\mathrm{e}^{\eta_j} + \eta_j + C,$$

for some constant C. Then

$$\begin{aligned}\frac{\partial}{\partial \eta_j} \log p(\eta_j \mid \textit{others}) &= \frac{3}{2} - \frac{2\mathrm{e}^{\eta_j}}{\mathrm{e}^{\eta_j} + c} - \frac{\mathrm{w}_j^2}{2\sigma^2}\mathrm{e}^{\eta_j} \\ &= \frac{3}{2} - 2\left(1 - \frac{1}{\mathrm{e}^{\eta_j} + c}\right) - \frac{\mathrm{w}_j^2}{2\sigma^2}\mathrm{e}^{\eta_j}.\end{aligned}$$

Finally,

$$\frac{\partial^2}{\partial \eta_j^2} \log p(\eta_j \mid others) = -\frac{2ce^{\eta_j}}{(e^{\eta_j}+c)^2} - \frac{w_j^2}{2\sigma^2} e^{\eta_j} < 0,$$

as desired. □

With this log-concavity, drawing samples from the posterior distribution of each η_j becomes a lot faster. Indeed, without log-concavity, the sampling is much slower as it needs to be done in a hierarchical fashion. To obtain a realization α_j from $p(\alpha_j \mid c)$, do the following:

$$\text{Draw} \quad g_j \sim \text{exponential}(c), \quad \text{then draw} \quad (\alpha_j \mid g_j) \sim \text{exponential}(g_j).$$

Unfortunately, this hierarchical approach encounters slower mixing than being able to draw directly from $p(\alpha_j \mid c)$ in one single sweep. From a practical standpoint, as stated earlier, this offers a more stable scheme since it is now driven by only one parameter c, and it is much easier to tune a single parameter than two as with the gamma prior structure. It is fair to note however that coding log-concavity samplers is not standard, at least not yet.

3. Alternative hierarchical model

3.1. Essential elements of the solution

Although Tipping's hierarchical structure ends up yielding a sparse representation, the complete freedom given to the distribution of the parameters leads to the impossibility to find a unique solution. Since the number of parameters grows with the sample size, this is a typical case of the Neyman–Scott problem [16]. In fact, the Neyman–Scott problem essentially means in our context that, with a prior like (9), the estimates of $\mathbf{w}$ are not consistent, unless one imposes some stochastic (exchangeable or partially exchangeable) structure on the coefficients w_i. If the dimension reduction is achieved via random-coefficient regression structure, one can get consistent estimates of $\mathbf{w}$, and that's precisely the main contribution of our paper, brought by way of an appropriately constructed prior structure. Indeed, the idea in this section is to add another layer in the hierarchical structure and to reparametrize in a way that reduces the dimensionality of the parameter space. Such a specification is an extension of Tipping's original work, with the advantage that it isolates a unique solution and provides a characterization of the level of sparsity through a correlation coefficient. We now specify a separate distribution for v. More specifically, we treat v as a fixed effect, and we put a constant prior on it. We now have

$$p(w_i \mid \alpha_i) = \mathcal{N}(w_i \mid 0, \alpha_i^{-1}). \tag{20}$$

Our proposed joint prior density in this case can be written as

$$p(v, \mathbf{w}, \sigma^2, \boldsymbol{\alpha}, \mu, \rho, \tau^2) = p(\mathbf{w} \mid \boldsymbol{\alpha}) p(\sigma^2) p(\boldsymbol{\alpha} \mid \mu, \rho, \tau^2) p(\mu) p(\rho) p(\tau^2), \tag{21}$$

where $p(\mathbf{w} \mid \boldsymbol{\alpha}) = \mathcal{N}_n(\mathbf{w} \mid 0, \Psi)$ is the joint prior of the weights $\mathbf{w}$, while $\Psi = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_n^{-1})$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$. Crucially, it turns out that a reparametrization of the vector $\boldsymbol{\alpha}$ leads to a more tractable way to handle our new prior structure. Let's reparametrize $\boldsymbol{\alpha}$ as $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, where $\eta_i = \log(\alpha_i)$. We assume the following prior for $\boldsymbol{\eta}$,

$$\boldsymbol{\eta} \sim \mathcal{N}_n(\mu \mathbf{1}_n, \tau^2 \boldsymbol{\Sigma}_n), \tag{22}$$

where

$$\boldsymbol{\Sigma}_n \equiv \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \rho & \vdots \\ \vdots & \rho & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = (1-\rho)\mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}. \tag{23}$$

Since the truly vital bit lies with the behavior of the α_i's, we shall start off by fixing the hyperparameters of the distribution of η_i, with the values of μ, τ^2 and ρ chosen according to explore their expected effect. First of all, it should be noted, that for plausible values of the sample size n such as $n \geq 50$, ρ will essentially remain in the interval $(0, 1)$, although our reference prior distribution rigorously requires it to be in $(-(n-1)^{-1}, 1)$. Now, with $\rho \in (0, 1)$, we can afford an interpretation of ρ as some form of mixing coefficient (weight), helping regulate the trade-off between total freedom of the α_i's and complete tightness of the α_i. Indeed, with

$$\boldsymbol{\Sigma}_n = (1-\rho)\mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}},$$

if ρ is very close to 1, then $\boldsymbol{\Sigma}_n$ is dominated by $\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}$, the result of which is extreme tightness and a return to a ridge-like situation. On the other hand, if ρ is too small (closer to 0), then the one returns to something like the original setting of essentially independent α_i's. Of course, the effect of τ^2 then has to be taken into account in both cases.

It is natural that with sparsity being one of the main goals of RVM, one would want to have moderately large values of ρ, large enough to pick up all the relevant vectors, but not so large as to make the extraction of the relevant points unstable. It should also make sense that τ^2 should be relatively large as well, since it somewhat acts as the multiplicating constant of a data-dependent g-Prior. In a sense, while ρ deals with the relative contribution of the joint effect of the information from all the α_i's, the value of τ^2 somewhat controls the magnitude of that information. A large τ^2 should therefore be expected if the information from the data dominates. Finally, a relatively small value of μ completes the picture, to achieve the kind of result intended by the underlying idea of the Relevant Vector Machine, namely many large values of α_i and very few small ones (relevant). Imagine a very large tail, but not large as to be uniform, with some support around the (small) mean. In fact, in the theoretical justification of our proposed structure and the discussion of the statistical properties of the posterior derived from it, we explore the details of the effect of all these hyperparameters.

3.2. Specification of the joint prior of $\boldsymbol{\eta}$

In order to fully specify our prior densities, the following identities hold: Let $\mathbf{A}$ be $p \times p$ invertible matrix, $\mathbf{u}$ and $\mathbf{v}$ both vectors in $\mathbb{R}^p$. The following identities hold true:

$$|\mathbf{A} + \mathbf{u}\mathbf{v}^{\mathsf{T}}| = |\mathbf{A}|(1 + \mathbf{v}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{u}), \tag{24}$$

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^{\mathsf{T}})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^{\mathsf{T}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{u}}. \tag{25}$$

Using the above identities on $\boldsymbol{\Sigma}_n$, we have

$$|\boldsymbol{\Sigma}_n| = |(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}| = (1-\rho)^n[1 + (n-1)\rho], \tag{26}$$

and

$$\boldsymbol{\Sigma}_n^{-1} = (1-\rho)^{-1}\left[\mathbf{I}_n - \frac{\rho\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}}{1 + (n-1)\rho}\right]. \tag{27}$$

Consequently, the prior density for $\boldsymbol{\eta}$ is given by

$$\begin{aligned} p(\boldsymbol{\eta} \mid \mu, \rho, \tau^2) &= \frac{1}{|2\pi\tau^2\boldsymbol{\Sigma}_n|^{1/2}} \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\eta} - \mu\mathbf{1}_n)^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\eta} - \mu\mathbf{1}_n)\right) \\ &= \frac{1}{(2\pi\tau^2)^{n/2}(1-\rho)^{n/2}[1 + (n-1)\rho]^{1/2}} \\ &\quad \times \exp\left\{-\frac{1}{2\tau^2(1-\rho)}(\boldsymbol{\eta} - \mu\mathbf{1}_n)^T\left[\mathbf{I}_n - \frac{\rho\mathbf{1}_n\mathbf{1}_n^T}{1 + (n-1)\rho}\right](\boldsymbol{\eta} - \mu\mathbf{1}_n)\right\}. \end{aligned} \tag{28}$$

We use independent conjugate priors for both μ and τ^2, namely

$$p(\mu \mid \delta) = \mathcal{N}(0, \delta^{-1}) \quad \text{and} \quad p(\tau^{-2}) = \mathsf{Ga}(c_1, d_1), \tag{29}$$

for some positive constants c_1 and d_1. We then use a reference prior for ρ, namely

$$p(\rho) = \frac{\sqrt{1+(n-1)\rho^2}}{(1-\rho)(1+(n-1)\rho)}, \tag{30}$$

where $-(n-1)^{-1} < \rho < 1$. We can also use a constant prior for ρ in the interval $(-(n-1)^{-1}, 1)$. Finally, we assume a Gamma (c_0, d_0) prior for σ^{-2}.

3.3. Derivation of full conditional posteriors

We now derive the full conditional posteriors of $(\tilde{\mathbf{w}}, \sigma^2, \boldsymbol{\eta}, \rho, \mu, \tau^2)$ given $\mathbf{y}$. Since we only added new layers to the prior, the full conditional posterior for $\tilde{\mathbf{w}}$ remains unchanged. Because $(\tilde{\mathbf{w}} \mid \sigma^2, \boldsymbol{\alpha}, \mathbf{y}) \propto \mathcal{N}(\tilde{\mathbf{H}}\tilde{\mathbf{w}}, \sigma^2\mathbf{I}_n)\mathcal{N}(0, \widetilde{\boldsymbol{\Psi}})$, where $\widetilde{\boldsymbol{\Psi}} = diag(0, \alpha_1^{-1}, \ldots, \alpha_n^{-1})$, we get

$$(\tilde{\mathbf{w}} \mid others) \sim \mathcal{N}_{n+1}\left((\tilde{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{H}} + \sigma^2\widetilde{\boldsymbol{\Psi}})^{-1}\tilde{\mathbf{H}}^{\mathrm{T}}\mathbf{y}, (\sigma^{-2}\tilde{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{H}} + \widetilde{\boldsymbol{\Psi}})^{-1}\right). \tag{31}$$

Since $[(\sigma^2)^{-1} \mid \tilde{\mathbf{w}}, \boldsymbol{\alpha}, \mathbf{y}] \propto \mathcal{N}(\tilde{\mathbf{H}}\tilde{\mathbf{w}}, \sigma^2\mathbf{I}_n)\mathrm{Ga}(c_0, d_0)$, we get

$$(\sigma^{-2} \mid others) \sim \mathrm{Ga}\left(c_0 + \frac{n}{2}, d_0 + \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{H}}\tilde{\mathbf{w}}\|^2\right).$$

We are reaching $\boldsymbol{\alpha}$ through its reparametrized version $\boldsymbol{\eta}$. Note that $\alpha_i = \mathrm{e}^{\eta_i}$. The joint posterior for $\boldsymbol{\eta}$ comes from

$$p(\boldsymbol{\eta} \mid others) \propto p(\mathbf{w} \mid \boldsymbol{\alpha})p(\boldsymbol{\eta} \mid \mu, \rho, \tau^2),$$

where

$$p(\mathbf{w} \mid \boldsymbol{\alpha}(\boldsymbol{\eta})) = \left(\prod_{i=1}^{n} \frac{\mathrm{e}^{\eta_i/2}}{\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \mathrm{e}^{\eta_i} w_i^2\right), \tag{32}$$

$$p(\boldsymbol{\eta} \mid \mu, \rho, \tau^2) \propto \exp\left\{-\frac{\sum_{i=1}^{n}(\eta_i - \mu)^2}{2\tau^2(1-\rho)} + \frac{\rho\left[\sum_{i=1}^{n}(\eta_i - \mu)\right]^2}{2\tau^2(1-\rho)[1+(n-1)\rho]}\right\}. \tag{33}$$

Fact 2. *The conditional posterior of η_i is log-concave.*

Proof. The conditional posterior of η_i is

$$[\eta_i \mid others] \propto \mathrm{e}^{\eta_i/2} \exp\left(-\frac{1}{2}\mathrm{e}^{\eta_i} w_i^2\right) \exp\left\{-\frac{(\eta_i - \mu)^2}{2\tau^2(1-\rho)} + \frac{\rho(\eta_i - \mu)^2}{2\tau^2(1-\rho)[1+(n-1)\rho]} + \frac{\rho(\eta_i - \mu)\sum_{j\neq i}(\eta_j - \mu)}{\tau^2(1-\rho)[1+(n-1)\rho]}\right\}.$$

Then for a constant C,

$$\log[\eta_i \mid others] = C + \frac{\eta_i}{2} - \frac{1}{2}\mathrm{e}^{\eta_i} w_i^2 - \frac{1+(n-2)\rho}{2\tau^2(1-\rho)[1+(n-1)\rho]}(\eta_i - \mu)^2 + \frac{\rho(\eta_i - \mu)\sum_{j\neq i}(\eta_j - \mu)}{\tau^2(1-\rho)[1+(n-1)\rho]},$$

where C is a constant. It is easy to see that

$$\frac{\partial}{\partial \eta_i}[\eta_i \mid others] = -\frac{1}{2}e^{\eta_i} w_i^2 - \frac{1+(n-2)\rho}{\tau^2(1-\rho)[1+(n-1)\rho]}. \tag{34}$$

Because $1+(n-2)\rho > 0$ for any $\rho \in (-(n-1)^{-1}, 1)$, (34) is then negative. The result holds. □

The reference prior of Eq. (30) allows us to write

$$\begin{aligned} p(\rho \mid others) &\propto p(\rho)p(\boldsymbol{\eta} \mid \mu, \rho, \tau^2) \\ &\propto \frac{p(\rho)}{(1-\rho)^{\frac{n}{2}}[1+(n-1)\rho]^{\frac{1}{2}}} \exp\left\{ -\frac{\sum_{i=1}^{n}(\eta_i-\mu)^2}{2\tau^2(1-\rho)} + \frac{\rho\left[\sum_{i=1}^{n}(\eta_i-\mu)\right]^2}{2\tau^2(1-\rho)[1+(n-1)\rho]} \right\}. \end{aligned}$$

The method of ratio of uniforms can be used to sample from this conditional posterior.

Recall that we have a constant prior for μ. Then,

$$\begin{aligned} p(\mu \mid others) &\propto p(\mu)p(\boldsymbol{\alpha} \mid \mu, \rho, \tau^2) \\ &= \exp\left\{ -\frac{\sum_{i=1}^{n}(\eta_i-\mu)^2}{2\tau^2(1-\rho)} + \frac{\rho\left[\sum_{i=1}^{n}(\eta_i-\mu)\right]^2}{2\tau^2(1-\rho)[1+(n-1)\rho]} \right\} \\ &\propto \exp\left\{ -\frac{n[1+(n-2)\rho]}{2\tau^2(1-\rho)[1+(n-1)\rho]}\left(\mu - n^{-1}\sum_{i=1}^{n}\eta_i\right)^2 \right\}. \end{aligned}$$

This is

$$p(\mu \mid others) \propto \mathcal{N}\left(n^{-1}\sum_{i=1}^{n}\eta_i, \frac{\tau^2(1-\rho)[1+(n-1)\rho]}{n[1+(n-2)\rho]} \right).$$

Using the fact that $|\tau^2\boldsymbol{\Sigma}| = (\tau^2)^n|\boldsymbol{\Sigma}|$ and $(\tau^2\boldsymbol{\Sigma})^{-1} = \tau^{-2}\boldsymbol{\Sigma}^{-1}$, it is easy to see that

$$\begin{aligned} p(\tau^{-2} \mid others) &\propto p(\tau^{-2})p(\boldsymbol{\eta} \mid \mu, \rho, \tau^2) \\ &\propto (\tau^{-2})^{c_1-1}\exp(-d_1\tau^{-2})(\tau^{-2})^{\frac{n}{2}} \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\eta}-\mu\mathbf{1}_n)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}-\mu\mathbf{1}_n) \right) \\ &= \mathrm{Ga}\left(\tau^{-2} \mid c_1 + \frac{n}{2}, d_1 + \frac{1}{2}(\boldsymbol{\eta}-\mu\mathbf{1}_n)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}-\mu\mathbf{1}_n) \right) \\ &= \mathrm{Ga}\left(\tau^{-2} \mid c_1 + \frac{n}{2}, d_1 + \frac{1}{2}\left\{ \frac{\sum_{i=1}^{n}(\eta_i-\mu)^2}{1-\rho} - \frac{\rho\left[\sum_{i=1}^{n}(\eta_i-\mu)\right]^2}{(1-\rho)[1+(n-1)\rho]} \right\} \right). \end{aligned}$$

Note that the inverses and the determinants and the inverses in the above posterior of τ^{-2} are easy to obtain from their special closed-forms.

4. Numerical simulations

To illustrate the performance of our method, let us reconsider Example 1 encountered earlier, where the underlying true function was

$$f(x) = -x + \sqrt{2}\sin\left(\pi^{3/2}x^2\right) \quad \text{with } x \in [-1, +1].$$

Fig. 6 (left) shows a fit obtained using the fully Bayesian implementation of the RVM under our prior specification. As before, $n = 99$, but now we only ran 500 MCMC iterations.

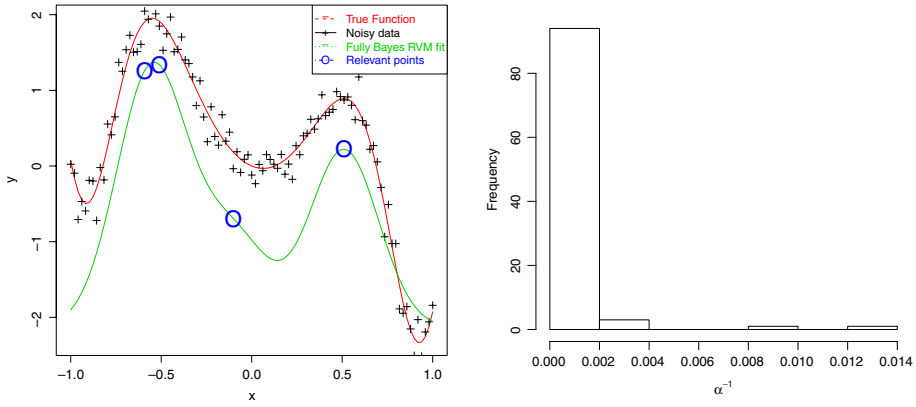

Fig. 6. Application of our proposed approach, with hyperparameters $\rho = 0.72$, $\tau = 16$, and $\mu = 1$. (left) Fit with relevant vectors; (right) histogram of the estimates of the α_i^{-1}'s.

5. Conclusion and discussion

We have proposed a novel hierarchical hyperprior structure for kernel regression that inherently has a structure capable of solving the Neyman–Scott problem suffered by the relevance vector machine.

Our contribution in this paper has been two-fold, namely the specification of a consistency inducing prior for kernel regression on the one hand, and the construction of a fully Bayesian computational scheme for estimation on the other hand. Although we have not provided a theoretical justification of our consistency conjecture in this paper, we have set up a computational framework that provides some empirical evidence as to the merit of our proposed solution. Due to time constraints, we have not provided the exhaustive computational demonstrations of the methods. However, as our plots have shown, our initial results point to the effectiveness of our proposed approach. We plan on providing thorough details of our numerical simulations in a subsequent paper. The price to pay in this case has been the computational complex inherited from the hierarchical structure. Indeed, because of the hierarchical nature of our prior structure, our computations are substantially slower than with the generic fully Bayesian treatment of RVM. To help circumvent the problem, we exploited some properties of our posterior distributions, namely the log-concavity, allowing us to somewhat explore the posterior much faster. It is fair to say however that much remains to be done in the way of computational efficiency.

Acknowledgements

The second author's research was supported by the NSF grant SES-0720229 and NIH grants R01-MH071418 and R01-CA112159.

References

[1] A. Barron, M.J. Schervish, L. Wasserman, The consistency of posterior distributions in nonparametric problems, The Annals of Statistics 27 (2) (1999) 536–561.
[2] G. Camps-Valls, et al. Relevance vector machines for sparse learning of biophysical parameters, in: Proceedings of SPIE, the International Society of Optical Engineering. Image and Signal Processing for Remote Sensing, vol. 5982, 2005, pp. 59820Z.1–59820Z.12.
[3] G. Camps-Valls, et al., Retrieval of oceanic chlorophyll concentration with relevance vector machines, Remote Sensing of Environment 105 (1) (2006) 23–33.
[4] S Chen, et al., The relevance vector machine technique for channel equalization application, IEEE Trans. Neural Netw. 20 (6) (2001) 1529–1532.
[5] T. Choi, R.V. Ramamoorthi, Remarks on consistency of posterior distributions, in: IMS Collections: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, vol. 3, IMS, 2008, pp. 170–186, doi:10.1214/074921708000000138.

[6] N. Dasgupta, et al. Relevance vector machine quantization and density function estimation: application to hmm-based multi-aspect text classification, Technical Report. Duke University, 2007.
[7] P. Diaconis, D. Freedman, On the Consistency of Bayes Estimates, The Annals of Statistics 14 (1) (1986) 1–26.
[8] P.W. Diaconis, D. Freedman, Consistency of Bayes estimates for nonparametric regression: normal theory, Bernoulli 4 (4) (1998) 411–444.
[9] A. D'Souza, et al., The bayesian backfitting relevance vector machine, in: Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.
[10] M.A.T. Figueiredo, Adaptive sparseness for supervised learning, IEEE Transactions on Pattern Analysis and Mach. Intell. 25 (1954) 1150–1159.
[11] J.P. Florens, A. Simoni, Regularized posteriors in linear ill-posed inverse problems, Technical Report. Toulouse School of Economics, Toulouse, France, 2008.
[12] E. Fokoué, Estimation of atom prevalence for optimal prediction, Contemporary Mathematics 443 (2008) 103–129.
[13] E. Fokoué, P. Goel, An optimal experimental design perspective on redial basis function regression, Technical Report. 2010. http://hdl.handle.net/1850/11694, Rochester Institute of Technology, Communication in Statistics: Theory and Methods (in press).
[14] S. Ghosal, J.K. Ghosh, Aad W. van der Vaart, Convergence rates of posterior distributions, The Annals of Statistics 28 (2) (2000) 500–531.
[15] B.J.K. Kleijn, A.W. van der Vaart, Misspecification in infinite-dimensional Bayesian statistics, The Annals of Statistics 34 (2) (2006) 837–877.
[16] J. Neyman, M. Scott, Consistent estimates based on partially consistent observations, Econometrika 16 (1948) 1–32.
[17] W. Ploberger, P.C.B. Phillips, Best empirical models when the parameter space is infinite dimensional, Technical Report, University of Rochester, Rochester, New York, USA, 2008.
[18] X. Shen, L. Wasserman, Rates of convergence of posterior distributions, Ann. Statist 29 (3) (2001) 687–714.
[19] C. Silva, B. Ribeiro, Combining active learning and relevance vector machines for text classification, in: Proceedings of the IEEE International Conference on Machine Learning Applications, 2007, pp. 130–135.
[20] C. Silva, B. Ribeiro, RVM ensemble for text classification, International Journal of Computational Intelligence Research 3 (1) (2007) 31–35.
[21] C. Silva, B. Ribeiro, Towards expanding relevance vector machines to large scale datasets, International Journal of Neural Systems 18 (1) (2008) 45–58. WSPC.
[22] A. Thayananthan, et al. Multivariate relevance vector machines for tracking, Technical Report, Cambridge University, 2008.
[23] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, Journal of Machine Learning Research 1 (2001) 211–244.
[24] S. Tripathi, R.S. Govindaraju, On Selection of kernel parameters in relevance vector machines for hydrologic applications, Stochastic Environmental Research and Risk Assessment 21 (2007) 747–764.
[25] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, 1995.
[26] G. Wahba, An introduction to model building with reproducing kernel hilbert spaces, Technical Report No. 1020, Department of Statistics, University of Wisconsin 1210 West Dayton St. Madison, WI 53706, USA, April 18, 2000.
[27] L. Wei, et al., Relevance vector machine for automatic detection of clustered microcalcifications, IEEE Transactions on Medical Imaging 24 (10) (2005) 1278–1285. http://www.ncbi.nlm.nih.gov/pubmed/16229415.
[28] R.J. Weiss, D.P.W. et Ellis, Estimating single channel source separation masks: relevance vector machine classifiers vs pitch-based masking, Technical Report, Dept. of Elec. Eng, Columbia University, New York, NY 10027, USA, 2005.
[29] D. Wipf, Srikantan Nagarajan, Beamforming using the relevance vector machine, in: Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 2007.
[30] W.S. Wong, et al. Using a sparse learning relevance vector machine in facial expression recognition, Technical Report, Man-Machine Interaction Group, Delft University of Technology, The Netherlands, 2005, eMail: L.J.M.Rothkrantz@ewi.tudelft.nl.
[31] L. Yinhai, et al. Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine, Genome Research, (2006) online, 2006.
[32] J. Yuan, et al., Integrating relevance vector machines and genetic algorithms for optimization of seed-separating process, Engineering Applications of Artificial Intelligence 20 (2007) 970–979.
[33] Z. Zhang, M.I. Jordan, D. Yeung, Posterior consistency of the silverman g-prior in bayesian model choice, Technical Report, Number xx, Department of Electrical Engineering and Computer Science, University of California, Berkeley, California, USA, 2008.