samsi

NSF • Duke • NCSU • UNC • NISS

A Prior for Consistent Estimation for The Relevance Vector Machine

Ernest Fokoue, Prem Goel and Dongchu Sun

Technical Report #2004-16
March 2004

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

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

A Prior for Consistent Estimation for The Relevance Vector Machine

Ernest Fokoué[*] Prem Goel[†] Dongchu Sun[‡]

Draft of First Version - March 2004

Abstract

The Relevance Vector Machine (RVM) provides an empirical Bayes treatment of function approximation by kernel basis expansion. In its original form **?**, RVM achieves a sparse representation of the approximating function by structuring a Gaussian prior distribution in a way that implicitly puts a sparsity pressure on the coefficients appearing in the expansion. RVM aims at retaining the tractability of the Gaussian prior while simultaneously achieving the assumed (and desired) sparse representation. This is achieved by specifying independent Gaussian priors for each of the coefficients. In his introductory paper, **?** shows that for such a prior structure, the use of independent Gamma hyperpriors yields a product of independent Student-t marginal prior for the coefficients, thereby achieving the desired sparsity. However, such a prior structure gives complete freedom to the coefficients, making it impossible to isolate a unique solution to the function estimation task.

At the other extreme, one could think of using a single hyperparameter for all the coefficients in the spirit of traditional regularized function estimation. With such a choice, a Gaussian prior distribution over the coefficients does not yield a sparse representation, and only a Laplacian prior in such a case does imply a sparse representation.

This paper aims at providing a prior structure that achieves a trade-off between the two extremes. The key idea here is to reduce the dimensionality of the hyperparameter space by specifying a prior structure that reflects the possibility of correlation between the hyperparameters of the coefficients distribution. With this, it is possible to isolate a unique solution.

Key words: *Consistency, Neyman-Scott Problem, Reference prior, sparsity*

1 Relevance Vector Regression

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

[*] *Ernest Fokoué is Postdoctoral Research Fellow in the SAMSI Data Mining and Machine Learning program. He is also Assistant Professor of Statistics at Ohio State University.* eMail: *epf@samsi.info*

[†] *Prem Goel is Professor, Department of Statistics, The Ohio State University.* eMail: *goel@stat.ohio-state.edu*

[‡] *Dongchu Sun is Professor, Department of Statistics, University of Missouri.* eMail: *dsun@stat.missouri.edu*

random variable Y_i can be expressed as a weighted sum of the form

$$Y_i = w_0 + \sum_{j=1}^{n} w_i K(\mathbf{x}_i, \mathbf{x}_j) + \epsilon_i \tag{1}$$

For notational convenience, equation (**??**) is often rewritten as

$$\mathbf{y} = \mathbf{\Phi}\tilde{\mathbf{w}} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^\mathsf{T}$, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^\mathsf{T}$, $\tilde{\mathbf{w}}^\mathsf{T} = (w_0, \mathbf{w}^\mathsf{T})$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\mathsf{T}$ and

$$\mathbf{\Phi} = \begin{bmatrix} 1 & K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ 1 & K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \tag{3}$$

In many applications of regression, it often turns out to be reasonable to assume that the noise terms are independent zero-mean Gaussian random variables with the same variance σ^2, i.e. $\epsilon_i \overset{iid}{\sim} \boldsymbol{\mathcal{N}}(0, \sigma^2)$. As a consequence, the likelihood function for an iid sample is Gaussian, ie

$$p(\mathbf{y} \mid w_0, \mathbf{w}, \sigma^2) = \boldsymbol{\mathcal{N}}(\mathbf{y} \mid \mathbf{\Phi}\tilde{\mathbf{w}}, \sigma^2 \mathbf{I}_n) \tag{4}$$

Note that in the above formulation the parameter vector $\boldsymbol{\theta} = (w_0, \mathbf{w}, \sigma^2)$ is $(n+2)-$dimensional while we only have n data points to be used for its estimation.

In fact, since the expansion is over all the vectors in the sample, there will be a gross overfitting. Besides, such an expansion over all the vectors in the sample, makes it hard to control the complexity of the derived model.

A common and indeed reasonable assumption in kernel regression is that many of the coefficients w's will turn out to be zero or at best of very negligible magnitude. This is the key justification for seeking a parsimonious version of equation (**??**).

The desired sparse representation is usually obtained by a careful choice of priors for $\mathbf{w}$ that reflect our belief in the fact that many of the w_i's will have negligible magnitude or even zero values with a consequence that only very few of the coefficients w_i will be needed in the expansion.

It is a well known fact in the literature that the double exponential (Laplace) prior for w_i's does indeed yield such a sparse representation. However, it is also well known that the non-differentiability of the Laplace density at zero poses many computational difficulties that render its use less attractive. **?** provides an insightful account of the use of the Laplace prior in a regression analysis of the type we are considering in this paper.

A natural candidate for the prior over $\mathbf{w}$ is the Gaussian distribution. However, such a prior in and of itself does not naturally yield a sparse representation. For such a prior, something else must be done to achieve sparsity.

1.1 Tipping's prior structure for RVM

Tipping specifies a separate independent zero-mean Gaussian prior for each coefficient w_i, ie

$$p(w_i \mid \alpha_i) = \boldsymbol{\mathcal{N}}(w_i \mid 0, \alpha_i^{-1}) \tag{5}$$

so that $p(\tilde{\mathbf{w}} \mid \boldsymbol{\alpha}) = \boldsymbol{\mathcal{N}}_n(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \mathsf{diag}(\alpha_0^{-1}, \alpha_1^{-1}, \cdots, \alpha_n^{-1})$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n)^{\mathsf{T}}$.

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

$$p(\alpha_i \mid a, b) = \mathsf{Ga}(\alpha_i \mid a, b) \tag{6}$$

the marginal prior for w_i is

$$\begin{aligned} p(w_i) &= \int p(w_i \mid \alpha_i) p(\alpha_i) d\alpha_i & (7) \\ &= \frac{b^a \Gamma(a + \frac{1}{2})}{(2\pi)^{\frac{1}{2}} \Gamma(a)} \left(b + w_i^2/2\right)^{-(a+\frac{1}{2})} & (8) \end{aligned}$$

With such a Student-t marginal prior for each w_i, the prior for the vector $\mathbf{w}$ is a product of independent Student-t distributions over the w_i's. Tipping uses a two-dimensional case to show that with such a prior the probability mass is concentrated both at the origin and along the "spines" where one of the weights is zero.

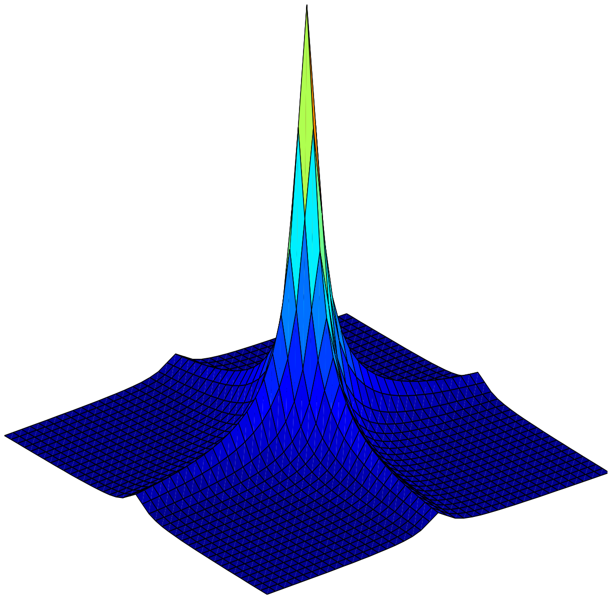

Figure 1: The 2-dimensional marginal prior for $\mathbf{w}$

1.2 A hierarchical prior structure for the RVM

Although Tipping's hierarchical structure ends up yielding a sparse representation, the complete freedom given to the distribution of the parameters leads to the impossibility to find a unique solution. Since the number of parameters grows with the sample size, this is a typical case of the Neyman-Scott problem **?**. In fact, Neyman-Scott show that with a prior like (**??**), the estimates of $\mathbf{w}$ are not consistent unless one imposes some stochastic (exchangeable or partially exchangeable) structure on the coefficients w_i's. If the dimension reduction is achieved via random-coefficient regression structure, one can get consistent estimates of $\mathbf{w}$.

The idea in this section is to add another layer in the hierarchical structure and to reparameterize in a way that reduces the dimensionality of the parameter space. Such a specification is an extension of Tipping's original work, with the advantage that it does isolate a unique solution and provides a characterization of the level of sparsity through a correlation coefficient. We now specify a separate distribution for w_0. More specifically, we treat w_0 as a fixed effect, and we put a constant prior on it. We now have

$$p(w_i \,|\, \alpha_i) = \boldsymbol{\mathcal{N}}(w_i \,|\, 0, \alpha_i^{-1}) \tag{9}$$

so that $p(\mathbf{w} \,|\, \boldsymbol{\alpha}) = \boldsymbol{\mathcal{N}}(0, \Psi)$, where $\Psi = \mathsf{diag}(\alpha_1^{-1}, \cdots, \alpha_n^{-1})$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)^{\mathsf{T}}$. The joint prior density in this case can be written as

$$p(w_0, \mathbf{w}, \sigma^2, \boldsymbol{\alpha}, \mu, \rho, \tau^2) = p(\mathbf{w} \,|\, \boldsymbol{\alpha}) p(\sigma^2) p(\boldsymbol{\alpha} \,|\, \mu, \rho, \tau^2) p(\mu) p(\rho) p(\tau^2). \tag{10}$$

Let's reparameterize $\boldsymbol{\alpha}$ as $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^T$, where $\eta_i = \log(\alpha_i)$. We assume that the following prior for $\boldsymbol{\eta}$,

$$\boldsymbol{\eta} \sim \boldsymbol{\mathcal{N}}_n(\mu \mathbf{1}_n, \tau^2 \boldsymbol{\Sigma}), \tag{11}$$

where

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \rho & \vdots \\ \vdots & \rho & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = (1-\rho)\mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}}. \tag{12}$$

Let $\boldsymbol{A} \in I\!\!R^{p \times p}$ and $\mathbf{u}$ and $\mathbf{v}$ both vectors in $I\!\!R^p$. The following identities hold true:

$$\begin{aligned} |\boldsymbol{A} + \mathbf{u}\mathbf{v}^{\mathsf{T}}| &= |\boldsymbol{A}|(1 + \mathbf{v}^{\mathsf{T}} \boldsymbol{A}^{-1} \mathbf{u}), & (13) \\ [\boldsymbol{A} + \mathbf{u}\mathbf{v}^{\mathsf{T}}]^{-1} &= \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\mathbf{u}\mathbf{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1 + \mathbf{v}^{\mathsf{T}} A^{-1} \mathbf{u}}. & (14) \end{aligned}$$

Using the above identities on $\boldsymbol{\Sigma} \in I\!\!R^{n \times n}$, we have

$$|\boldsymbol{\Sigma}| = |(1-\rho)\mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}}| = (1-\rho)^n [1 + (n-1)\rho], \tag{15}$$

and

$$\boldsymbol{\Sigma}^{-1} = (1-\rho)^{-1} \left[\mathbf{I}_n - \frac{\rho \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}}}{1 + (n-1)\rho} \right]. \tag{16}$$

Consequently,

$$\begin{aligned} p(\boldsymbol{\eta} \,|\, \mu, \rho, \tau^2) &= \frac{1}{|2\pi\tau^2\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{\eta} - \mu \mathbf{1}_n)^T (\tau^2 \boldsymbol{\Sigma})^{-1} (\boldsymbol{\eta} - \mu \mathbf{1}_n) \right) \\ &= \frac{1}{(2\pi\tau^2)^{n/2} (1-\rho)^{n/2} [1 + (n-1)\rho]^{1/2}} \\ &\quad \exp\left\{ -\frac{1}{2\tau^2(1-\rho)} (\boldsymbol{\eta} - \mu \mathbf{1}_n)^T \left[\boldsymbol{I}_n - \frac{\rho \mathbf{1}_n \mathbf{1}_n^T}{1 + (n-1)\rho} \right] (\boldsymbol{\eta} - \mu \mathbf{1}_n) \right\}. \qquad (17) \end{aligned}$$

A possible choice would be the use of noninformative priors for both $\boldsymbol{\mu}$ and τ^2, namely

$$p(\mu, \tau^2) \propto \frac{1}{\tau^2}.$$

However, in this paper, we instead resort to conjugate priors for both μ and τ^2, namely

$$p(\mu \,|\, \delta) = \boldsymbol{\mathcal{N}}(0, \delta^{-1}) \quad \text{and} \quad p(\tau^{-2}) = \mathsf{Ga}(c_1, d_1), \tag{18}$$

for some positive constants c_1 and d_1.

We then use a reference prior for ρ, namely

$$p(\rho) = \frac{\sqrt{1 + (n-1)\rho^2}}{(1-\rho)(1+(n-1)\rho)}, \tag{19}$$

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

2 Derivation of full conditional posteriors

We now derive the full conditional posteriors of $(\tilde{\mathbf{w}}, \sigma^2, \boldsymbol{\eta}, \rho, \mu, \tau^2)$ given $\boldsymbol{y}$.

2.1 Conditional posterior for $\tilde{\mathbf{w}}$

Since we only added new layers to the prior, the full conditional popsterior for $\tilde{\mathbf{w}}$ remains unchanged. Because $(\tilde{\mathbf{w}} \,|\, \sigma^2, \boldsymbol{\alpha}, \mathbf{y}) \propto \boldsymbol{\mathcal{N}}(\boldsymbol{\Phi}\tilde{\mathbf{w}}, \sigma^2 \boldsymbol{I}_n) \boldsymbol{\mathcal{N}}(0, \widetilde{\boldsymbol{\Psi}})$, where $\widetilde{\boldsymbol{\Psi}} = diag(0, \alpha_1^{-1}, \cdots, \alpha_n^{-1})$, we get

$$(\tilde{\mathbf{w}} \,|\, others) \sim \boldsymbol{\mathcal{N}}_{n+1}\left((\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \sigma^2 \widetilde{\boldsymbol{\Psi}})^{-1} \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{y},\ (\sigma^{-2} \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \widetilde{\boldsymbol{\Psi}})^{-1} \right). \tag{20}$$

2.2 Conditional posterior for σ^{-2}

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

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

2.3 Conditional posterior for $\boldsymbol{\alpha}$ or $\boldsymbol{\eta}$

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

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

where

$$p(\mathbf{w} \,|\, \boldsymbol{\alpha}(\boldsymbol{\eta})) \quad = \quad \left(\prod_{i=1}^{n} \frac{e^{\eta_i/2}}{\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} e^{\eta_i} w_i^2 \right), \tag{21}$$

$$p(\boldsymbol{\eta} \,|\, \mu, \rho, \tau^2) \quad \propto \quad \exp\left\{ -\frac{\sum_{i=1}^{n} (\eta_i - \mu)^2}{2\tau^2(1-\rho)} + \frac{\rho \left[\sum_{i=1}^{n} (\eta_i - \mu)\right]^2}{2\tau^2(1-\rho)[1+(n-1)\rho]} \right\}. \tag{22}$$

Fact 1 *The conditional posterior of η_i is log-concave.*

PROOF. The conditional posterior of η_i is

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

Then for a constant C,

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

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

$$\frac{\partial}{\partial\eta_i}[\eta_i \,|\, others] \;=\; -\frac{1}{2}e^{\eta_i}w_i^2-\frac{1+(n-2)\rho}{\tau^2(1-\rho)[1+(n-1)\rho]}. \qquad (23)$$

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

2.4 Conditional posterior for ρ

The reference prior of equation (**??**) allows us to write

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

A approximate discretised version of this conditional posterior density can be used.

2.5 Conditional posterior for μ

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

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

This is

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

2.6 Conditional posterior for τ^{-2}

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

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

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

3 Numerical Experiments

3.1 First example

$$f(\mathbf{x}) = \frac{\sin(\mathbf{x})}{\mathbf{x}} \quad \text{with} \quad \mathbf{x} \in [-10, 10] \tag{24}$$

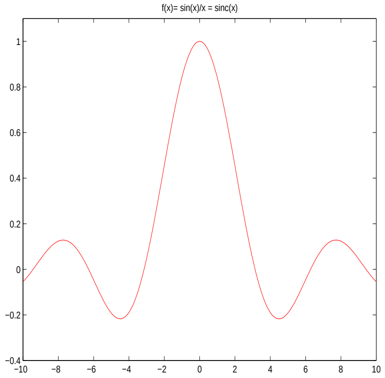

Figure 2: The sinc in one dimensional

For this example, we generated 200 input points from a uniform distribution in $[-10, 10]$, and then we formed the corresponding response values as $y_i = f(\mathbf{x}_i) + \epsilon_i$ where the independent noise terms ϵ_i followed a zero mean Gaussian distribution with standard deviation 0.2.

The thin plate spline kernel in this case is simply

$$K(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|^2 \log\left(\|\mathbf{x}_i - \mathbf{x}_j\|\right) \tag{25}$$

The Laplace kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2r}\right) \tag{26}$$

The Gaussian Radial Basis function kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2r^2}\right) \tag{27}$$

The ”linear spline” kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = 1 + \mathbf{x}_i\mathbf{x}_j + \mathbf{x}_i\mathbf{x}_j \min(\mathbf{x}_i, \mathbf{x}_j) - \frac{\mathbf{x}_i + \mathbf{x}_j}{2} \min(\mathbf{x}_i, \mathbf{x}_j)^2 + \frac{\min(\mathbf{x}_i, \mathbf{x}_j)^3}{3} \tag{28}$$

3.2 Second example

$$f(\mathbf{x}) = (1 - \mathbf{x}^2) \exp\left(-\frac{\mathbf{x}^2}{2}\right) \quad \text{with} \quad \mathbf{x} \in [-7, 7] \tag{29}$$

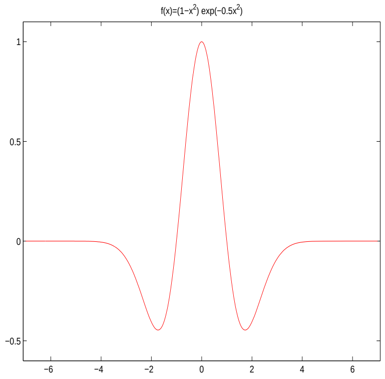

Figure 3: The Mexican hat in one dimensional

References

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

Figueiredo, M.A.T. (1954). Adaptive Sparseness for Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **25**, 1150-1159.

Neyman, J. and Scott, M. (1954). The Neyman-Scott Problem. *The Annals of Statistics*, **xx**, xxxx-xxxx.