School of Engineering, Computer Science
and Mathematics,
The University of Exeter. EX4 4QF. UK.
http://www.ex.ac.uk/secam

Smooth Relevance Vector Machine: a smoothness prior extension of the RVM
(submitted draft version – do not quote)

Alexander Schmolck **Richard Everson**

January 30, 2007

Abstract

Enforcing sparsity constraints has been shown to be an effective and efficient way to obtain state-of-the-art results in regression and classification tasks. Unlike the support vector machine (SVM) the relevance vector machine (RVM) explicitly encodes the criterion of model sparsity as a prior over the model weights. However the lack of an explicit prior structure over the weight variances means that the degree of sparsity is to a large extent controlled by the choice of kernel (and kernel parameters). This can lead to severe overfitting or oversmoothing – possibly even both at the same time (e.g. for the multiscale Doppler data). We detail an efficient scheme to control sparsity in Bayesian regression by incorporating a flexible noise-dependent smoothness prior into the RVM. We present an empirical evaluation of the effects of choice of prior structure on a selection of popular data sets and elucidate the link between Bayesian wavelet shrinkage and RVM regression. Our model encompasses the original RVM as a special case, but our empirical results show that we can surpass RVM performance in terms of goodness of fit and achieved sparsity as well as computational performance in many cases. The code is freely available.

Index terms: sparse regression, kernel regression, smoothness prior, relevance vector machine.

Contents

1 Introduction

In nonlinear regression a function of interest y is approximated by a linear combination of the input vector, $\mathbf{x}$, projected onto a (typically fixed) set of nonlinear basis functions, $\{\boldsymbol{\phi}_m\}_{m=1}^M$:

$$y(\mathbf{x}) = w_0 + \sum_{m=1}^{M} w_m \boldsymbol{\phi}_m(\mathbf{x}) \tag{1}$$

Thus, provided with a set of N training input vectors $\{\mathbf{x}_n\}_{n=1}^N$ and corresponding targets t_n, the task is to find the $M+1$ weights w_m that will yield the most faithful approximation to y. For simplicity the following exposition will assume that the data is mean centered, so that we can omit the bias w_0 and work with M weights. Choosing $\boldsymbol{\phi}(\mathbf{x}) \equiv \mathbf{x}$ regains linear regression. Frequently the basis functions are derived from kernels centered at each of the observations $\boldsymbol{\phi}_m(\mathbf{x}) = K(\mathbf{x}_m, \mathbf{x})$ in which case the regression is known as *kernel regression*, however, the basis functions may be quite general functions, including for example wavelets, and may form an over-complete dictionary.

Writing the true signal, $\mathbf{y}$, as an N-vector and w_m, the weights, as an M-vector, (1) is conveniently written as $\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}$, with the basis functions arranged as the columns of the $N \times M$ design matrix $\boldsymbol{\Phi}$. Employing the standard assumption of zero-mean Gaussian noise in the target observations, we have:

$$\mathbf{t} = \mathbf{y} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N) \tag{2}$$

Demanding a sparse representation in the space spanned by a suitable set of such basis functions provides a general strategy to adjust the bias/variance trade-off in regression and classification problems, as is evinced by the state-of-the-art results achieved by support vector machines (SVMs) in a variety of domains (e.g. Schölkopf and Smola, 2002). An important additional benefit of sparsity is that it also often translates into significant computational savings.

1.1 Sparse Bayesian regression

Whilst in SVM regression a desirable level of sparsity has to be brought about indirectly by determining an error or margin parameter via a cross-validation scheme, the Bayesian formulation of the regression problem in the relevance vector machine (RVM) (Tipping, 2000, 2001; Faul and Tipping, 2002; Tipping and Faul, 2003) allows for a prior structure that explicitly encodes the desirability of sparse representations.

This is done by complementing the standard likelihood function (which follows directly from the above assumptions):

$$p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\|(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})\|^2}{2\sigma^2}\right) \tag{3}$$

with an "automatic relevance determination" prior (MacKay, 1992) over the weights:

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = (2\pi)^{-\frac{M}{2}} \prod_{m=1}^{M} \alpha_m^{\frac{1}{2}} \exp\left(-\frac{1}{2}\alpha_m w_m^2\right) \tag{4}$$

that has the effect of "switching off" basis functions for which there is no evidence in the data (more on this in sec. 2).

Whilst $p(\boldsymbol{\alpha})$ is effectively uniform[1], a standard inverse gamma prior is placed over the noise variance σ^2:

$$p(\sigma^2) = \mathcal{IG}(\sigma^2 \mid g, h) = \frac{h^g}{\Gamma(g)} \sigma^{-2(g+1)} e^{-h/\sigma^2} \tag{5}$$

where g and h are fixed hyperparameters, usually set to some uninformative value (e.g., $g = h = 10^{-4}$).

It should be stressed that in this scheme g and h are the only "true" hyperparameters, in the sense that unlike everything else that is introduced in extending the standard regression model (3) by a hierarchical prior (4), (5), they are additional parameters that require specification by the user (and in the absence of any prior information about $p(\sigma|g, h)$ just setting them to some uninformative default will work fine). All the other variables ($\boldsymbol{\alpha}$ etc.) are just nuisance parameters that can be integrated out or determined by the Bayesian approach.

Thus everything else, including ultimately $\hat{\mathbf{y}}$, the mean posterior prediction that we wish to obtain is determined by the values of $\boldsymbol{\Phi}$, $\mathbf{t}$, g and h.

This is the beauty of the Bayesian paradigm – it allows one to reap the benefits of a probabilistic approach, without burdening the model with additional externally-determined parameters (unlike the SVM, there is no need to expensively determine a regularization parameter via cross-validation and furthermore confidence intervals, likelihood values and posterior probabilities for the solution can easily be obtained).

The learning of the model parameters proceeds by an elegant type II likelihood maximization scheme in which values of $\boldsymbol{\alpha}$ and σ that maximize the log marginal likelihood $\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{t} \mid \boldsymbol{\alpha}, \sigma^2)$ are found iteratively (Faul and Tipping, 2002; Tipping and Faul, 2003). Weights w_m for which the learned precision α_m is large are effectively switched off because w_m is constrained to be close to zero.

1.1.1 Shortcomings of the classical RVM

Although the RVM carries the benefits of a probabilistic formulation, unfortunately it still does not go far enough in its Bayesian encoding of the sparsity constraint — in practice one finds that in spite of (4), the choice of highly resolving kernels for data which do not need the many degrees of freedom offered by these kernels will still result in severe overfitting. This overfitting is illustrated in Figure 1 by using a symmlet wavelet basis for regression to the Sinc data set ($N = 128$, SNR $= 2$) (Tipping, 2001). As the top-left plot shows, the multi-scale nature of the symmlet kernel results in drastic overfitting. Employing, for example, Gaussian or linear spline (lspline; pictured) basis functions[2] results in a good fit (bottom-left) and an apparently sparse solution: only 7 of the 128 available lspline basis functions are not switched off (have $\alpha_m < \infty$) compared with 127 symmlet basis functions.

Closer examination, however, reveals that the Gaussian and lspline bases, which do not contain high frequency basis functions, simply have difficulty fitting the noise. In the case of the Gaussian kernel (which with a kernel width that gives good results for the Sinc data yields a very ill-conditioned design matrix[3]), this is already apparent in the least squares estimate: choosing coefficients w_m to minimize $E = \|\mathbf{t} - \sum_m w_m \boldsymbol{\phi}_m\|$ yields a substantial error that is in

[1] A Jeffreys' prior is apparently advocated in (Tipping, 2001), but (Tipping, 2000; Faul and Tipping, 2002; Tipping and Faul, 2003) and the published implementation use a uniform prior. This is discussed further in see sec. 5.1.

[2] See sec. A.1 for details of the kernel functions; the lspline examples use $r = 3.0$.

[3] Condition number $\kappa = 6 \times 10^{18}$ for $N = 128$, $r = 3.0$.

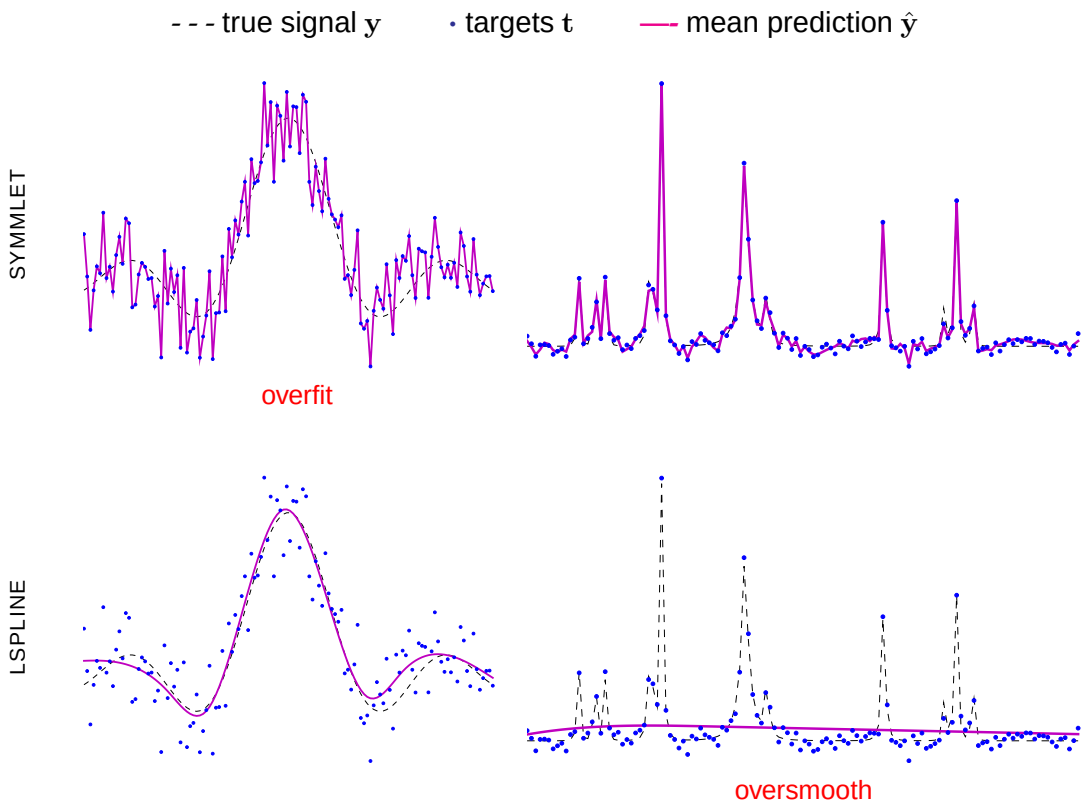


Figure 1: Classical RVM. The effect of kernel choice on the smoothness of the regression result (Sinc data *left*, Bumps data *right*) when there is no prior over $\boldsymbol{\alpha}$. Choosing a flexible symmlet-wavelet kernel (*top row*) results in drastic overfitting for the Sinc data set (*top left*; N=128, SNR=2.0). To obtain the appropriate level of smoothing for the Sinc data one has to resort to a different kernel type, such as lspline (*bottom left*). However an lspline kernel cannot resolve the Bumps data (*bottom right*; N=128, SNR=7.0) at all.

fact larger than $\|\mathbf{y} - \sum_m w_m \boldsymbol{\phi}_m\|$. The case is a bit more subtle for the lspline kernel, because unlike the Gaussian kernel, a lspline kernel with a kernel width that gives good results for (classical) RVM regression can still exactly represent $\mathbf{t}$, just as the symmlet kernel.

However, as illustrated by Figure 2, on comparing individual basis functions from both kernels, it becomes clear that whereas the symmlet basis offers multiscale resolution and hence can fit a large proportion of the noise with relatively few components, a much greater number of the exclusively low-frequency basis functions in the lspline kernel are needed to fit a comparable proportion of the noise.

Consequently the relatively mild enforcement of sparsity of the classical RVM scheme, which proves insufficient for symmlet kernels, is already enough to prevent overfitting for lspline and Gaussian kernels.

The sparse regression provided by the RVM with lspline kernels thus partially depends on a propitious choice of kernel. Although the aforementioned gauss or lspline kernels are sufficiently resolving for the Sinc data, they cannot resolve data with genuine high frequencies such as the Bumps data (Donoho and Johnstone, 1994) which is therefore severely oversmoothed (bottom-right of Figure 1), although it poses no problems for the symmlet basis (top-right).

In other words, it is apparent that a crucial aspect of sparsity control (kernel choice) remains

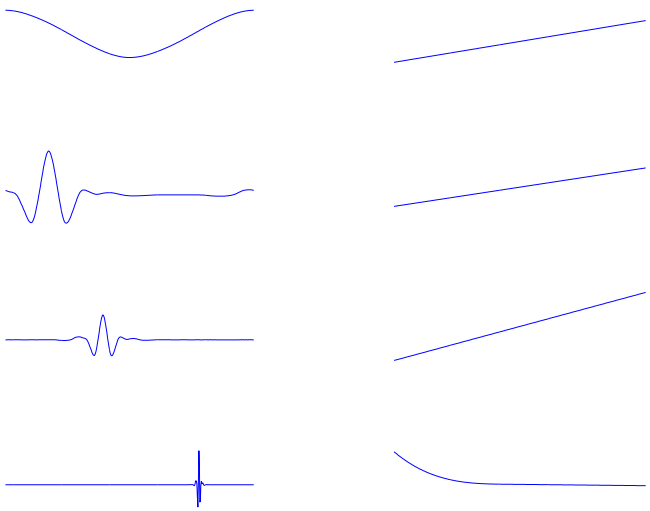


Figure 2: Basis functions 1, 10, 23, 230 from $N = 512$ symmlet (left) and lspline kernels (right). Whereas the symmlet kernel contains components at all frequencies, the lspline kernel only offers low-frequency components. This explains why the classical RVM's relatively weak sparseness enforcement suffices for lspline kernels, but not symmlets.

outside the principled probabilistic framework. Choosing kernel type (and in some cases width parameters) via cross validation-schemes is not just cumbersome and wasteful on data and computational resources, it also typically offers only a crude approach to sparsity control, making it for example difficult or impossible to obtain good results with standard kernels for multi-resolution data (for a good example see Figure 5 later).

1.1.2 Amending the RVM; outlook and overview

Fortunately a strength of Bayesian models is their inherent extensibility by means of additional prior structure; here we examine how to incorporate a wavelet-shrinkage inspired, noise-dependent *smoothness prior* for RVM models without degrading the efficiency of Tipping and Faul's (2003) fast RVM scheme (in fact performance can in many cases be *significantly* improved due to increased sparsity and the gained ability to obtain good results with wavelet kernels which allow efficient ($O(N)$) implementations of operations which are cubic in the general case). In brief, our prior is of the form, $p(\alpha_m \mid \sigma^2) \propto e^{-c/(1+\sigma^2\alpha_m)}$ (where c is a constant that controls the level of smoothing) and as is visible from inspection of Figure 3 (c.f. 1), or indeed the formula itself, it greatly promotes sparsity.

Having outlined the motivation for better prior-controlled sparsity control and briefly introduced our proposed prior in this section, we discuss our smoothness prior in more detail in section 2. In section 3 we show how the efficient scheme for learning the parameters in (Faul and Tipping, 2002) may be simply adapted to incorporate the smoothness prior. Results on a variety of standard datasets showing that it effectively controls sparsity are provided in section 4, followed by a discussion of alternative priors and the summary and conclusion in 5. More detailed theoretical discussions and proofs are relegated to the appendix. A preliminary report on this work appeared in (Schmolck and Everson, 2005).

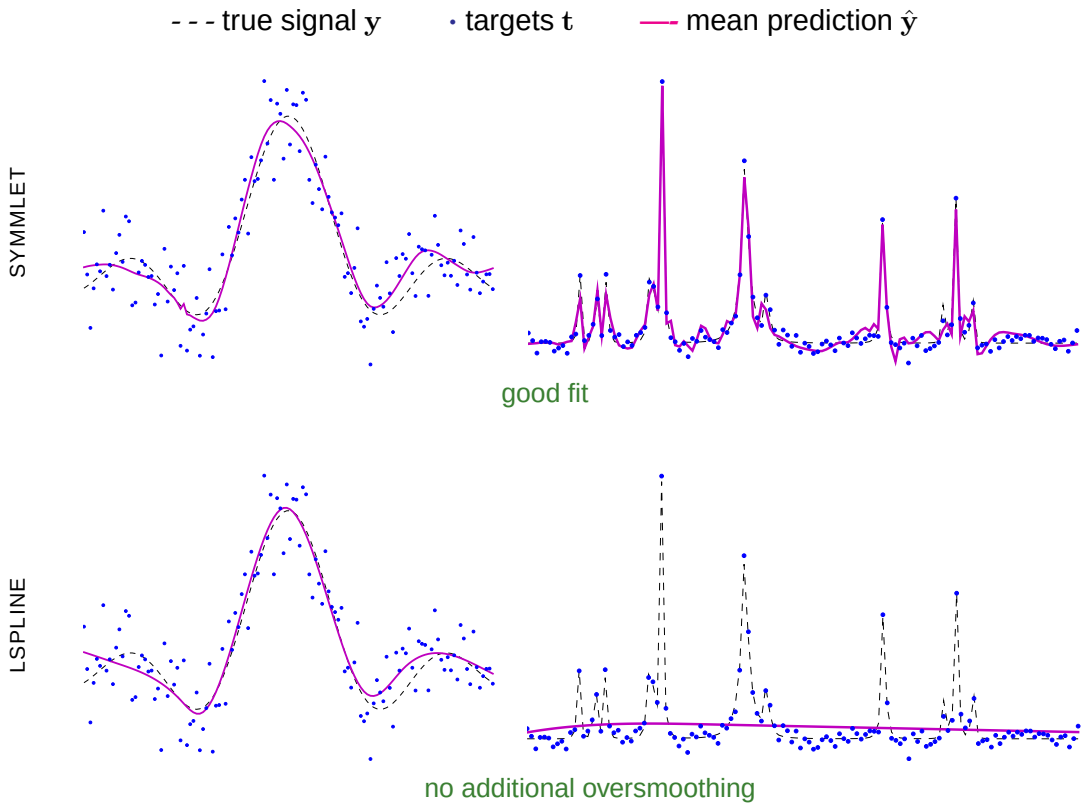


Figure 3: sRVM. The smoothness prior means that enforcing sparsity is no longer mostly relegated to the choice of kernel. A symmlet kernel (*top row*) no longer results in drastic overfitting for the Sinc data set (*on the left*). The bottom row shows that the smoothness prior typically has no adverse effect when smoothing is already mandated by the kernel. The data sets are identical to Figure 1.

2 The smoothness prior

On its own (4) does not appear to strongly favour sparsity, but of course the overall effect on the weights depends on the prior assigned to $\boldsymbol{\alpha}$. Here we consider only priors of the form $p(\boldsymbol{\alpha}, \sigma^2) = \prod_{m=1}^{M} p(\alpha_m \,|\, \sigma^2) p(\sigma^2)$. Then, the effective prior on w_m is found from:

$$p(w_m \,|\, \sigma^2) = \int p(w_m \,|\, \alpha_m) p(\alpha_m \,|\, \sigma^2)\, d\alpha_m \tag{6}$$

When the prior is a Gamma density, $p(\alpha_m \,|\, \sigma^2) = p(\alpha_m) = \Gamma(\alpha_m)^{-1} b^a \alpha_m^{a-1} e^{-b\alpha_m}$ with hyper-parameters a and b, then $p(w_m)$ is a Student-t density. Tipping (2001) presents a nice graphical illustration that the joint distribution $p(w_1, w_2)$ of two Student-t densities concentrates probability mass close to zero values of w_1 and w_2 rather than in regions where both w_1 and w_2 are non-zero, thus encouraging sparse solutions. In fact, the product of any two super-Gaussian[4] prior densities $p(w_m)$ in combination with a Gaussian noise model favours posterior solutions for which one or the other or both w_m are close to zero. This may be seen by noting that the log likelihood (3) is quadratic in $\mathbf{w}$ so that if $\log p(w_1, w_2) = \log p(w_1) + \log p(w_2) \approx w_1^q + w_2^q$ with $q < 2$, then as either coefficient moves away from the coordinate axis the log likelihood decays more rapidly than the log prior, thus encouraging a sparse posterior solution.

[4] Densities whose tails decay more slowly than Gaussians.

The expression (6) with the ARD prior (4) shows that $p(w_m)$ is a scale mixture of Gaussians and so under quite general conditions has positive kurtosis (Clarkson and Barrett, 2001; Lam and Goodman, 2000). It appears, therefore, since almost any prior on $\boldsymbol{\alpha} \mid \sigma^2$ will favour sparsity to some extent, that there is considerable freedom in its choice.

As it is empirically clear that the $p(\mathbf{w})$ resulting from a uniform $p(\boldsymbol{\alpha} \mid \sigma^2)$ (henceforward **None** prior) does not enforce sparsity strongly enough for flexible kernel types (Figure 1), a well-founded, sparser prior over $\boldsymbol{\alpha} \mid \sigma^2$ is desirable. Since the question of existing and proposed prior types for the RVM is somewhat convoluted, we postpone a more extensive discussion till section 5.1, and concentrate for now on a smoothness prior.

As our desire for sparse $\mathbf{w}$ is ultimately grounded in beliefs about the complexity and structure of the signal $\mathbf{y}$, it is in a way natural to work one's way backwards, viz to fashion the prior $p(\boldsymbol{\alpha} \mid \sigma^2)$ so that the mean posterior prediction $\hat{\mathbf{y}}$ reflects these beliefs.

Given the posterior over the weights

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) &= \frac{p(\mathbf{w}|\mathbf{t}, \sigma^2) p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \\ &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned} \tag{7}$$

with

$$\boldsymbol{\Sigma} = (\sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathrm{diag}(\boldsymbol{\alpha}))^{-1} \tag{8}$$

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} \tag{9}$$

we obtain

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}\boldsymbol{\mu} = (\boldsymbol{\Phi}\sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\mathbf{t} \equiv \mathbf{S}\mathbf{t} \tag{10}$$

where $\mathbf{S}$ is known as the *smoothing matrix* (Hastie and Tibshirani, 1990). Note that without the term $\mathrm{diag}(\boldsymbol{\alpha})$, which can be regarded as a regularization term, $\mathbf{S}\mathbf{t}$ would just be the projection of $\mathbf{t}$ into the column space of $\boldsymbol{\Phi}$, or equivalently, the least squares estimate

$$\hat{\mathbf{y}}_{\mathrm{LS}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{t} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\dagger}\mathbf{t} \tag{11}$$

Thus $\mathbf{S}$ computes the *regularized* or smoothed projection of $\mathbf{t}$. Furthermore the special case where all α_i are identical is equivalent to ridge-regression (Hoerl and Kennard, 1970) with regularization parameter $\lambda = \sigma^2\alpha$ (the larger λ, the smoother the estimate, the more all w_i are *shrunk* towards zero compared to the least squares estimate). The "ridge-regression prior" $p(\mathbf{w}|\alpha) = \mathcal{N}(0, \alpha^{-1}\mathbf{I})$ is naturally also just a special case of the ARD prior (4). Of course indiscriminately shrinking coefficients for relevant as well as irrelevant basis functions is unattractive, but ridge regression has the convenient property that the amount of shrinking can be easily quantified.

However, even with an ARD prior, it is possible to quantify the degree of smoothing imposed by the model by a single number: the degrees of freedom of $\mathbf{S}$, given by its trace:

$$\mathrm{DF} = \mathrm{tr}\,\mathbf{S} \tag{12}$$

It is helpful to consider the case of orthonormal basis functions (such as wavelets) so that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \mathbf{I}_M$ and the trace of the smoothing matrix is seen to be:

$$\mathrm{DF} = \mathrm{tr}\,\mathbf{S} = \sum_{m=1}^{M} (1 + \sigma^2\alpha_m)^{-1} \tag{13}$$

In (13) it is evident that basis functions with $\alpha_m \to \infty$ make no contribution to the degrees of freedom, whereas functions with $\alpha_m = 0$ contribute fully; DF thus counts the number of active basis functions, where the extent to which they are active is measured relative to the noise magnitude.

These equations make the related roles of $\boldsymbol{\alpha}$ and $\operatorname{tr}\mathbf{S}$ for sparsity control and smoothing apparent. As all $\alpha_m \to 0$ we approach the least squares estimate (11). In this case $\operatorname{tr}\mathbf{S} = N$, there is no smoothing and the model interpolates the data (indeed for orthormal or invertible $\mathbf{\Phi}$, the least squares estimate is just $\mathbf{t}$). Conversely, as $\alpha_m \to \infty$ the corresponding component $\boldsymbol{\phi}_m$ is turned off ($w_m = 0$). Here $\operatorname{tr}\mathbf{S} = 0$ and the mean posterior estimate is zero.

As the model typically has roughly as many (for square $N \times M$ design matrix $\mathbf{\Phi}$, $M = N$) or more parameters (for overcomplete $\mathbf{\Phi}$, $N < M$) as training examples, the least squares estimate (all $\alpha_m \to 0$) will almost always result in drastic overfitting (never mind severe computational headaches[5]). Conversely all $\alpha_m \to \infty$ will just yield a constant prediction as all $w_m = 0$. But since the RVM associates an unique hyperparameter α_m with each weight w_m a suitable prior over $\boldsymbol{\alpha} \,|\, \sigma^2$ will bring about just the right amount of smoothing for each individual component when we maximize the posterior probability over the weights $p(\mathbf{w} \,|\, \mathbf{t}, \boldsymbol{\alpha}, \sigma^2)$. We expect most components to be turned off, hence most α_m to be ∞ and thus their corresponding weights w_m to be 0, but the few relevant components will have finite α_m and $w_m \neq 0$.

2.1 Finding a suitable prior over α or wavelet shrinkage to the rescue

Similar observations lead Holmes and Denison (1999) to choose the following prior structure for encoding sparsity beliefs for the related problem of wavelet shrinkage:

$$p(\boldsymbol{\alpha} \,|\, \sigma^2) \propto e^{-c\mathrm{DF}} \tag{14}$$

Since DF may be regarded as the effective number of parameters in the regression problem with $\mathbf{\Phi}$ fixed, different choices for the hyperparameter c may be related to different classical model choice criteria (Holmes and Denison, 1999):

$$c = \begin{cases} 0 & \textbf{None}, \text{ Bayes factor (so the classical RVM is just a special case)} \\ 1 & \textbf{AIC}, \text{ Akaike information criterion} \\ \log(N)/2 & \textbf{BIC}, \text{ Bayesian information criterion} \\ \log(N) & \textbf{RIC}, \text{ Risk inflation criterion} \end{cases}$$

Thus we are left with 4 different weight variance priors, from least smoothing to most smoothing as follows: **None, AIC, BIC, RIC**.

Using (13) to compute DF even in the non-orthogonal case yields a convenient prior expression for an individual α_i that does not depend on any of the other $\alpha_{j\neq i}$ and we adopt this form throughout:

$$p(\alpha_i \,|\, \sigma^2) \propto e^{-c\sum_{i=1}^{M}(1+\sigma^2\alpha_m)^{-1}} \simeq e^{-c\mathrm{DF}} \tag{15}$$

[5] Apart from numerical issues, the asymptotic complexity of all standard direct solutions are cubic in M (see, e.g. (Golub and van Loan, 1989)). However, sparse greedy matrix approximations might be used to enhance the convergence rates (Smola and Schmölkopf, 2000), and iterative schemes with improved convergence properties have also been developed; in particular backfitting (Hastie and Tibshirani, 1990) which D'Souza et al. (2004) adapted to a Bayesian EM framework to obtain an $O(MN)$ complexity RVM implementation.

This approximation finds justification beyond computational and analytical expediency. Firstly, we have obtained good empirical results with this prior even when orthonormality is not present (e.g. with various spline kernels and even with overcomplete dictionaries with $M = 2N$; see e.g. Figure 7). Secondly, we also ran some tests where, during model runs, we simultanously computed the true DF expression and compared it to the approximation above and have obtained very similar results. Thirdly our prior structure exerts strong pressure to exclude[6] redundant components, hence although the basis functions might not be orthogonal the eventually *included* basis functions can be expected to be typically near-orthogonal.

A noteworthy and distinguishing characteristic of the smoothness prior is its noise dependency. The degrees of freedom amounts to a count of the number of active basis functions. As the noise increases, the DF decrease, i.e. $\mathbf{S}$ becomes more strongly smoothing. This is what one would intuitively expect to happen: everything else staying fixed, if there is no noise ($\sigma^2 = 0$) then the observations $\mathbf{t}$ ought to equal the true signal $\mathbf{y}$ and so should the posterior estimate $\hat{\mathbf{y}}$, thus $\mathbf{S}$ must be the identity. However as the level of noise increases, more and more of the targets $\mathbf{t}$ has to be explained by the noise and hence $\mathbf{S}$ should become more smoothing as the noise level increases.

Note that this means that basis functions with smaller α_m have greater prior support when the noise is larger, which may appear counter-intuitive at first.

Although (15) is an improper prior, a proper prior may be obtained by restricting α_m to a finite range $[L, H]$. In this case we may write:

$$p(\alpha_i \,|\, \sigma^2) = \begin{cases} Z e^{-c\sum_{m=1}^{M}(1+\sigma^2\alpha_m)^{-1}} & L \leq \alpha_m \leq H \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and the normalization constant Z is given by

$$\begin{aligned} Z =& (\sigma^{-2} + H)\exp(-c/(1+\sigma^2 H)) - (\sigma^{-2} + L)\exp(-c/(1+\sigma^2 L)) \\ &+ c\sigma^{-2}[\mathrm{Ei}(c/(1+\sigma^2 L)) - \mathrm{Ei}(c/(1+\sigma^2 H))] \end{aligned} \tag{17}$$

where Ei denotes the exponential integral function (Arfken, 1985). In the work reported here we choose $L = 10^{-10}$, $H = 10^{10}$. This expression is however only needed when we would like to obtain the posterior probability of the result (see A.6 for how to do so efficiently). Furthermore we note that when $\sigma^2 H/c \gg 1$ then $Z \to H$.

To summarize: The smoothness prior clearly favours large α_m, thus encoding a belief that the weight w_m should be close to zero and consequently a sparse solution. The prior has a desirable dependency on the noise. The constant c controls the smoothness prior's severity (with the least severe prior $c = 0$ reducing to the classical RVM's uniform prior). It is also easy to prove that this prior results in a scale invariant posterior (i.e. our model will give the same answers if we rescale $\mathbf{t} \to k\mathbf{t}$ and simultaneously $\sigma \to k\sigma$, see A.2).

Apart from the observations $\mathbf{t}$ and the choice of kernel $\mathbf{\Phi}$, the hyperparameters g, h and c are the only parameters to be externally specified by the user. Moreover we find that $c = \log(N)/2$, the value for **BIC** makes a good default for c, while we can generally, in the absence of any prior belief about the likely shape of $p(\sigma)$, just set g, h to 10^{-4} or some other uninformative value.

How to effectively learn the other parameter values and estimates is the topic of the next section.

[6] i.e. to set the corresponding $\alpha_i = \infty$, peek ahead to sec. 3.2 (14, 15) to see that the corresponding ϕ_i play indeed no role

3 Implementation

3.1 Overview of implementation properties and strategy

The efficient calculation of MAP point estimates for the model parameters that we are about to detail is based on the Tipping and Faul (2003) "fast RVM" scheme and rests mostly on three facts:

1. Our sparsity prior structure ensures that for most real data sets the posterior will peak in regions with mostly infinite α_m and as discussed an infinite value for α_m is equivalent to the exclusion of ith component from the model, so only $S \ll M$ of the coefficients w_m will be nonzero.
2. Although the introduction of this sparsity prior structure means that some expressions no longer have convenient closed form solutions, the solutions are still easily and efficiently found by simple numerical methods in all instances and all the important desirable properties of the fast RVM that are detailed in (Faul and Tipping, 2002) remain unaffected by the inclusion of the smoothness prior.
3. In particular it is still possible to determine the relevance of a basis function not currently included in the model (so that components can be included one by one, starting with an empty model) and to derive expressions for all quantities of interest that only depend on S and not M. Consequently, the computational complexity scales cubically with the number of included components S, rather than the number of basis functions M.

With the **None** prior and uniform $p(\boldsymbol{\alpha} \,|\, \sigma)$ maximization of $\log p(\boldsymbol{\alpha} \,|\, \sigma, \mathbf{t})$ is equivalent to maximizing the log marginal likelihood $\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{t} \,|\, \boldsymbol{\alpha}, \sigma)$, which can be efficiently effected by the elegant type II maximum likelihood scheme described in Faul and Tipping (2002) and Tipping and Faul (2003). The key idea is to write

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i) \tag{18}$$

in order to separate out the contribution of the ith basis function $\boldsymbol{\phi}_i$ into the term $\ell(\alpha_i)$ which depends solely on α_i and a term $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ that is independent of α_i. Maximization of $\mathcal{L}(\boldsymbol{\alpha})$ then proceeds by successive maximizations of $\ell(\alpha_i)$ for a sequence of components. With the **None** prior the maximizing $\alpha_i^\star = \mathrm{argmax}_{\alpha_i} \ell(\alpha_i)$ is found in closed form, so that the maximization is particularly cheap.

In our case, the addition of the smoothness prior means that rather than the log likelihood, we seek to maximize the log posterior:

$$\hat{\mathcal{L}}(\boldsymbol{\alpha}) \equiv \log p(\mathbf{t} \,|\, \boldsymbol{\alpha}, \sigma^2) + \log p(\boldsymbol{\alpha} \,|\, \sigma^2) \tag{19}$$

$$= \mathcal{L}(\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha} \,|\, \sigma^2) \tag{20}$$

Due to the multiplicative prior structure, $p(\boldsymbol{\alpha} \,|\, \sigma^2) = \prod_i p(\alpha_i \,|\, \sigma^2)$ the dependence of $\hat{\mathcal{L}}(\boldsymbol{\alpha})$ on α_i can still be isolated, and although the additional term requires that the optimal

$$\hat{\alpha}_i = \operatorname*{argmax}_{\alpha_i} \hat{\ell}(\alpha_i) = \operatorname*{argmax}_{\alpha_i} \left[\ell(\alpha_i) + \log p(\alpha_i \,|\, \sigma)\right] \tag{21}$$

is found numerically, rather than analytically as in Tipping and Faul (2003), the extension is straightforward and has the desired properties. In particular:

1. There still is at most one local maximum for $\ell(\alpha_i)$.

2. $\hat{\alpha}_i \geq \alpha_i^\star$, in other words the sRVM or smoothness prior MAP solution is always at least as sparse as the RVM or ML solution (see appendix A.3), but typically is much sparser for flexible kernels.

3. It adds virtually no computational overhead, but allows enormous computational savings in many cases – this is because the complexity of a single step in the model is essentially $O(S^3)$ and the sparsity prior will always produce at least as small an S as the **None** prior, but often only a fraction. The increased sparsity also makes it feasible to use wavelet kernels for many tasks without fear of overfitting which further reduces the complexity per step to $O(N)$, because all matrix multiplications disappear[7]; we have empirically verified that approximately linear in N per-step behaviour obtains in our implementation for $4096 \leq N \leq 524288$.

We now give further details of the maximization scheme, although proofs are relegated to the appendix.

3.2 Maximizing the marginal posterior

For clarity we first recapitulate the decomposition of the log marginal likelihood used by Tipping and Faul's original scheme before we describe the modifications needed to incorporate the smoothness prior. The following subscripts will be used: $_S$ denotes the value of a variable with only the S selected components included whereas $_{-i}$ denotes the value of a variable with the ith component removed.

The log marginal likelihood

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}) &= \log p(\mathbf{t} \,|\, \boldsymbol{\alpha}, \sigma^2) && (22) \\ &= \log \int p(\mathbf{t} \,|\, \mathbf{w}, \sigma^2) p(\mathbf{w} \,|\, \boldsymbol{\alpha}) d\mathbf{w} && (23) \\ &= -\frac{1}{2}\left[N \log 2\pi + \log |\mathbf{C}| + \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right] && (24) \end{aligned}$$

with

$$\mathbf{C} = \sigma^2 \mathbf{I} + \boldsymbol{\Phi} \text{diag}(\boldsymbol{\alpha}^{-1}) \boldsymbol{\Phi}^T \tag{25}$$

can be decomposed to isolate the terms involving a particular α_i. Writing

$$\begin{aligned} \mathbf{C} &= \sigma^2 \mathbf{I} + \sum_{m \neq i} \alpha_m^{-1} \boldsymbol{\phi}_m \boldsymbol{\phi}_m^T + \alpha_i^{-1} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T && (26) \\ &\equiv \mathbf{C}_{-i} + \alpha_i^{-1} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T && (27) \end{aligned}$$

the log likelihood may be reformulated, using standard matrix identities for inverse and determinant of $\mathbf{C}$, as (18) where:

$$\mathcal{L}(\boldsymbol{\alpha}_{-i}) = -\frac{1}{2}\left[N \log 2\pi + \log |\mathbf{C}_{-i}| + \mathbf{t}^T \mathbf{C}_{-i}^{-1} \mathbf{t}\right] \tag{28}$$

and

$$\ell(\alpha_i) = \frac{1}{2}\left[\log \alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right] \tag{29}$$

[7] The covariance matrix becomes the identity due to orthonormality, $\boldsymbol{\Sigma}$ becomes diagonal and other multiplications by $\boldsymbol{\Phi}$ is can be replaced by the equivalent, but much more efficient, discrete wavelet transform.

The quantities

$$s_i \equiv \boldsymbol{\phi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\phi}_i \tag{30}$$

and

$$q_i \equiv \boldsymbol{\phi}_i^T \mathbf{C}_{-i}^{-1} \mathbf{t} \tag{31}$$

respectively measure the degree to which $\boldsymbol{\phi}_i$ overlaps other basis functions in the solution (its "sparsity") and its "quality," namely its correlation with the model error with $\boldsymbol{\phi}_i$ excluded: $q_i = \sigma^{-2} \boldsymbol{\phi}_i^T (\mathbf{t} - \hat{\mathbf{y}}_{-i})$.

Now it is easy to maximize $\mathcal{L}(\boldsymbol{\alpha})$ with respect to α_i. Faul and Tipping (2002) show that $\ell(\alpha_i)$ has a single unique maximum at

$$\alpha_i^\star = \begin{cases} \frac{s_i^2}{q_i^2 - s_i} & \text{if } q_i^2 > s_i \\ \infty & \text{otherwise} \end{cases} \tag{32}$$

Kernels for which $q_i^2 < s_i$ are effectively excluded from the model and the elegance of the Tipping and Faul (2003) fast RVM scheme derives from the fact that s_i and q_i can be calculated from quantities involving only the $S \ll M$ included components.

Maximization of the log marginal posterior with respect to a single α_i can be achieved by maximizing

$$\hat{\ell}(\alpha_i) = \ell(\alpha_i) - \frac{c}{1 + \sigma^2 \alpha_i}. \tag{33}$$

The derivative of $\hat{\ell}(\alpha_i)$ is

$$\hat{\ell}'(\alpha_i) = \frac{1}{2} \left[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + s_i} - \frac{q_i^2}{(\alpha_i + s_i)^2} \right] + \frac{c}{(1 + \sigma^2 \alpha_i)^2} \tag{34}$$

$$= \frac{P(\alpha_i)}{2\alpha_i (\alpha_i + s_i)^2 (\alpha_i + \sigma^{-2})^2} \tag{35}$$

where $P(\alpha_i)$ is cubic in α_i. Simple closed form solutions to the roots of $P(\alpha_i) = 0$ are not available, however it is simple and computationally cheap to numerically find the roots of P. Since $\lim_{\alpha_i \to \infty} \hat{\ell}(\alpha_i) = 0$ the maximum of $\hat{\ell}(\alpha_i)$ may occur at infinite α_i, corresponding to the ith basis function being "switched off". A basis function ϕ_i is active if the maximum occurs for $\alpha_i < \infty$. With the smoothness prior the posterior may have more than one turning point but, as shown in Appendix A.3, there can be at most one maximum $\hat{\alpha}_i < \infty$ and $\alpha_i^\star < \hat{\alpha}_i$ showing that the smoothness prior always has the effect of making the solution sparser. Figure 4 shows the four possible cases that may arise. Most severely, (*top-left*) the prior can null the maximum in the likelihood. Alternatively, (*top-right*) the posterior has a, possibly local, maximum at finite $\hat{\alpha}_i > \alpha_i^\star$; when there is a single turning point $\hat{\alpha}_i > \lim_{\alpha_i \to \infty} \hat{\ell}(\alpha_i) = 0$ and the basis function is active; if there are two turning points in $\hat{\ell}(\hat{\alpha}_i)$ the global maximum may be at the turning point (*bottom-left*) or at infinite α_i. It is straightforward during learning to distinguish between these last two cases by evaluating $\hat{\ell}(\hat{\alpha}_i)$.

In brief, maximization of $\hat{\mathcal{L}}(\boldsymbol{\alpha})$ therefore proceeds by successively choosing (at random) a component i to include in the model, maximizing $\hat{\ell}(\alpha_i)$ with respect to α_i and reestimating the parameters $\boldsymbol{\Sigma}_S$, $\boldsymbol{\mu}_S$, $\mathbf{s}$ and $\mathbf{q}$ which depend upon α_i ($\mathbf{s}$ and $\mathbf{q}$ are M vectors of the sparsity and

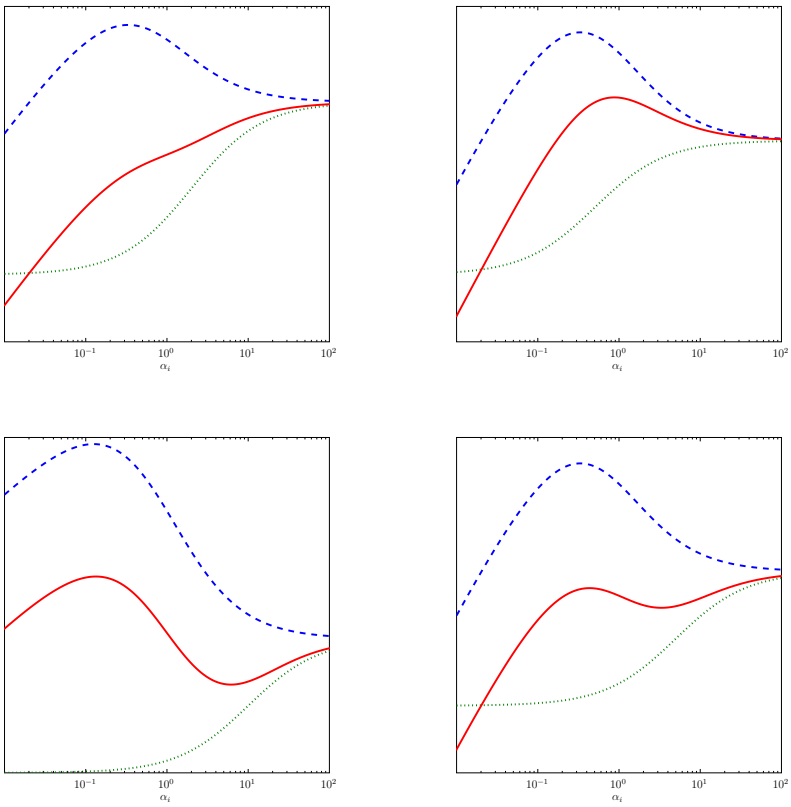


Figure 4: Log posteriors $\hat{\ell}(\alpha_i)$ (solid), log likelihoods $\ell(\alpha_i)$ (dashed), and log prior $-c(1+\sigma^2\alpha_i)^{-1}$ (dotted) plotted versus $\log \alpha_i$ for the four possible cases with a smoothness prior. *Top-left:* prior nulls maximum in posterior; *Top-right:* single turning point with $\hat{\alpha}_i$ finite; *Bottom-left:* two turning points in posterior and $\hat{\ell}(\hat{\alpha}_i) > \lim_{\alpha_i \to \infty} \hat{\ell}(\alpha_i) = 0$; *Bottom-right:* two turning points in posterior, but $\hat{\ell}(\hat{\alpha}_i) < 0$.

quality indices q_i and s_i of the corresponding α_i). Since the posterior is increased by maximization of each individual $\hat{\ell}(\alpha_i)$ such a sequence of maximizations terminates when a, possibly local, maximum of the posterior is located at which the inclusion or deletion of any single component can only decrease the $\hat{\mathcal{L}}(\boldsymbol{\alpha})$. Tipping and Faul (2003) use this computationally efficient sequential maximization as the basis of the *fast* RVM. Although in (Faul and Tipping, 2002) an attempt is made to prove that "sequential optimization of individual α_i cannot lead to a stationary point from which a joint maximization over all $\boldsymbol{\alpha}$ may have escaped", it appears that the proof is flawed and this desirable property may indeed often *not* hold. However, by examining the Hessian of $\hat{\mathcal{L}}(\boldsymbol{\alpha})$ at the maximum it is straightforward to show that this property does hold for orthonormal basis functions (such as wavelet kernels) regardless of the imposition of a smoothness prior (Appendix A.4).

3.3 Noise reestimation

Again whereas the classical RVM can employ an analytical update rule (Tipping, 2001, eq 46) for σ^2 from setting $\frac{\partial \mathcal{L}}{\partial \sigma^{-2}} = 0$, the introduction of the smoothness prior term means we have to resort to a numerical scheme. As noted above, provided that $\sigma^2 H/c \gg 1$ the normalisation

Algorithm 1 The sRVM algorithm.

1: $\sigma^2 \leftarrow 0.1 \times \mathrm{var}(\mathbf{t})$ *initialization for σ^2*
2: $\boldsymbol{\alpha} \leftarrow [\infty \cdots \infty]^T$ *start with the empty model*
3: $i \leftarrow \mathrm{argmax}\,(\|\boldsymbol{\phi}_m^T \mathbf{t}\|/\|\boldsymbol{\phi}_m\|)$ *pick an i that stands a good chance of being relevant*
4: $\mathcal{S} \leftarrow \{i\}$ *include it in the set of included components*
5: update($\alpha_i, \boldsymbol{\Sigma}_S, \boldsymbol{\mu}_S, \mathbf{s}, \mathbf{q}$) *compute initial values for all model paramters*
6: $R \leftarrow 10$ *reestimate noise every R steps*
7: step $\leftarrow 1$ *already made the first step*
8: `until` converged()
9: $i \leftarrow$ randint(M) *pick a random component i*
10: DID-NOTHING $\leftarrow$ False
11: `if` ($q_i^2 - s_i > 0$ `and` ... has-real-positive-root(numerator($\hat{\ell}'(\alpha_i)$)) *if component i is relevant*
12: `unless` $\alpha_i < \infty$ *unless it is already included*
13: $\mathcal{S} \leftarrow \mathcal{S} \cup \{i\}$ *add it*
14: `else`
15: `if` $\alpha_i < \infty$ *the component is irrelevant but currently included*
16: $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$ *delete it*
17: `else` *component is and was irrelevant*
18: DID-NOTHING $\leftarrow$ True *no need for action*
19: `unless` DID-NOTHING *otherwise update everything*
20: step $\leftarrow$ step + 1
21: update($\alpha_i, \boldsymbol{\Sigma}_S, \boldsymbol{\mu}_S, \mathbf{s}, \mathbf{q}$) *update the model parameters*
22: `if` step mod $R = 0$
23: reestimate(σ^2)

term in the prior Z (17) is effectively constant, and we therefore numerically solve:

$$\frac{\partial \hat{\mathcal{L}}(\boldsymbol{\alpha}, \sigma^2)}{\partial \sigma^{-2}} \approx \frac{1}{2}\left[N\sigma^2 - \|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2 - \sigma^2 \sum_m (1 - \alpha_m \Sigma_{mm})\right] - c\sum_m \frac{\alpha_m}{(\sigma^{-2} + \alpha_m)^2} + (g-1)\sigma^2 - h = 0 \quad (36)$$

Good initial guesses for starting the numerical solution are either the $c = 0$ linear solution or the previous value of σ^2, and we find that convergence is rapid. A graphical analysis of (36) shows that the smoothness prior term serves always to increase the effective noise variance as may be expected because a sparser solution requires more of the error to be explained by noise. However, for even moderate amounts of data, we find that the estimate of σ^2 is dominated by the marginal likelihood terms, being insensitive to the noise prior $p(\sigma^2|g, h)$ and hence the choices for g and h.

3.4 The algorithm

The outline of the algorithm, based on the fRVM algorithm, is as follows (c.f. pseudocode in Alg. 1).

We start out with a model that includes only a single component (a good default choice is the component that has the largest projection onto the target $\mathbf{t}$) (lines 1-7). Then, until the model has converged (line 8), at each iteration a candidate component is picked at random (line 9). If upon testing this component turns out to be neither currently included (i.e. $\alpha_i = \infty$) nor relevant (i.e. inclusion would not increase $\hat{\ell}(\alpha_i)$ or equivalently the overall posterior $\hat{\mathcal{L}}(\boldsymbol{\alpha})$)

nothing is done (line 18f). In all other cases the model is updated: α_i and all other relevant parameters are reestimated (line 21); only the noise estimation is not updated on each step (lines 22f) to prevent spurious oscillation.

There are thus 4 possible cases that can occur after the relevance of some component i given the current state of the model has been established:

- The component is currently not included but is still deemed irrelevant, so nothing happens.
- The component is currently not included, but since inclusion would increase $\hat{\ell}$, it is included (line 13).
- The component is currently included and α_i is updated to reflect the current state of the model which can either mean:
  - deletion: α_i is set to ∞ if doing so does not reduce $\hat{\ell}$ (line 16);
  - (mere) reestimation: the value of α_i is set to some finite value, possibly the same that it already had (line 21).

It should be noted that whereas only a single α_i of all the $\boldsymbol{\alpha}$ is updated on each step, *all* the other parameters, i.e. all the M q_m and s_m as well as the S elements of $\boldsymbol{\mu}_S$ and $S \times S$ elements of $\boldsymbol{\Sigma}_S$ are completely recalculated.[8]

3.4.1 The convergence criterion

Due to the greedy nature and itemwise update of the algorithm finding, a good convergence criterion requires a bit of tweaking to prevent premature convergence while at the same time avoiding endless iterations close to the solution.

We follow Tipping and Faul (2003) in requiring that the differences between successful values for any α_i in logarithmic space must be less than 10^{-6} and that all components currently deemed to be relevant are actually included in the model.

Whilst this suffices in many cases, we have found it useful to add some additional requirements and as a consequence in practice the scheme is a bit more complex than that depicted in Alg. 1. Most importantly, before we declare convergence we ensure that all component α_m are re-evaluated and that the noise reestimation remains stable over several past estimates. We also test that $\hat{\mathcal{L}}$ no longer increases noticeably. Full details may be found in the implementation which is available from `<http://www.dcs.ex.ac.uk/~reverson/sRVM>`.

3.5 Future directions

Although empirically the greediness of the fRVM type II MAP scheme on which we base this work does not seem to be much of in an issue under most scenarios (including the spline kernel examples that have dominated the RVM literature and our use of wavelet kernels), under the stress-test of using overcomplete dictionaries local maxima can start to become a problem.

Whilst we have found it useful to increase exploration by ad hoc measures[9] in these cases, the

[8] The efficient formulae $s_m = \frac{\alpha_m S_m}{\alpha_m - S_m}$ where $S_m = \sigma^{-2}\boldsymbol{\phi}_m^T\mathbf{t} - \sigma^{-4}\boldsymbol{\phi}_m^T\boldsymbol{\Phi}_S\boldsymbol{\Sigma}_S$, and $q_m = \frac{\alpha_m Q_m}{\alpha_m - Q_m}$ where $Q_m = \sigma^{-2}\boldsymbol{\phi}_m^T\mathbf{t} - \sigma^{-4}\boldsymbol{\phi}_m^T\boldsymbol{\Phi}_S\boldsymbol{\Sigma}_S\boldsymbol{\Phi}_S^T\mathbf{t}$ are used for calculating $\mathbf{q}$ and $\mathbf{s}$ (Tipping and Faul, 2003).

[9] e.g. by initially also including α_i for which $\hat{\ell}(\alpha_i) \leq \hat{\ell}(\infty)$.

added flexibility of the sRVM for kernel choice appears to call for a principled way to deal with local maxima. We are therefore currently investigating an alternative non-greedy MCMC formulation that samples from the posterior.

4 Results

4.1 Simple data

As Figure 3 shows, we find that use of the smoothness prior typically yields substantial improvements for tasks where overfitting is a problem due to the multi-scale resolution of the kernel, while it generally has no appreciable negative impact when overfitting is not an issue.

4.2 Multiscale data

In Figure 5 we can clearly see the advantages of smoothness control via prior structure as opposed to kernel choice: with the sRVM a multiscale signal can receive just the right level of smoothing to fit the signal, but not the noise, at each scale (Figure 5 *bottom*), whereas the RVM's dependence on kernel choice for sparsity control and thus smoothing means choosing between the evils of oversmoothing the high-frequency structure (Figure 5 *top*) or overfitting the low-frequency structure (Figure 5 *middle*).

The effect of the smoothness prior on sparsity is most clearly visualized by comparing "shrinkage plots" for **None** and **BIC** (Figure 6) for the Doppler data.

4.3 Heterogenous data and overcomplete dictionaries

Another attractive ability of the sRVM is to automatically choose the right locally fitting components from an overcomplete dictionary. This can be used to obtain very good results for signals that are hetergenous to the extent that particular regions can be well represented sparsely by a particular (non-custom) kernel whilst another standard kernel (or combination of kernels) would yield much better results for other regions. Figure 7 presents an illustrative toy example in which an overcomplete dictionary of thin-plate spline (tpspline) kernels and Haar wavelets kernels are shown to provide an effective sparse representation of the concatenation of the smooth Sinc data and the step-like Blocks data (Donoho and Johnstone, 1994). Likewise, data such as the HeaviSine dataset (ibid.), with small discontinuity regions but mostly smooth and continuous overall can also profit from a similar overcomplete dictionaries.

Overcomplete dictionaries constructed from morphologically diverse kernels have also found applications to blind source separation problems, where the morphological differences in individual signal components allow such components to be largely represented by morphologically similar parts of the overcomplete dictionary which can be leveraged to effect the separation (Bobin et al., 2005). Although the sRVM is clearly not designed with that task in mind, in this specific, simple example we obtain near-perfect separation of the Blocks and Sinc signal components by discarding the tpspline and Haar contribution respectively (see Figure 8 *right*). By contrast, the plain RVM cannot achieve this separation (see Figure 8 *left*), apart from needing 419 instead of 88 components and achieving only a MSE of 0.024 instead of 0.009.

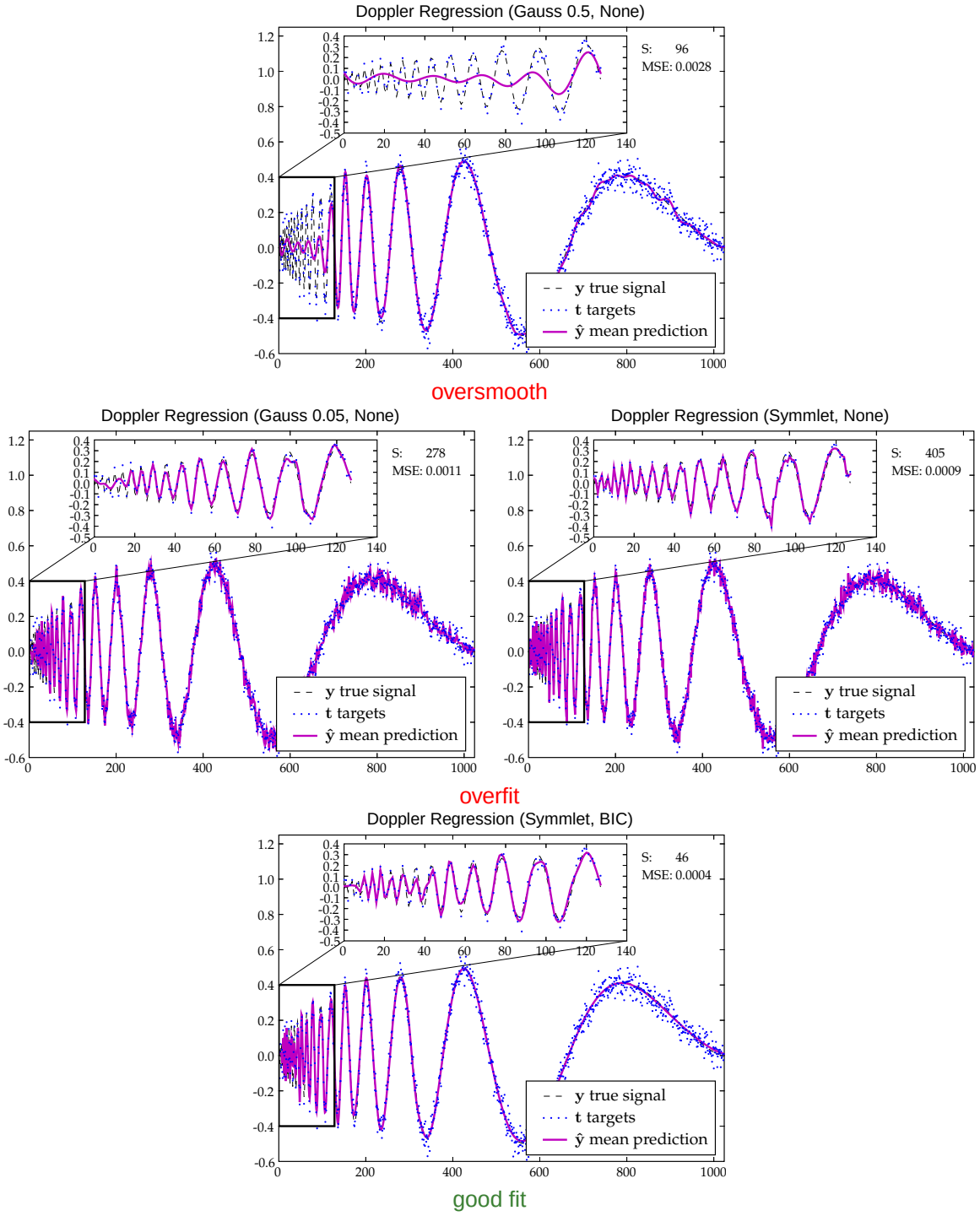


Figure 5: Multiscale resolution data like Doppler ($N = 1024$, SNR $= 7.0$) defeats the RVM (*top*, *middle*), but not the sRVM (*bottom*), demonstrating the limitedness of sparsity control via kernel (parameter) choice (here in *top* and *middle* panels via the width parameter r of a Gaussian kernel (r is respectively 0.5 and 0.05) as well kernel choice between gauss (*top* and *left middle* vs. the parameterless symmlet *right middle*)). Such smoothness control acts globally, whereas only part of the signal is respectively fine scale/large scale, so that even though overfitting already starts to become apparent in the *top* panel, the fine scale information on the left side is still severely oversmoothed. Decreasing kernel width (*middle*) to improve resolution sufficiently to fit the fine scale details on the left is seen to be tied to drastic overfitting in the right part of the plot. By contrast, a smoothness prior in combination with a multi-resolution kernel achieves an adaptive level of smoothing (*bottom*). Without a smoothness prior, again drastic overfitting would occur (not pictured, but c.f. **Figure 1** top left).

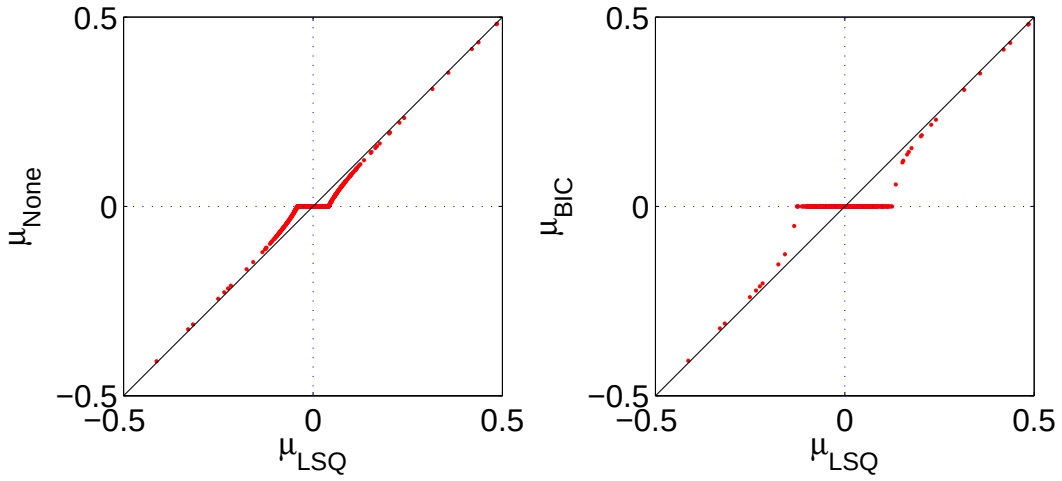
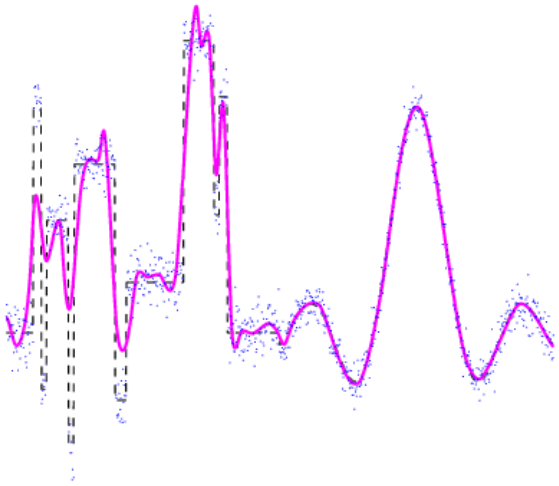
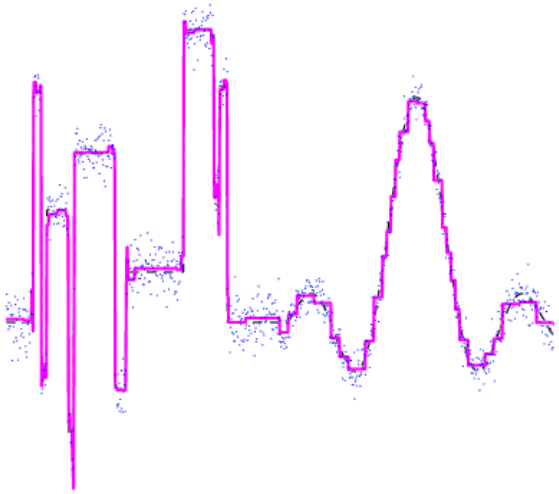


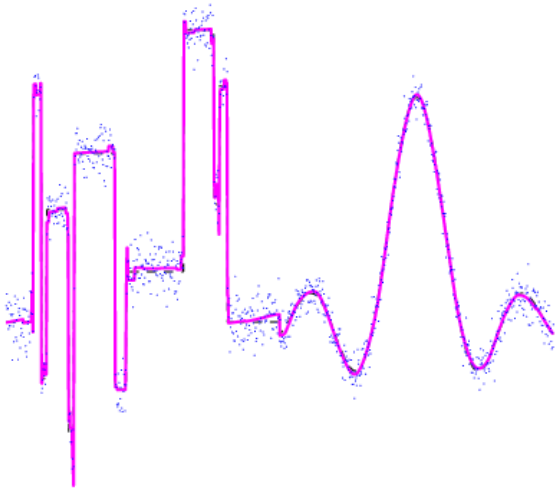
Figure 6: Shrinkage plots. Plotting the least squares estimate of the weights $\boldsymbol{\mu}_{LSQ}$ against the posterior weight estimates obtained with **None** (*left*) and **BIC** (*right*) priors clearly shows that the **BIC** smoothness prior is much more effective at weeding out small, irrelevant components by setting them to 0. These plots correspond to the middle right and bottom panel, respectively, of Figure 5) and are clipped to $|\mu_m| \leq 0.5$ (the few larger components are essentially unaffected by shrinkage and thus lie on the diagonal).



(a) sRVM 3.0 tpspline kernel (MSE: 0.379, S: 72)
left half (Blocks) of signal cannot be properly resolved



(b) sRVM haar kernel (MSE: 0.020, S: 75)
now the right half (Sinc) shows staircase artifacts



(c) sRVM overcomplete haar+tpspline dictionary (MSE: 0.009, S: 88)
the sRVM automatically finds the right basis functions for each region of the signal

Figure 7: The sRVM (here with **RIC** prior) makes it possible to obtain very good results by using overcomplete dictionaries. The example data ("BlocksSinc") is constructed by concatenating two signals with very different characteristics: Blocks and Sinc and adding Gaussian noise (SNR: 7.0). Whilst no standard kernel will give ideal results for this combination, thin-plate splines (tpsplines) are well suited for smooth, continuous curves such as Sinc (a), whilst the step-like nature of Haar wavelets makes them the ideal candidate for the Blocks subset (b). However, thanks to the smoothness prior, the sRVM can do a remarkably good job at automatically picking the appropriate components for each part of the signal from an overcomplete dictionary obtained by concatenating both these kernels together (also see

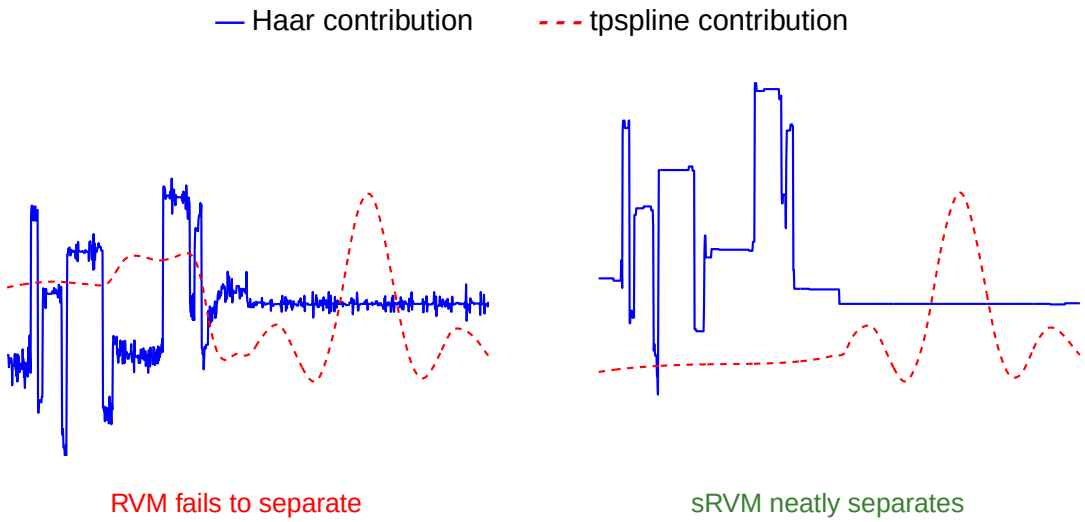


Figure 8: Contributions of the tpspline and Haar subparts of the overcomplete dictionary for BlocksSinc signal for sRVM and RVM. Evidently the sRVM is able to pick fitting components for each morphologically distinct part of the overall signal – indeed by discarding all the tpspline or Haar contributions to $\hat{\mathbf{y}}$ one essentially obtains a clean separation into Blocks and Sinc.

.

It has to be noted, however, that although with symmlet or spline kernels we generally achieve bit-identical or near-identical results regardless of the way component inclusion proceeds, overcomplete kernels appear to expose some limitations of the fRVM scheme on which we build. Apart from numerical issues caused by the overcompleteness, getting trapped in different local maxima starts to become a problem, so that we see more variability in the results than we do under simpler scenarios.

4.4 Summary statistics

Table 1 shows for a number of standard datasets the sparsity, measured by the number of included components S, and the MSE between the mean prediction $\hat{\mathbf{y}}$ and the true signal $\mathbf{y}$. Clearly the **None** prior is generally insufficiently severe to control the sparsity for multiresolution kernels, while the smoothness priors provide sufficient smoothing and thus permit σ^2 to be correctly estimated. On the other hand it is evident from the gauss 3.0 (see A.1 for definitions of all used kernels) examples that the smoothness priors do not result in missestimation when smoothing is already enforced by the kernel. Further, the Doppler data in the second half of the table demonstrates that even if the true value for σ^2 is given so that incorrect noise estimation is not an issue, the **None** prior is still too weak to bring about the desired level of sparsity.[10]

Lastly, although **BIC** rarely obtains the best answer, it is typically not too far off which recommends it as the default choice.

[10]With this exception all results in this paper, including those in Figure 5, were obtained using noise estimation.

σ^2 is estimated

Bumps SNR=2.0 (**$\sigma^2 = 0.119$, $N = 128$**)

Kernel	Prior	S	MSE	σ^2_{MAP}
symmlet	None	127.0±0.0	0.119±0.001	0.000±0.000
symmlet	**AIC**	36.3±7.8	**0.088±0.008**	0.121±0.037
symmlet	BIC	11.9±2.5	0.153±0.034	0.262±0.038
symmlet	RIC	2.6±1.4	0.320±0.051	0.450±0.068

Bumps SNR=7.0 (**$\sigma^2 = 0.010$, $N = 128$**)

Kernel	Prior	S	MSE	σ^2_{MAP}
symmlet	None	127.0±0.0	0.010±0.000	0.000±0.000
symmlet	**AIC**	61.9±6.0	**0.009±0.001**	0.010±0.004
symmlet	BIC	19.2±4.9	0.081±0.024	0.106±0.029
symmlet	RIC	6.4±1.3	0.203±0.024	0.238±0.018

Sinc SNR=2.0 (**$\sigma^2 = 0.031$, $N = 128$**)

Kernel	Prior	S	MSE	σ^2_{MAP}
gauss 3.0	None	5.7±0.7	0.004±0.001	0.032±0.001
gauss 3.0	**AIC**	5.4±1.1	**0.004±0.001**	0.034±0.001
gauss 3.0	BIC	5.2±0.6	0.005±0.001	0.035±0.002
gauss 3.0	RIC	4.9±1.0	0.005±0.001	0.036±0.002
symmlet	None	127.0±0.0	0.031±0.000	0.000±0.000
symmlet	AIC	28.9±5.9	0.012±0.003	0.020±0.004
symmlet	BIC	9.1±1.9	0.006±0.002	0.032±0.003
symmlet	**RIC**	6.2±0.6	**0.006±0.001**	0.036±0.002

σ^2 is given

Doppler SNR=2.0 (**$\sigma^2 = 0.031$, $N = 1024$**)

Kernel	Prior	S	MSE
symmlet	None	367.3±10.5	0.00067±0.00002
symmlet	AIC	130.5±6.9	0.00041±0.00002
symmlet	**BIC**	56.6±2.8	**0.00026±0.00002**
symmlet	RIC	42.2±1.5	0.00036±0.00002

Table 1: Empirical comparisons of different priors on standard datasets. Results are averaged over 10 runs (with different noise ϵ on each run). Results with lowest MSE appear in **bold**.

5 Discussion

Whilst we have concentrated on the fRVM framework (Faul and Tipping, 2002), since our implementation is based on it, it is worth mentioning that the fRVM is by no means the only attempt to provide a scheme that is computationally more efficient than the original, "slow" RVM (Tipping, 2000) and might be adapted to incorporate a smoothness prior; we draw attention to the Subspace EM (SSEM) algorithm (Quiñonero-Candela, 2004) and a version based on a Bayesian interpretation of backfitting (D'Souza et al., 2004).

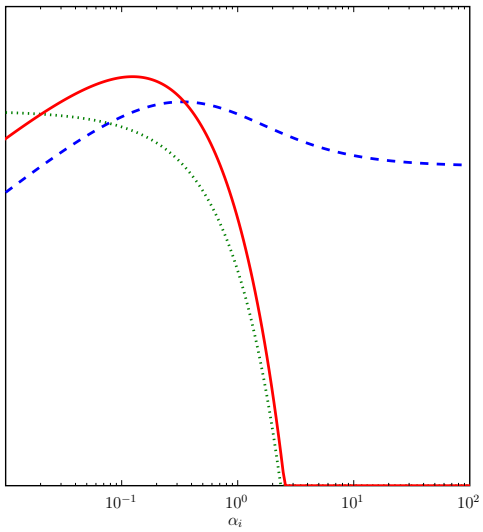


Figure 9: Log posteriors $\hat{\ell}(\alpha_i)$ (solid), log likelihoods $\ell(\alpha_i)$ (dashed), and log gamma prior (dotted) plotted versus $\log \alpha_i$, showing that with $q_i = 1$, $s_i = 2$, $a = 1$ and $b = 2$ the MAP α_i is less sparse than the maximum likelihood solution.

Before examining more closely the issue of different choices for $p(\boldsymbol{\alpha})$, we mention other work pertinent to RVM learning. Wipf and Rao (2004) provide a principled justification for approximating the hyperparameter posterior $p(\boldsymbol{\alpha}, \sigma^2 \,|\, \mathbf{t})$ with the point estimates $\boldsymbol{\alpha}_{\mathrm{MAP}}$ and σ_{MAP}. Quiñonero-Candela (2004) offers an augmentation to the RVM at the prediction stage to ameliorate the problem of artificially low predictive variances for test-points that are far off the "centres" of the "relevance vectors" (i.e. the final set of basis functions for well-localized kernels such as gauss) – an issue that may be regarded as an undesirable side-effect of sparsity. Figueiredo (2003) provides an illuminating perspective on sparse Bayesian learning and presents an Expectation Maximisation algorithm for learning the coefficients $\mathbf{w}$ directly, treating the $\boldsymbol{\alpha}$ as hidden data.

5.1 Other prior choices for $\boldsymbol{\alpha}$

As we noted in section 2, any super-Gaussian prior on each $p(w_m)$ will encourage sparseness or shrinkage. A natural prior that has been used to promote sparsity in a variety of contexts is the Laplacian prior, $p(w_m) \propto e^{-|w_m|}$, which leads to the LASSO (least absolute shrinkage and selection operator) scheme (Tibshirani, 1996), although in this context the prior is introduced as the penalty in a penalized likelihood. As Figueiredo (2003) shows, the Laplacian prior on $p(w_m)$ may be obtained via a hierarchical scheme, like ours, in which a $p(w_m)$ arises as a scale mixture of Gaussians with an exponential prior on the precisions: $p(\alpha_m \,|\, \gamma) \propto e^{-\gamma\alpha/2}$, where γ is a hyperparameter. In fact, Figueiredo abandons the exponential/Laplacian scheme in favour of a Jeffreys' prior on α_m, namely $p(\alpha_m) \propto 1/\alpha_m$, which in turn results in a similar very heavy-tailed prior on the coefficients: $p(w_m) \propto 1/|w_m|$. The attractions of the Jeffreys' prior result from the fact that it is a *non-informative* prior (Bernardo and Smith, 1994): first, it is scale invariant and, secondly, there are no (hyper)parameters to adjust. Before examining the Jeffreys' prior in more detail we first discuss the Gamma prior which has the Jeffreys' prior as a limiting case.

The Gamma prior

$$p(\alpha_m \mid a, b) = \frac{b^a}{\Gamma(a)} \alpha_m^{a-1} e^{-b\alpha_m} \tag{37}$$

has two hyperparameters, $a > 0$ and $b > 0$, which respectively control the shape and width of the density. This prior, which leads to a Student-t $p(w_m)$, is considered by (Tipping, 2001; Wipf and Rao, 2005). However it is not clear how the hyperparameters a and b are to be chosen, except by cross-validation which is a data and time consuming procedure or via a variational approach, which, in the formulation that corresponds closest to the classical RVM Bishop and Tipping (2000), is neither computationally efficient nor of clear practical value[11]. Furthermore, as illustrated in Figure 9, it is possible for the Gamma prior with particular values of a and b to yield $\hat{\alpha}_i < \alpha_i^\star$, that is a MAP α_i that is less sparse than the $\alpha_m^\star$ which maximizes the likelihood alone. By contrast the smoothness prior always results in $\hat{\alpha}_i > \alpha_i^\star$ and we point out that the smoothness prior is strictly increasing and so always assigns increasing weight to increasing precisions (c.f. Figure 4).

The Jeffreys' prior, a uniform density on the logarithmic scale, is obtained in the limit $a, b \to 0$, which (Tipping, 2001) appears to advocate for the RVM although the fRVM (Tipping and Faul, 2003) clearly uses a uniform prior in "un-logged space"—called the **None** prior here. In this limit the Student-t density for $p(w_i)$ becomes $1/|w_i|$. However, the analogous component-wise maximization scheme leads to models in which all components are active when $a < 1/2$ because $\hat{\ell}(\alpha_i)$ is maximized at $\alpha_i = 0$ regardless of the values of s_i, q_i and b (see Appendix A.5). Approaching the Jeffreys' prior by $p(\alpha_i) \propto \alpha_i^\zeta$ as $\zeta \to -1$ leads to the same conclusion: for $\zeta < -1/2$ every component is active because $\hat{\ell}(\alpha_i)$ is maximized at $\alpha_i = 0$ (Appendix A.5).

Thus although the scale invariance and the absence of hyperparameters of Jeffrey's prior is appealing, the type II MAP solution sought here does not accommodate it. However, the smoothness prior, which is noise dependent and therefore confers scale invariance with invariant SNR, has a single hyperparameter and is readily interpretable in terms of the solution sparseness, the degrees of freedom in the smoothing matrix.

Before we proceed to discuss $\boldsymbol{\alpha}$-priors from a wavelet-shrinkage perspective, first a short digression that readers already familiar with wavelet shrinkage may prefer to skip; in-depth treatments of wavelets and wavelet shrinkage can be found in Mallat (1999) and Jansen (2001).

Wavelet shrinkage We have already mentioned that the Discrete Wavelet Transform (DWT), which provides an orthogonal decomposition of a signal into components localized in both frequency and time, is extremely fast – $O(N)$. Algebraically, however, the DWT of a vector $\mathbf{t}$ is simply equivalent to $\mathbf{W}^T\mathbf{t}$, where $\mathbf{W}$ is an orthogonal matrix.

Thus, since the discrete wavelet transform is an orthogonal linear operator, it is easily verfied that it maps stationary white noise on the targets to stationary white noise of the same amplitude on the wavelet coefficients. On the other hand, for "reasonable" noise-free curves (i.e. signals that can be well approximated by piecewise polynomials of a small degree) and a suitable choice of wavelet, wavelet transforms are said to be *decorrelating*. In other words whilst noise enters equally into all wavelet coefficients, the true signal carried by the targets will be mostly concentrated in but a few.

[11] See (Tipping, 2001, footnote 6) or <http:www.miketipping.com/index.php?page=rvm>: "Note that the 'variational' relevance vector machine is pretty much identical to the non-variational version, but is a lot slower to train."

This suggests the following template for wavelet-based denoising: transform to wavelet space ($\mathbf{w_t} = \mathbf{W}^T\mathbf{t}$), somehow cull or *shrink* those coefficients that contain largely noise leaving the signal carrying ones mostly intact and transform back ($\hat{\mathbf{y}} = \mathbf{W}\,\mathrm{shrink}(\mathbf{w_t})$).

But taking $\mathbf{\Phi} = \mathbf{W}$, $\mathbf{W}\,\mathrm{shrink}(\mathbf{w_t})$ is really just a regularized projection from (10) – with regularization coefficients $\sigma^2\boldsymbol{\alpha}$ controlling the amount of shrinkage and the added bonus that the covariance matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ becomes $\mathbf{I}$ and multiplications by $\mathbf{\Phi}$ can be carried out in linear time. So as long as our RVM-way of determining $\sigma^2\boldsymbol{\alpha}$ gives values that also work well for wavelets, wavelet shrinkage can be recognised a special case; as we have shown, this is the case for the sRVM, but not the plain RVM which will hopelessly overfit.

A number of approaches to shrinking the wavelet coefficients have been devised. A straightforward idea is to just set to zero those coefficients whose absolute values remains below a certain threshold τ, i.e. set $\alpha_i = \infty$ for all $|w_i| < \tau$ (*hard thresholding*). Additionally reducing all the other coefficients towards zero by said threshold τ is another, often preferable, alternative (*soft thresholding*; inter alia it gives a continuous shrinkage curve which is analytically more convenient) (Donoho and Johnstone, 1994).

$\boldsymbol{\alpha}$ priors for wavelets Apart from these "classical" wavelet shrinkage approaches (Jansen, 2001, see, e.g.), a variety of Bayesian schemes have also been applied to yield graduated shrinkage. These, like our scheme, commonly impose a heavy-tailed prior, such as a Student-t (Vidakovic, 1998b) or mixture of two zero-mean Gaussians (Chipman et al., 1997), on the wavelet coefficients. See (Vidakovic, 1998a) and (Denison et al., 2002, sec. 3.4) for extensive reviews. The Holmes and Denison (1999) smoothness prior is also suitable for non-wavelet kernels because, unlike most popular wavelet shrinkage priors, it is not dependent on the wavelet length scale or level. Holmes and Denison reject such level dependence as inconsistent with the knowledge that noise enters additively across all components, but there is, in principle, no reason not to incorporate priors in the RVM that only work in conjunction with certain kernel types.

Finally we note that a further alternative hierarchical prior to address the under-determination of the $\mathbf{w}$ is explored in (Fokoué et al., 2004), while Girolami and Rogers (2005) (and references therein) pursue a completely different avenue: a Bayesian treatment of kernel construction itself.

5.2 Summary and Conclusion

We have presented a straightforward extension to the RVM that imposes a more stringent prior on the variance of the weights in nonlinear regression, and we have described an efficient algorithm for maximizing the marginal posterior probability of the model. The RVM with a smoothness prior is also easily adapted to handle classification problems.

From a theoretical perspective we have seen that unlike other proposed prior types (such as the implicit uniform prior in the original RVM implementation, or a Gamma prior) the smoothness prior we presented is noise-dependent in a principled fashion (data/kernel rescaling whilst keeping the SNR fixed does not change the result and, as one would expect, setting $\hat{\sigma}^2$ to a multiple or fraction of the real σ^2 in experiments results respectively in a sparser or less sparse regression).

Further, our results indicate that symmlets with a smoothness prior make an attractive default choice for RVM regression tasks: the combination is flexible enough to be suitable for a large variety of signals, requires no additional kernel parameters to be determined by cross-validation (e.g. scale for Gaussian kernels). The hyperparameter c could be optimised by cross-validation, but our experiments show that the BIC choice works well for a wide range of problems. The sRVM has attractive computational characteristics resulting from the properties of wavelets. In particular the matrix-multiplication by kernel columns can be carried out by the mathematically equivalent but much more efficient discrete wavelet transform ($O(N)$!); this implies that no $N \times M$ design matrix needs to be constructed and held in memory and that the per-step time complexity drops from cubic in S to linear in N. Furthermore numerical robustness also tends to be better than for many other kernels.

This might seem to beg the question "why not just use wavelet shrinkage to start with?" – of course there are limitations of wavelets that other types of kernels do not share (the data must be equally spaced) and although symmlets perform well across a wide range of signals one might find in practice, it is difficult to beat the performance of less general kernels for tasks for which they are particularly well suited (e.g. lsplines for Sinc-like data).

But the deeper point is that the RVM updated with a smoothness prior (sRVM) can be profitably regarded as a *generalization* of wavelet shrinkage.

Figure 7 demonstrates that we can even obtain the best of both worlds in *one and the same experiment* by using an overcomplete dictionary composed of different kernel types (such as Haar wavelets and thin-plate splines) that each capture certain aspects of the overall signal particularly well and then rely on the sRVM to automatically select a sparse representation from this overcomplete dictionary.

In other words a chief attraction of the sRVM is that spans a bridge between the RVM and related methods on the one hand and wavelet shrinkage on the other, yielding a powerful synthesis.

References

Arfken, G. (1985). *Mathematical methods for physicists*. Academic Press.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. Wiley.

Bishop, C. and Tipping, M. (2000). Variational relevance vector machines. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53.

Bobin, J., Moudden, Y., Starck, J.-L., and Elad, M. (2005). Multichannel morphological component analysis. In *Proceedings of Spars05*, pages 103–106, Rennes, France.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92:1413–1421.

Clarkson, E. and Barrett, H. (2001). High-pass filters give histograms with positive kurtosis. *Optics Letters*, 26(16):1253–1255.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.

Denison, D., Holmes, C., Mallick, B., and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

D'Souza, A., Vijayakumar, S., and Schaal, S. (2004). the Bayesian backfitting relevance vector machine. In *proceedings of the international conference on machine learning (ICML 2004)*.

Faul, A. and Tipping, M. (2002). Analysis of sparse Bayesian learning. In *Advances in Neural Information Processing Systems*, volume 14.

Figueiredo, M. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159.

Fokoué, E., Goel, P., and Sun, D. (2004). A Prior for Consistent Estimation for The Relevance Vector Machine. Technical report, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, USA.

Girolami, M. and Rogers, S. (2005). Hierachic Bayesian models for kernel learning. In *22nd International Conference on Machine Learning (ICML 2005)*, pages 241–248, Bonn.

Golub, G. H. and van Loan, C. E. (1989). *matrix computations*. John Hopkins Press, 2nd edition.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall, London.

Hoerl, A. and Kennard, R. (1970). "ridge regression: Biased estimation for nonorthogonal problems". *Technometrics*, 12:55–67.

Holmes, C. and Denison, G. (1999). Bayesian wavelet analysis with a model complexity prior. In *Bayesian statistics 6: Proceedings of the sixth Valencia international meeting*, pages 769–776, Oxford.

Jansen, M. (2001). *Noise reduction by wavelet thresholding*. Springer, New York.

Lam, E. and Goodman, J. (2000). A mathematical analysis of the DCT coefficient distributions for images. *IEEE Trans. Image Processing*, 9:1661–1666.

MacKay, D. (1992). The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736.

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic Press, 2nd edition.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge, 2 edition.

Quiñonero-Candela, J. (2004). *Learning with Uncertainty – Gaussian Processes and Relevance Vector Machines*. PhD thesis, Technical University of Denmark, Lyngby, Denmark.

Roweis, S. (1999). Matrix identities. Available from `http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf`.

Schmolck, A. and Everson, R. (2005). Smoothness priors for sparse Bayesian regression. In *Workshop on Signal Processing with Adaptative Sparse Structured Representations*, Rennes, France.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT Press, Cambridge, Mass.

Smola, A. and Schmölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In Langley, P., editor, *Proceedings of the 17th International Conference on Machine Learning*, pages 911–918, San Fransisco. Morgan Kaufman.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58:267–288.

Tipping, M. (2000). The relevance vector machine. In Solla, A., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 652–658, Cambridge, Mass. MIT Press.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Tipping, M. and Faul, A. (2003). Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003*.

Vidakovic, B. (1998a). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93:173–179.

Vidakovic, B. (1998b). Wavelet-based non-parametric Bayes Methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 133–155. Springer.

Wipf, D. and Rao, B. (2004). Perspectives on sparse Bayesian learning. In *Advances in Neural Information Processing Systems 16*.

Wipf, D. and Rao, B. (2005). Finding sparse representations in multiple response models via Bayesian learning. In *Proceedings of Spars05*, pages 155–158, Rennes.

A Appendix

A.1 Kernel functions

For completeness, here we list the kernels generating the basis functions used in this paper.

$$K_{\text{lspline}}(x_m, x_n) = 1 + r^{-2}x_m x_n + r^{-3}x_m x_n \min(x_m, x_n) - \frac{x_m + x_n}{2} r^{-3} \min(x_m, x_n)^2 + r^{-3}\frac{\min(x_m, x_n)^3}{3} \tag{38}$$

$$K_{\text{gauss}}(x_m, x_n) = \exp(-(x_m - x_n)^2/r^2) \tag{39}$$

$$K_{\text{tpspline}}(x_m, x_n) = |x_m - x_n|^2 r^{-2} \log(|x_m - x_n + \delta_{mn}|/r) \tag{40}$$

Here r sets the width of the kernel. The quoted values of r are relative to data defined so that $-10 \leq x \leq 10$. We use the shorthand "gauss 3.0" etc. in the text, to denote the gaussian kernel defined below with a width parameter $r = 3.0$.

The symmlet family of wavelets is due to Daubechies, all our examples use the symmlet8 wavelet (the 8 here is not a width parameter) from that family as defined in (Daubechies, 1992). For a general discussion of wavelet bases such as Haar and symmlets see e.g. (Mallat, 1999).

A.2 Scale invariance for constant SNR

The sRVM scheme is invariant to scaling of the signal amplitude provided that the noise variance is also scaled so that the SNR remains constant. Suppose that the targets are scaled so that $\mathbf{t} \to k\mathbf{t}$ and the noise is scaled so that $\sigma^2 \to k^2\sigma^2$, then if the coefficient precisions α_i are scaled as $\alpha_i \to k^{-2}\alpha_i$ then the log posterior $\mathcal{L}(\boldsymbol{\alpha}) + \log p(\alpha \,|\, \sigma^2)$ changes only by an additive constant. To see this, note that the matrix $\mathbf{C}$ (25) becomes

$$\tilde{\mathbf{C}} = k^2\sigma^2\mathbf{I} + \sum_m \alpha_m^{-1} k^2 \boldsymbol{\phi}_m \boldsymbol{\phi} \tag{41}$$

$$= k^2\mathbf{C} \tag{42}$$

Consequently, from (22), $\mathcal{L}(\boldsymbol{\alpha})$ becomes:

$$\tilde{\mathcal{L}}(\boldsymbol{\alpha}) = -\frac{1}{2}\left[N \log 2\pi + 2M \log k + \log|\mathbf{C}| + (k\mathbf{t}^T)(k^{-2}\mathbf{C}^{-1})(k\mathbf{t})\right] \tag{43}$$

$$= \mathcal{L}(\boldsymbol{\alpha}) + 2M \log k \tag{44}$$

When the α_i are rescaled along with σ^2, it is clear that DF in (13) remains unchanged. Thus the prior is invariant to simultaneous rescaling of $\boldsymbol{\alpha}$ and σ^2 and so the log posterior is

$$\log p(\boldsymbol{\alpha}/k^2, k^2\sigma^2 \,|\, k\mathbf{t}) = \log p(\boldsymbol{\alpha}, \sigma^2 \,|\, \mathbf{t}) + 2M \log k \tag{45}$$

Maximum a posteriori solutions for $\boldsymbol{\alpha}$ in the scaled case are thus just the MAP solutions in the unscaled case divided by k^2.

A.3 Uniqueness of local maximum

Here we show that $\hat{\ell}(\alpha_i)$ (33) can have at most a single maximum $\hat{\alpha}_i < \infty$ and $\alpha_i^\star < \hat{\alpha}_i$. We drop the explicit indication of which basis function is being dealt with and it is convenient to work in terms of the noise precision $\beta = \sigma^{-2}$.

The derivative of $\hat{\ell}(\alpha)$ is given by (34) in which the cubic polynomial $P(\alpha) = B_3\alpha^3 + B_2\alpha^2 + B_1\alpha + B_0$ has coefficients

$$B_0 = s^2\beta^2 \tag{46}$$

$$B_1 = s\beta^2 + 2\beta s^2 - \beta^2 q^2 + 2s^2 c\beta \tag{47}$$

$$B_2 = 2s\beta + s^2 - 2\beta q^2 + 4s\beta c \tag{48}$$

$$B_3 = s - q^2 + 2c\beta \tag{49}$$

We note that the numerator of (34) is positive for all $\alpha > 0$ so the turning points of $\hat{\ell}(\alpha)$ can be found by examining $P(\alpha)$. A crucial quantity turns out to be the sign of B_3 and we treat positive and negative cases separately. First note that

$$2\hat{\ell}'(\alpha) = \frac{1}{2\alpha(\alpha+s)^2(\alpha+\beta)^2}\left\{[s - q^2 + 2c\beta]\alpha^3 + \text{H.O.T.}\right\} \tag{50}$$

so that the gradient at infinite α is always zero. Also $\lim_{\alpha\to\infty}\hat{\ell}(\alpha) = 0$ and so $\hat{\ell}(\alpha)$ is asymptotic to zero from below if $B_3 > 0$ or from above if $B_3 < 0$. As $\alpha \to 0$ then $\hat{\ell}(\alpha) \to -\infty$.

A.3.1 Asymptote from below:

Since $B_3 > 0$ the graph of $P(\alpha) \to \pm\infty$ as $\alpha \to \pm\infty$, and $P(0) = B_0 > 0$. Consequently, $P(\alpha)$ has a least one root for $\alpha < 0$. It must therefore either have zero or two positive roots.

If P has no positive roots the maximum of $\hat{\ell}(\alpha)$ is achieved at infinity (e.g. top-left Figure 4).

If there are two positive roots, one corresponds to a maximum and the other to a minimum. Ignoring the degenerate case of an inflexion point, the root for smaller α must be the maximum and the root for larger α is a minimum. The maximum may be greater or less than the asymptotic value, as illustrated by the bottom row of Figure 4.

A.3.2 Asymptote from above:

In this case since $B_3 < 0$ there is at least a single positive root of $P(\alpha)$. Since $\hat{\ell}(\alpha)$ is asymptotic to zero from above at infinity this root must correspond to a maximum. However, we must ensure that there cannot be 3 positive roots.

The derivative of $\hat{\ell}(\alpha)$ can be written as the sum of the derivatives of $\ell(\alpha)$ and $\log p(\alpha_i \,|\, \sigma^2)$ as follows:

$$\hat{\ell}' = \frac{[s^2 + (s - q^2)\alpha](\beta+\alpha)^2 + 2c\beta\alpha(s+\alpha)^2}{2\alpha(\alpha+s)^2(\alpha+\beta)^2} \tag{51}$$

$$\equiv \frac{P_0(\alpha) + 2c\beta\alpha(s+\alpha)^2}{2\alpha(\alpha+s)^2(\alpha+\beta)^2} \tag{52}$$

where $P_0(\alpha)$ is the cubic $P(\alpha)$ with $c = 0$. It has a root at $\alpha^\star$, which corresponds to the maximum in the likelihood, and there is an additional repeated root at $\alpha = -\beta$.

The term $2c\beta\alpha(s+\alpha)^2$ arising from the prior is a cubic with a root at $\alpha = 0$ and a repeated root at $\alpha = -s < 0$. It is clearly positive for all $\alpha > 0$. There is therefore a root of $P(\alpha)$ at some $\hat{\alpha} > \alpha^\star$ because $P(\alpha) \geq P_0(\alpha)$ for all $\alpha > 0$. Since $P_0(\alpha)$ is monotonically decreasing for $\alpha > \alpha^\star$, while $c\beta\alpha(s+\alpha)^2$ is monotonically increasing they can only intersect once (at $\hat{\alpha}$) so there can be no roots for $\alpha > \hat{\alpha}$ and we conclude that there can only be a single maximum of $\hat{\ell}(\alpha)$ for $\alpha > 0$ in this case, which is illustrated in the bottom right panel of figure 4.

A.4 Saddle-points of $\hat{\mathcal{L}}$

In order to determine the nature of the maxima of $\hat{\mathcal{L}}$ the Hessian of $\mathcal{L}$ is required. As shown in (Faul and Tipping, 2002), the off-diagonal terms of the Hessian are:

$$\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\alpha}})}{\partial\alpha_i \partial\alpha_j} = \frac{\boldsymbol{\phi}_i^T \mathbf{C}^{-1} \boldsymbol{\phi}_j}{2\alpha_i^2 \alpha_j^2} \left[\boldsymbol{\phi}_i^T \mathbf{C}^{-1} \boldsymbol{\phi}_j - 2(\boldsymbol{\phi}_i^T \mathbf{C}^{-1}\mathbf{t})(\boldsymbol{\phi}_j^T \mathbf{C}^{-1}\mathbf{t})\right] \quad i \neq j \tag{53}$$

When the basis functions are orthogonal the matrices $\mathbf{C}$ and therefore $\mathbf{C}^{-1}$ are diagonal (25). Consequently $\boldsymbol{\phi}_i^T \mathbf{C}^{-1} \boldsymbol{\phi}_j$ and hence the off-diagonal terms of the Hessian are zero. At a solution located through the maximization procedure the diagonal elements of the Hessian corresponding to finite α_i maxima are necessarily negative. The Hessian is therefore positive semi-definite, with the zeros corresponding to the infinite α_m, switched-off components. Unfortunately, the demonstration (Faul and Tipping, 2002) that the Hessian of the log marginal likelihood is negative semi-definite with general basis functions appears to be flawed and we are unable to provide any assurance that joint optimization of two or more α_i might not yield a better result than successive maximization with respect to each.

A.5 Approaching the Jeffreys' prior

With $p(\alpha_i) = Z\alpha^\zeta$ the contribution of the ith component to the posterior becomes:

$$\hat{\ell}(\alpha_i) = \frac{1}{2}\left[\log\alpha_i - \log(\alpha_i + s) + \frac{q^2}{\alpha_i + s}\right] + \log Z + \zeta\log\alpha_i \tag{54}$$

$$= \frac{1}{2}\left[(1+2\zeta)\log\alpha_i - \log(\alpha_i + s) + \frac{q^2}{\alpha_i + s}\right] + \log Z \tag{55}$$

Setting the derivative to 0 to find the MAP solution $\hat{\alpha}_i$ we get:

$$\hat{\ell}'(\alpha_i) = \frac{1}{2}\left[\frac{1+2\zeta}{\alpha_i} - \frac{1}{\alpha_i + s} - \frac{q^2}{(\alpha_i + s)^2}\right] = 0 \tag{56}$$

From this it becomes apparent that $\hat{\ell}$ has no turning points for finite, positive α_i if $\zeta < -\frac{1}{2}$ because α_i, q^2 and s are always positive. Furthermore, in this case since $\hat{\ell}'(\alpha_i) < 0$ the maximum of $\hat{\ell}(\alpha_i)$ is achieved at $\alpha_i = 0$. As a Jeffreys' prior corresponds to $\zeta = -1$, one would always obtain a model in which all the components are active.

With a Gamma prior for α_m (37) the derivative of the contribution to the log posterior from the ith basis function is:

$$\hat{\ell}'(\alpha_i) = \frac{1}{2}\left[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + s} + \frac{q^2}{\alpha_i + s}\right] + \frac{a-1}{\alpha_i} - b \tag{57}$$

$$= \frac{1}{2}\left[\frac{2a-1}{\alpha_i} - \frac{1}{\alpha_i + s} - \frac{q^2}{(\alpha_i + s)^2} - 2b\right] \tag{58}$$

In this case is it clear that $\hat{\ell}$ has no turning points for finite, positive α_i if $a < 1/2$, although there will be no positive α_i for larger a when b is larger. Consequently the Gamma prior forces all components to be active when $a < 1/2$, which of course includes the Jeffreys' prior in the limit $a, b \to 0$.

A.6 Efficiently calculating the full likelihood and posterior

As $\mathbf{C}$ is a $N \times N$ matrix, its inversion and computation of the determinant are very costly procedures – $O(N^3)$. Fortunately with the help of the Woodbury-Sherman-Morrison matrix inversion and determinant identities (see e.g. (Press et al., 1992) or (Roweis, 1999)) and using (8) and (21) the marginal log likelihood may instead be advantageously expressed as:

$$\mathcal{L} = -\frac{1}{2}\left[N\log(2\pi) + N\log\sigma^2 - \sum_S \log\alpha_s + \log|\mathbf{\Sigma}| + \left(\sigma^{-2}\mathbf{t}^T\mathbf{t} - \boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \tag{59}$$

The efficient expression for $\hat{\mathcal{L}}$, the log posterior is then given by

$$\hat{\mathcal{L}} = \mathcal{L} + \sum_S \frac{-c}{1 + \sigma^2\alpha_s} + S\log Z + \log\mathcal{IG}(\sigma^2 \mid g, h) \tag{60}$$

and may be computed in $O(S^3)$ time, where S is the number of *included* components (as the matrix $\mathbf{\Sigma}$ is $S \times S$). Although the stated algorithm does not depend on it this expression is useful for obtaining the posterior likelihood of the solution the sRVM algorithm finds and may also be used for convergence testing.

B Notation and Glossary

Matrices are bold-uppercase ($\mathbf{\Phi}$), vectors are bold lowercase and columns from matrices are the subscripted bold lowercase equivalent ($\boldsymbol{\phi}_i$ for the ith column of $\mathbf{\Phi}$). Hats refer to posteriors (e.g. $\hat{\mathbf{t}}$), $_{\setminus i}$ means with the influence of the ith component removed.

$\mathbf{y}^{N\times 1}$	the true signal (2)
$\mathbf{t}^{N\times 1}$	the observations or targets (2)
$\boldsymbol{\epsilon}$	the error (the difference between observations and true signal) (2)
σ	the standard deviation of the error (2), (5)
β	the noise precision ($\beta = \sigma^{-2}$), (sec. A.3)
$\hat{\mathbf{y}}^{N\times 1}$	the posterior mean prediction for the true signal (10)
$\mathbf{\Phi}^{N\times M}$	the design matrix, viz. the kernel, viz. a dictionary of basis functions (3)
$\mathbf{\Phi}_S^{N\times S}$	the selected components of the design matrix
N	the number of target observations
M	the number of basis functions viz. components
S	the number of included basis functions viz. components (or non-zero w_m or finite α_m)
$\mathcal{S}$	the set of currently included components
$\mathbf{w}^{M\times 1}$	the weights ($\mathbf{\Phi w} = \mathbf{y}$)
$\mathbf{\Sigma}^{M\times M}$	the posterior covariance matrix of the weights (8)
$\mathbf{\Sigma}_S^{S\times S}$	the posterior covariance matrix of the included weights
$\boldsymbol{\mu}^{M\times 1}$	the posterior mean of the weights (9)
$\boldsymbol{\alpha}^{M\times 1}$	the precisions of the weights (4)
$\mathcal{L}(\boldsymbol{\alpha})$	the log-likelihood of the observations, $\log p(\mathbf{t} \mid \boldsymbol{\alpha}, \sigma^2)$, (22)
$\ell(\alpha_i)$	the contribution of the ith component to the likelihood $\mathcal{L}(\boldsymbol{\alpha})$ (18), (29)
$\mathcal{L}(\boldsymbol{\alpha}_{\setminus i})$	the log-likelihood of the observations without the contribution of the ith component(22),(59)
$\hat{\mathcal{L}}(\boldsymbol{\alpha})$	the log-posterior of the weight precisions $\boldsymbol{\alpha}$ (60)
$\hat{\ell}(\alpha_i)$	the contribution of the ith component to the posterior $\hat{\mathcal{L}}(\boldsymbol{\alpha})$ (33)
$\alpha_i^\star$	the value of α_i that maximizes $\hat{\ell}$ (21)
c	the hyperparameter that controls the degree of smoothing in the smoothness prior $p(\boldsymbol{\alpha}\|\sigma^2)$ (14)
g, h	hyperparameters for the inverse Gamma distribtuion over σ^2 (5)
DF	the degrees of freedom of the smoothing matrix $\mathbf{S}$ hence an indicator of the complexity of the model (12), (13)
$\mathbf{S}^{N\times N}$	the smoothing matrix (10)
$\mathbf{C}^{N\times N}$	covariance matrix of the data likelihood (25)
$\mathbf{C}_{\setminus i}^{N\times N}$	$\mathbf{C}$ without the influence of the ith component (25)
s_i, q_i	convenience variables that can be thought of as indices of sparsity and quality of component i (30), (31)