SMOOTHNESS PRIORS FOR SPARSE BAYESIAN REGRESSION

Alexander Schmolck *Richard Everson*

School of Engineering, Computer Science and Mathematics
University of Exeter. UK

ABSTRACT

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 smoothing is to a large extent controlled by the choice of kernel. This can lead to severe overfitting (or oversmoothing).

We detail an efficient scheme to incorporate flexible sparsity priors into the RVM and present an empirical evaluation of the effects of choice of prior structure on a selection of popular data sets and elucidate the link between wavelet shrinkage and RVM regression.

We find that a smoothness prior with symmlet wavelets yields good performance across a wide spectrum of problems for low computational costs as leveraging special properties of wavelets allows for considerable computational savings.

1. INTRODUCTION

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

$$y(\mathbf{x}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) \tag{1}$$

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 weights w_m that will yield the most faithful approximation to y.

Writing the targets as an N-vector and w_m, the weights, as an M-vector, (2) is conveniently written as $\mathbf{y} = \mathbf{\Phi}\mathbf{w}$, with design matrix $\mathbf{\Phi}$. 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 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/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):

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

with an 'automatic relevance determination' prior

[MacKay, 1992] over the weights:

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

that has the effect of 'switching off' basis functions for which there is no evidence in the data.

A standard inverse gamma prior is placed over the noise variance σ^2:

$$p(\sigma^2) = \mathcal{IG}(\sigma^2 \mid a, b) \qquad (5)$$

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

Finally, the values of σ and $\boldsymbol{\alpha}$ are determined by a type II likelihood maximization scheme [Tipping, 2001, Tipping and Faul, 2003].

Unfortunately the RVM in a sense 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 does not need that many degrees of freedom will still result in severe overfitting (see Fig. 1), so that a crucial aspect of sparsity control (kernel choice) remains outside the principled probabilistic framework.

Fortunately a strength of Bayesian models is their inherent extensibility by means of additional prior structure; here we examine a *smoothness prior* for RVM models. See [Girolami and Rogers, 2005] for another possible avenue: a Bayesian treatment of kernel construction itself.

1.2. A detailed look at sparsity priors

On its own (4) does not appear to strongly favour sparsity, but of course the overall effect depends on the prior $p(\boldsymbol{\alpha} \mid \sigma^2)$. 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 (Fig. 1), a well-founded, sparser prior over $\boldsymbol{\alpha} \mid \sigma^2$ is desirable.

As our desire for sparse $\mathbf{w}$ is ultimately grounded in beliefs about the complexity and structure of the signal $\mathbf{t}$, 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 $\hat{\mathbf{t}}$ reflects these beliefs.

Given the the posterior over the weights

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2) &= \frac{p(\mathbf{w}|\mathbf{t}, \sigma^2) p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\alpha, \sigma^2)} \\ &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned} \qquad (6)$$

with

$$\boldsymbol{\Sigma} = (\sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathrm{diag}(\boldsymbol{\alpha}))^{-1} \qquad (7)$$

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} \qquad (8)$$

we obtain

$$\hat{\mathbf{t}} = \boldsymbol{\Phi}\boldsymbol{\mu} = (\boldsymbol{\Phi}\sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\mathbf{t} \equiv \mathbf{S}\mathbf{t} \qquad (9)$$

where $\mathbf{S}$ is known as the smoothing matrix. The smoothing power of $\mathbf{S}$ is typically quantified by its degrees of freedom given by its trace [Hastie and Tibshirani, 1990], so

$$\mathrm{DF} = \mathrm{tr}\mathbf{S} = \sum_{i=1}^{N} (1 + \sigma^2\alpha_i)^{-1}. \qquad (10)$$

These observations lead Holmes and Denison [1999] to choose the following prior structure for encoding sparsity beliefs for the related problem of wavelet shrinkage:

$$p(\alpha_i \mid \sigma^2) \propto e^{-c\left(1+\sigma^2\alpha_i\right)^{-1}}. \qquad (11)$$

Holmes and Denison relate different choices for the parameter c to different classical model choice criteria:

c	
0	**None**, Bayes Factor
1	**AIC**, Akaike information criterion
$\ln(N)/2$	**BIC**, Bayesian information criterion
$\ln(N)$	**RIC**, Risk inflation criterion

Note that basis functions with smaller α_i have greater prior support when the noise is smaller. It should further be noted that a uniform $p(\boldsymbol{\alpha} \mid \sigma^2)$ as in the original RVM implementation is just a special case of the above smoothness prior with $c = 0$.

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

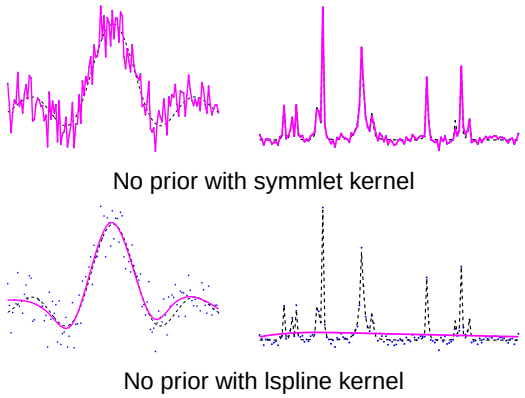

Figure 1: The effect of kernel choice on the smoothness of the regression result when there is no prior over $\boldsymbol{\alpha}$. Legend: dots:data $\mathbf{t}$; dashed line:true signal $\mathbf{y}$; solid line:prediction $\hat{\mathbf{t}}$. In the classical RVM choosing a flexible symmlet-wavelet kernel results in drastic overfitting for the Sinc data set (top row 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. However an lspline kernel cannot resolve the Bumps data (on the right; N=128, SNR=7.0) at all.

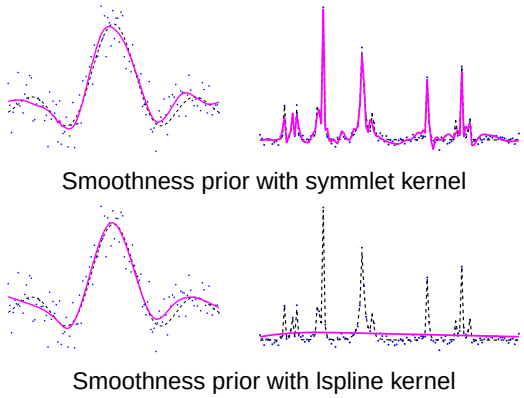

Figure 2: The smoothness prior means that enforcing sparsity is no longer mostly relegated to the choice of kernel. A symmlet kernel no longer results in drastic overfitting for the Sinc data set (on the left). The bottom row shows that the smoothness prior typically has no adverse effect when smoothing is already mandated by the kernel. The data sets are identical to Fig. 1.

2. IMPLEMENTATION

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

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i) \tag{12}$$

in order to separate out the contribution of the ith basis function $\boldsymbol{\phi}_i$ into a term for which the computational effort of maximization scales with the number S of basis functions included in the model, rather than with M, the total number of basis functions.

The addition of the smoothness prior means that $\mathcal{L} \not\propto \ln p(\boldsymbol{\alpha}, \sigma \mid \mathbf{t})$, but although the required additional term requires that the optimal

$$\alpha_i^* = \max_{\alpha_i} \left[\ell(\alpha_i) + \ln p(\alpha_i, \sigma)\right] \tag{13}$$

is found numerically, rather than analytically as in Tipping and Faul [2003], the extension is straightforward and has the desired properties. Similar adjustments have to be made for the noise reestimation; details will be given in a forthcoming paper.

The RVM with a smoothness prior is also easily adapted to handle classification problems.

3. RESULTS

As Fig. 2 shows, we find that use of the smoothness prior typically yields substantial improvements for tasks where overfitting is a problem due to the multiscale resolution of the kernel, while it generally has no appreciable negative impact when overfitting is not an issue .

Table 1 shows for a number of standard datasets the sparsity, measured by the number of included components S, and the MSE between $\hat{\mathbf{t}}$ and the true signal y. Clearly the **None** prior is 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.

Bumps SNR=2.0 ($\sigma^2 = 0.119$)				
Kernel	Prior	S	MSE	σ^2_{MAP}
symmlet	None	127.0± 0.0	0.127±0.018	0.000±0.000
symmlet	**AIC**	98.1±16.1	**0.120±0.021**	0.008±0.009
symmlet	BIC	13.3± 3.0	0.145±0.023	0.217±0.036
symmlet	RIC	4.4± 2.1	0.269±0.061	0.373±0.091
Bumps SNR=7.0 ($\sigma^2 = 0.010$)				
Kernel	Prior	S	MSE	σ^2_{MAP}
symmlet	None	127.1± 0.3	0.010±0.001	0.000±0.000
symmlet	**AIC**	86.0±10.2	**0.009±0.002**	0.003±0.002
symmlet	BIC	41.0± 5.8	0.019±0.007	0.022±0.009
symmlet	RIC	8.5± 1.8	0.165±0.028	0.175±0.031
Sinc SNR=2.0 ($\sigma^2 = 0.031$)				
Kernel	Prior	S	MSE	σ^2_{MAP}
gauss	None	5.4± 1.0	0.004±0.001	0.033±0.005
gauss	**AIC**	5.5± 1.1	**0.004±0.001**	0.034±0.006
gauss	BIC	5.1± 0.9	0.005±0.001	0.034±0.006
gauss	RIC	4.8± 0.8	0.005±0.001	0.034±0.006
symmlet	None	127.0± 0.0	0.033±0.005	0.000±0.000
symmlet	AIC	75.2±17.2	0.028±0.006	0.004±0.003
symmlet	BIC	9.2± 2.0	0.007±0.002	0.031±0.005
symmlet	**RIC**	6.1± 0.3	**0.006±0.001**	0.036±0.006

Table 1: Empirical comparisons of different priors on standard datasets. Results are averaged over 10 runs.

4. DISCUSSION

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) and has attractive computational characteristics resulting from the properties of wavelets (the matrix-multiplication by kernel columns can be carried out by the mathematically equivalent but much more efficient discrete wavelet transform – in particular this implies that no $N \times M$ design matrix needs to be constructed and held in memory; there are further simplifications due to orthonormality and numerical robustness also tends to be better than for many other kernels). This might seem to beg the question why not just 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) but the deeper point is that the RVM updated with a smoothness prior (sRVM) can be profitably regarded as a *generalization* of wavelet shrinkage.[1]

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

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

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

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

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

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

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

M. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems*, volume 12, pages 652–658, 2000.

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

M.E. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of Artificial Intelligence and Statistics*, 2003.

[1] The Holmes and Denison [1999] smoothness prior is directly suitable for other types of kernels than wavelets because, unlike most popular wavelet shrinkage priors, it is not level-dependent. 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.