3-15-2010

An optimal experimental design perspective on redial basis function regression

Ernest Fokoue

Prem Goel

An Optimal Experimental Design Perspective on Radial Basis Function Regression

Ernest Fokoué ERNEST.FOKOUE@GMAIL.COM
Center for Quality and Applied Statistics
Rochester Institute of Technology
Rochester, NY 14623, USA

Prem Goel GOEL@STAT.OSU.EDU
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Editor: xxxx

Abstract

This paper provides a new look at radial basis function regression that reveals striking similarities with the traditional optimal experimental design framework. We show theoretically and computationally that the so-called relevant vectors derived through the relevance vector machine (RVM) and corresponding to the centers of the radial basis function network, are very similar and often identical to the support points obtained through various optimal experimental design criteria like D-optimality. This allows us to provide a statistical meaning to the relevant centers in the context of radial basis function regression, but also opens the door to a variety of ways of approach optimal experimental design in multivariate settings.

Keywords: Radial Basis Function Regression, Relevance Vector Machine, Sensor Selection, Marginal Likelihood, Maximum A Posterior (MAP), Sparsity, D-Optimality.

1. Introduction

Radial Basis Function (RBF) regressors have been extensively used in multivariate settings for more than two decades by both numerical analysts and statisticians. The Relevance Vector Machine (RVM) introduced in (1) as a Bayesian counterpart to the popular Support Vector Machine is just one instance of the RBF machinery that has had tremendous success in the Machine Learning community thanks to its simplicity and applicability. Initially promoted on the strength of its counter-intuitive yet effective way of achieving a sparse representation in data space, RVM turned out to also provide very competitive performances in prediction, specifically outperforming the generalization abilities of Support Vector Regression. The original RVM paper (1) was entirely motivated by the search for a sparse functional representation of the prediction mechanism in the Bayesian learning framework, with an emphasis on the derivation of accurate yet fast predictions. Immediately following the publication of (1), the number of applications of the RVM approach grew steadily. Signal processing applications were some of the earliest uses of RVM with notable papers by (8), and later (22) to name a few. Environmentalists, Remote Sensing Engineers and agricultural scientists have also extensively applied RVM in various contexts as can be seen in (7), (11), (20) and (19). Interestingly, there have been many applications of RVM to Image Processing, with notable papers by (16), (21), and (24). Two areas of applications have recently seen a surge of interest in RVM, namely Text Classification with papers by (14), (15), (13) and (18), and MicroArray Data Analysis in the emerging field of genomics, with papers like (12). There has also been an increase of interest in the development of extensions of RVM and its connection to other techniques (9) and (10). Despite this relatively large number of successful applications of the Relevance Vector Machine, Not much, if anything, has been studied to provide a statistical characterization of the relevant vectors that helps to see why the Relevance Vector Machine is so successful in prediction. The goal of this paper is to provide such a statistical characterization, namely to argue that relevant vectors are indeed analogous to support (design) points in the context of D-optimality. For simplicity and clarity of exposition, the motivating examples and illustrations are univariate, thereby providing both visual and theoretical insights into our argument. Finally, the characterization provided here, although useful in its own right, also hints on what the relevant points mean in high dimensional spaces, thereby providing a way to indirectly perform predictive D-optimality via RVM in multivariate settings. The rest of this paper is organized as follows: section 2 gives a brief review of optimal experimental design for linear models, with an emphasis on the maximum a posteriori (MAP) estimation and the corresponding design problem. A connection is made between optimal experimental design and sensor selection (4). Section 3 briefly introduces the essential building blocks of the Relevance Vector Machine (RVM) with a hint on saturated designs. The connection is then made between RVM and sensor selection, with an emphasis the most similar aspects of the two methodologies. Section 4 explores two simulated examples of univariate regression and shows the striking similarities between the solutions found by the two methods. Section 5 gives some concluding remarks along with ideas for a much more complete theoretical account of the optimal design perspective of the relevance vector machine.

2. Optimal experimental design for linear models

Let $\mathbf{x}_j^\top \equiv (x_{j1}, x_{j2}, \cdots, x_{jp})$ denote a p-dimensional vector of some observable characteristics of interest. Consider a p-dimensional vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^\top$ of regression coefficients, then assume that a response (measurement) Y_j of interest at point $\mathbf{x}_j$ can be written as

$$Y_j = \mathbf{x}_j^\top \boldsymbol{\beta} + \epsilon_j, \qquad j = 1, \cdots, n.$$

Throughout this paper, we shall assume that the ϵ_j's are i.i.d $\mathbf{N}(0, \sigma^2)$. Note also that, for simplicity, we have restricted ourselves to a model that passes through the origin. Under this homoscedastic noise model, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$ of $\boldsymbol{\beta}$ is such that

$$\hat{\boldsymbol{\beta}}_{\mathsf{MLE}} = \left(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^\top\right)^{-1} \sum_{j=1}^{n} y_j \mathbf{x}_j \quad \text{and} \quad \mathsf{cov}(\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}) = \sigma^2 \left(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^\top\right)^{-1}$$

In traditional optimal experimental design, one has a set of n potential points of measurement or sensors, and the goal is to choose those k sensors or points of measurement that yield the "best" estimation of $\boldsymbol{\beta}$. For instance, with $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$ being an unbiased estimator, a reasonable criterion for measuring the goodness of $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$ will naturally be based on its covariance matrix. In fact, we will see later that all the three criteria used for measuring the optimality of the design will be based on functions of the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$. The problem in optimal experimental design is then two-fold: *(i) Which $k \ll n$ sensors or points of measurement to choose out of the n possible ones; and (ii) How many times can each chosen sensor be used, while making the total number of uses at most equal to k.* One of the most commonly used optimality criteria is the so-called D-optimality that seeks to choose those points that minimize the determinant of the covariance matrix of $\hat{\beta}_{\mathsf{MLE}}$. In order words, if each $\pi_j,\ j = 1, 2, \cdots, n$ represents the frequency of use of measurement point j, then a k-point D-optimal design is obtained as a solution to the relaxed sensor selection convex optimization problem

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\displaystyle\sum_{j=1}^{n} \pi_j \mathbf{x}_j \mathbf{x}_j^\top\right) \\ \textsf{Subject to} & 0 \le \pi_j \le 1, \qquad j = 1, \cdots, n \quad \text{and} \quad \displaystyle\sum_{j=1}^{n} \pi_j = k \end{array} \tag{1}$$

The k-point D-optimal design is therefore the subset $\xi = \{i_1, i_2, \cdots, i_k\} \subseteq \{1, 2, \cdots, n\}$ that corresponds to set of sensors or measurements with the k largest values of π_j. (4) proposes an approximate solution obtained by making the constraint $\pi_j \in (0, 1)$ implicit in the objective function so that the resulting convex optimization problem is

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\displaystyle\sum_{j=1}^{n} \pi_j \mathbf{x}_j \mathbf{x}_j^\top\right) + \kappa\left[\displaystyle\sum_{j=1}^{n} \log(\pi_j) + \sum_{j=1}^{n} \log(1 - \pi_j)\right] \\ \textsf{Subject to} & \displaystyle\sum_{j=1}^{n} \pi_j = k. \end{array} \tag{2}$$

In the Bayesian framework, if one uses a Gaussian prior $\boldsymbol{\beta} \sim \mathbf{N}(0, \boldsymbol{\Phi})$, then the corresponding Maximum A Posteriori (MAP) estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\mathsf{MAP}} = \left(\sigma^{-2}\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^\top + \boldsymbol{\Phi}^{-1}\right)^{-1} \sum_{j=1}^{n} y_j \mathbf{x}_j \quad \text{and} \quad \mathsf{cov}(\hat{\boldsymbol{\beta}}_{\mathsf{MAP}}) = \left(\sigma^{-2}\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^\top + \boldsymbol{\Phi}^{-1}\right)^{-1}$$

The corresponding approximate relaxed sensor selection problem is therefore

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\sigma^{-2}\displaystyle\sum_{j=1}^{n} \pi_j \mathbf{x}_j \mathbf{x}_j^\top + \boldsymbol{\Phi}^{-1}\right) + \kappa\left[\displaystyle\sum_{j=1}^{n} \log(\pi_j) + \sum_{j=1}^{n} \log(1 - \pi_j)\right] \\ \textsf{Subject to} & \displaystyle\sum_{j=1}^{n} \pi_j = k. \end{array} \tag{3}$$

It is interesting to note that equation (3) is a special case of the more general setting

$$\begin{aligned} &\textsf{Maximize} && \log\det\left(\sigma^{-2}\sum_{j=1}^{n}\pi_j\mathbf{x}_j\mathbf{x}_j^{\top}+\mathbf{\Phi}^{-1}\right)+a\sum_{j=1}^{n}\log(\pi_j)+b\sum_{j=1}^{n}\log(1-\pi_j) \\ &\textsf{Subject to} && \sum_{j=1}^{n}\pi_j = k, \end{aligned} \tag{4}$$

which could be thought of as a specification of an independent $\mathsf{Beta}(a+1, b+1)$ prior distribution on each π_j, namely assuming that

$$p(\pi_j|a,b) \propto \pi_j^{(a+1)-1}(1-\pi_j)^{(b+1)-1}.$$

Indeed, the choice of $\kappa = a = b$ with κ made small, is the most appropriate in this context since one wants to select a given point with the highest confidence and therefore would prefer values of π_j that are extreme, meaning either close to 1 or close to 0. The Beta with $a = b = \kappa$ with $\kappa \leq 0.5$ achieves just that, as the following Wikipedia Figure (1) of the density of a Beta distribution shows.

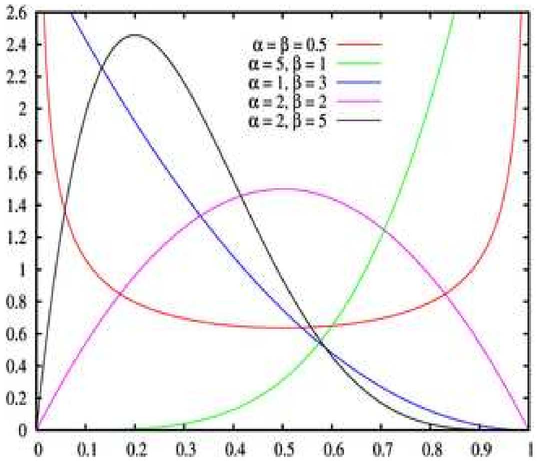

Figure 1: Density function of a Beta distribution.

Note on Figure (1) that the density of the Beta is highest at both extremes 0 and 1 when the two parameters are equal to 0.5. The objective function of equation (3) therefore has the potential of yielding a solution that does select the optimal design points. Along the same lines, one could also make the case for using any other prior that helps achieve selection. For instance, a Gamma prior with parameters that emphasize the selection of the few most important points can also be used. Note on Figure (2) that the Gamma density is highest at 0 when the number of degrees of freedoms is set to 1. This shows that the use of the Gamma distribution as in the objective function of equation (5) can be resorted to for the selection of support points, especially when one expects the number of D-optimal support points to be very small.

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\sigma^{-2}\sum_{j=1}^{n}\pi_j\mathbf{x}_j\mathbf{x}_j^\top+\boldsymbol{\Phi}^{-1}\right)+(a-1)\sum_{j=1}^{n}\log(\pi_j)-b\sum_{j=1}^{n}\pi_j \\ \textsf{Subject to} & \sum_{j=1}^{n}\pi_j=k. \end{array} \tag{5}$$

It is worth noting that the use of either the Beta or the Gamma distribution does not require the Bayesian framework, since this is not related to the distribution of $\boldsymbol{\beta}$ but instead to the indicators π_j's.

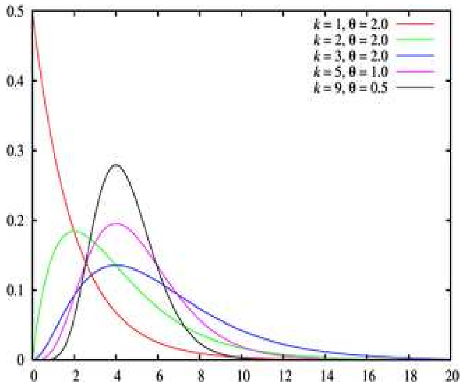

Figure 2: Density function of a Gamma distribution.

3. The Relevance Vector Machine for Regression

Given $D=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_n,y_n)\}$, the Relevance Vector Machine posits that

$$Y_j=w_0+\mathbf{h}_j^\top\mathbf{w}+\epsilon_j, \qquad j=1,\cdots,n \tag{6}$$

where $\mathbf{w}=(w_1,w_2,\cdots,w_n)^\top$ is the n-dimensional vector of weights, and the n-dimensional vector $\mathbf{h}_j^\top\equiv(K(\mathbf{x}_j,\mathbf{x}_1),K(\mathbf{x}_j,\mathbf{x}_2),\cdots,K(\mathbf{x}_j,\mathbf{x}_n))$ is built from some kernel function like

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2r^2}\right) \quad \text{or} \quad K(\mathbf{x}_i,\mathbf{x}_j)=(1+\mathbf{x}_i^\top\mathbf{x}_j)^d$$

that measures the similarity (or dissimilarity) between the vectors $\mathbf{x}_i^\top\equiv(x_{i1},x_{i2},\cdots,x_{ip})$ and $\mathbf{x}_j^\top\equiv(x_{j1},x_{j2},\cdots,x_{jp})$. Also, ϵ_j is assumed to be i.i.d $\mathbf{N}(0,\sigma^2)$. The essence of RVM however comes through the specification of the hyperprior distribution on the weights w_j. First of all, RVM assumes that $[w_j|\alpha_j]\overset{iid}{\sim}\mathbf{N}(0,\alpha_j^{-1})$, which results in a Gaussian marginal likelihood for $\mathbf{y}=(y_1,y_2,\cdots,y_n)^\top$, namely

$$p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)\propto\left[\det\left(\sigma^2\mathbf{I}+\sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right)\right]^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I}+\sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y}\right\}, \tag{7}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)$. With independent gamma priors $[\alpha_j|a,b] \overset{iid}{\sim} \mathsf{Gamma}(a,b)$, i.e.,

$$p(\alpha_j|a,b) \propto \alpha_j^{a-1}\exp(-b\alpha_j) \quad \text{and} \quad p(\boldsymbol{\alpha}|a,b) = \prod_{j=1}^{n} p(\alpha_j|a,b),$$

and assuming for simplicity that σ^2 is known, the resulting RVM objective function is

$$\begin{aligned}\mathcal{E}(\boldsymbol{\alpha}) &= -\frac{1}{2}\log\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right) - \frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y} \\ &+ (a-1)\sum_{j=1}^{n}\log\alpha_j - b\sum_{j=1}^{n}\alpha_j. \end{aligned} \tag{8}$$

From an optimization perspective, the Relevance Vector Machine problem at hand is

$$\begin{array}{ll}\mathsf{Maximize} & -\frac{1}{2}\log\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right) - \frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y} \\ & + (a-1)\sum_{j=1}^{n}\log\alpha_j + b\sum_{j=1}^{n}\alpha_j \\ \mathsf{Subject\ to} & \alpha_j > 0 \qquad j = 1, \cdots, n. \end{array} \tag{9}$$

Essentially, the variances α_j^{-1} are used by RVM as indicators of the relevance, with values closer to zero meaning irrelevance. This fact will be used to motivate an alternative to the Gamma prior. More specifically, if instead of a Gamma prior on α_j one uses a $\mathsf{Beta}(\kappa{+}1, \kappa{+}1)$ prior on α_j^{-1} , namely

$$p(\alpha_j^{-1}|\kappa) \propto [\alpha_j^{-1}]^{(\kappa+1)-1}[1-\alpha_j^{-1}]^{(\kappa+1)-1} \quad \text{and} \quad p(\boldsymbol{\alpha}|\kappa) = \prod_{j=1}^{n} p(\alpha_j^{-1}|\kappa),$$

then the problem becomes

$$\begin{array}{ll}\mathsf{Maximize} & -\frac{1}{2}\log\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right) - \frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I} + \sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y} \\ & + \kappa\left[\sum_{j=1}^{n}\log(\alpha_j^{-1}) + \sum_{j=1}^{n}\log(1-\alpha_j^{-1})\right] \\ \mathsf{Subject\ to} & 0 < \alpha_j^{-1} < 1, \qquad j = 1, \cdots, n \end{array} \tag{10}$$

Clearly, by letting $\pi_j = \alpha_j^{-1}$, equation (10) becomes

$$\begin{array}{ll}\mathsf{Maximize} & -\frac{1}{2}\log\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right) - \frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I} + \sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y} \\ & + \kappa\left[\sum_{j=1}^{n}\log(\pi_j) + \sum_{j=1}^{n}\log(1-\pi_j)\right] \\ \mathsf{Subject\ to} & 0 < \pi_j < 1, \qquad j = 1, \cdots, n, \end{array} \tag{11}$$

which in many ways is similar in form to the optimization problem of (2) derived from the traditional D-optimal design framework. In fact, if we consider performing D-optimal design on the kernel expansion "model" of equation (6), the corresponding optimization problem is

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right) + \ \kappa\left[\sum_{j=1}^{n}\log(\pi_j)+\sum_{j=1}^{n}\log(1-\pi_j)\right] \\ \textsf{Subject to} & 0<\pi_j<1, \qquad j=1,\cdots,n, \end{array} \tag{12}$$

where as indicated earlier, the use of a value for κ less than $\frac{1}{2}$, yields optimal values of π_j that are either close to 0 (irrelevance) or close to 1 (relevance). In fact, in its most generic form as presented in (4), the second portion of the objective function is absent, so that we shall use the term generic D-0ptimality criterion to refer to the problem

$$\begin{array}{ll} \textsf{Maximize} & \log\det\left(\sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right) \\ \textsf{Subject to} & \sum_{j=1}^{n}\pi_j = 1 \quad \text{and} \quad 0<\pi_j<1, \qquad j=1,\cdots,n, \end{array} \tag{13}$$

One of the most important aspects here is the following: while equations (12) and (13) provide convex optimization problems, and therefore unique solutions, equation (8) is well known not to have a unique solution (1). Let $\lambda \geq 0$ be a nonnegative real number representing the precision (inverse of variance) of a random variable.

Lemma 1 *Consider the function* $g(\lambda)=\exp\left(-\frac{1}{2}\mathbf{x}^2\lambda\right)$, *and the function*

$$f(\lambda)=\sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{1}{2}\mathbf{x}^2\lambda\right)=\sqrt{\frac{\lambda}{2\pi}}g(\lambda).$$

If $\lambda \geq 1/\mathbf{x}^2$ *then* $\sup_\lambda f(\lambda)=\sup_\lambda g(\lambda)$.

Consider once again the marginal likelihood of equation (7). As a consequence of the above lemma, to find the maximizer of

$$p(\boldsymbol{\alpha}|\mathbf{y},\sigma^2)\propto p(\boldsymbol{\alpha}|\kappa)p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2),$$

it suffices to find the maximizer of

$$q(\boldsymbol{\alpha}|\mathbf{y},\sigma^2)\propto p(\boldsymbol{\alpha}|\kappa)q(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$$

where

$$q(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)\propto\exp\left\{-\frac{1}{2}\mathbf{y}^\top\left[\sigma^2\mathbf{I}+\sum_{j=1}^{n}\alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y}\right\},$$

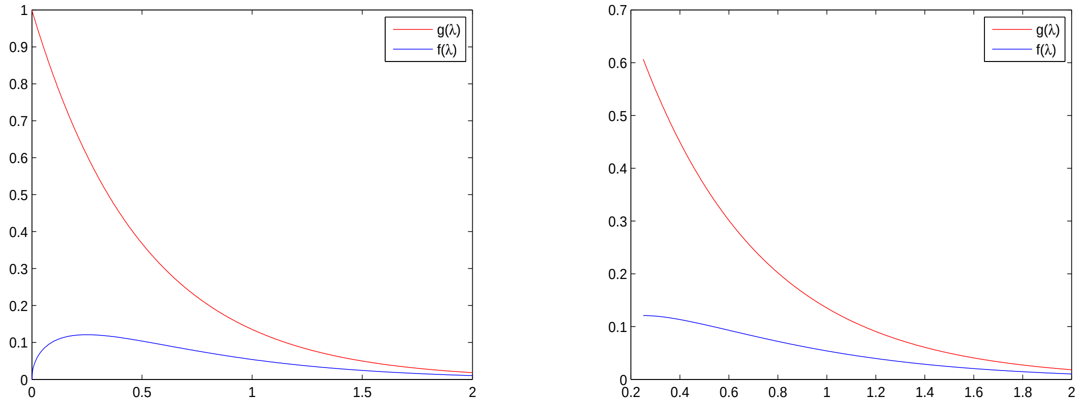

Figure 3: (left) $\sup_{\lambda} f(\lambda) \neq \sup_{\lambda} g(\lambda)$ (right) With $\lambda \geq 1/\mathbf{x}^2$, $\sup_{\lambda} f(\lambda) = \sup_{\lambda} g(\lambda)$

provided that

$$\mathbf{y}^\top \left[\sigma^2\mathbf{I} + \sum_{j=1}^{n} \alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1} \mathbf{y} = 1.$$

Clearly, $\max_{\boldsymbol{\alpha}} q(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) = 1$, which is attained if $\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n} \alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right)^{-1} = 0$. Unfortunately, this maximizer is different from the desired

$$\operatorname*{argmax}_{\boldsymbol{\alpha}} p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2).$$

As the above lemma shows however, the maximizer of $q(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ is found by minimizing $\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n} \alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right)^{-1}$ under the above mentioned constraint, so that the maximizers of $q(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ and $p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)$ coincide if one solves the problem

$$\begin{array}{ll}\underset{\boldsymbol{\alpha}}{\textsf{Maximize}} & \log\det\left(\sigma^2\mathbf{I} + \displaystyle\sum_{j=1}^{n} \alpha_j^{-1}\mathbf{h}_j\mathbf{h}_j^\top\right) \\ \textsf{Subject to} & \mathbf{y}^\top \left[\sigma^2\mathbf{I} + \displaystyle\sum_{j=1}^{n} \pi_j\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1} \mathbf{y} = 1, \quad \text{and} \quad 0 < \pi_j < 1, \qquad j = 1,\cdots,n.\end{array} \tag{14}$$

Reverting to $\pi_j = \alpha_j^{-1}$ and using the $\mathsf{Beta}(\kappa+1, \kappa+1)$ distribution to induce selection, we now have the convex optimization problem,

$$\begin{array}{lll}\underset{\boldsymbol{\alpha}}{\mathsf{Maximize}} & \log\det\left(\sigma^2\mathbf{I} + \sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right) + \kappa\left[\sum_{j=1}^{n}\log(\pi_j) + \sum_{j=1}^{n}\log(1-\pi_j)\right] & \\ \mathsf{Subject\ to} & \mathbf{y}^\top\left[\sigma^2\mathbf{I} + \sum_{j=1}^{n}\pi_j\mathbf{h}_j\mathbf{h}_j^\top\right]^{-1}\mathbf{y} = 1, \quad \text{and} \quad 0 < \pi_j < 1, & j = 1, \cdots, n.\end{array} \tag{15}$$

Interestingly, the above reasoning is very similar to the underlying principle of D-optimality, with the only bit being the constraint linked to the response vector $\mathbf{y}$. This last bit should not surprise, since the motivating framework of the Relevance Vector Machine is Bayesian implying that inference is based on posterior quantities which must be conditional on having observed the data. Worth noting also is the fact that the convex optimization problem of equation (4) regularized the weighted information matrix with $\sigma^2\mathbf{I}$ using the noise variance. This provides a device that might help ward off some potential numerical instability due to some kernels. Finally, like the typical D-optimality criterion, our derived estimation scheme does not require any explicit manipulation of the dimensionality of the input vectors $\mathbf{x}_j$'s. The convex optimization problem of equation (12) can be thought of the non-Bayesian counterpart of the more Bayesian approach of equation (4). This establishes that the kernel expansion does indeed does provide a framework for dealing with optimal experimental design for arbitrary models provided that the model can be expressed via a kernel. This is particularly interesting because the kernel regression setting handles both linear and nonlinear problems without any added modelling work.

4. Numerical demonstrations and simulations

Example 1: In order to gain insights into the similarities and the differences between D-optimal support points and relevant vectors, we first consider a simple univariate function

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

With this, our data consists of pairs $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, where the $\mathbf{x}_i$'s are equally spaced points in $[-1, +1]$. From a traditional D-optimal design standpoint, we need to specify a model in order to form the data matrix. A natural candidate in this case is the polynomial regression model. A quick snoop at the scatterplot suggests that an 8th polynomial could capture the underlying function, i.e.,

$$Y_j = \beta_0 + \beta_1\mathbf{x}_j + \beta_2\mathbf{x}_j^2 + \cdots + \beta_8\mathbf{x}_j^8 + \epsilon_j.$$

For the relevance vector machine, we used the gaussian radial basis function kernel, and found the bandwidth of $r = 0.5$ to be adequate for this data.
For simplicity, the noise variance σ^2 is assumed known and fixed at 0.2^2. As far as the similarities go, most of the points are identical for both methods. Regarding the differences, the relevance vector machine yields fewer points, for the obvious reason that it applies an

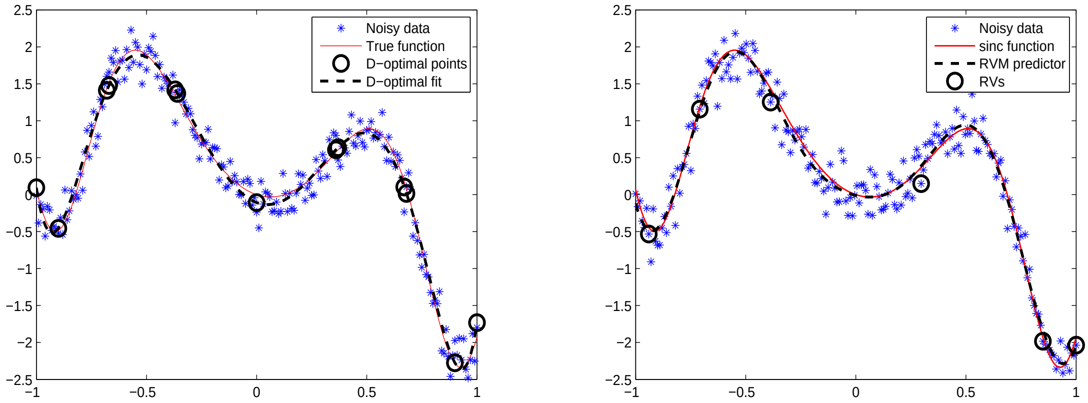

Figure 4: (left) D-optimal support points; (right) RVM relevant vectors.

extra constraint driven by the response values and therefore achieves more tuning. Besides, it is important to recall that the strongest motivation behind RVM is sparsity (fewer relevant vectors), while D-optimality sets out to find a k-point design. The number is fixed in one case, while the minimum number is sought in the other.

Example 2: As our second example, we take a look at the commonly used sinc function

$$f(\mathbf{x}) = \frac{\sin 10\mathbf{x}}{10\mathbf{x}} \qquad \text{with} \quad \mathbf{x} \in [-1, +1].$$

For this example, our noise variance is still $\sigma^2 = 0.2^2$, but our response variable is now expressed as a weighted sum of *Legendre* or *Chebyshev* orthogonal polynomials to which we add the homoscedastic gaussian noise $\epsilon \sim \boldsymbol{N}(0, \sigma^2)$ as before. Figure (5) shows the results obtained from both the D-optimality criterion (left) and the Relevance vector machine approach (right). Again, while it is obvious that the two methods are looking for the points that most affect the variance of the estimates of the parameters, it seems clear that RVM retains fewer points than D-optimality. The reason is that the results presented here are obtained using the generic D-optimality criterion of equation (13). We solved this using CVX, a package for specifying and solving convex programs (5; 6). Once the D-optimality criterion is enriched with the selection inducing Beta as in equation (12), a more sparse solution should be expected. Also, the complete reformulation of equation should produce results that are fairly identical to the output from the Relevance Vector Machine of (1).

5. Conclusion, discussion and future work

We have shown in this paper that the statistical problem underlying the now very popular Relevance Vector machine can essentially formulated as an adaptive D-optimal design problem. The formulation derived in this paper provides a crucial advantage in that the problem is now a convex optimization task with the guarantee of a unique solution, as opposed to original RVM that is known not to yield a unique solution. Our immediate future work is to numerically implement the new formulation and also use our derived scheme on real

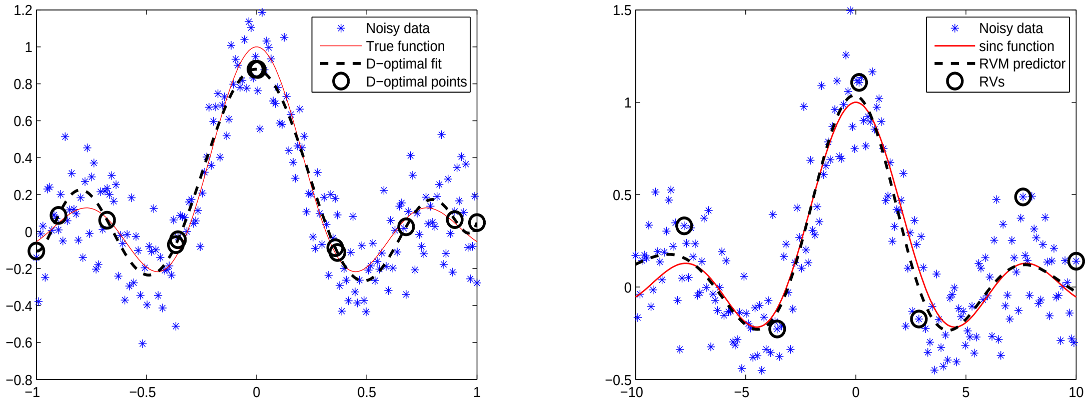

Figure 5: (left) D-optimal support points; (right) RVM relevant vectors.

life problems. Another aspect worth exploring is the reconstruction of the primal problem corresponding to the dual definition of the RVM. Much later, we hope to investigate the theoretical aspects of this connection a little further, and also consider exploring how this affects Relevance Vector Classification.

References

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

[2] Figueiredo, M.A.T. (1994). Adaptive Sparseness for Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **25**,1150-1159.

[3] Boyd, S. and L. Vandenberghe (2004). Convex Optimization. *Cambridge University Press*, 2004.

[4] Joshi, S. and Boyd, S (2008). Sensor Selection via Convex Optimization. *IEEE Transactions on Signal Processing*, To appear.

[5] Grant, M. and Boyd, S (2008) CVX: Matlab software for disciplined convex programming. `http://stanford.edu/~boyd/cvx`, December 2008.

[6] Grant, M. and Boyd, S. (2008) Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar), V. Blondel, S. Boyd, and H. Kimura, editors, pages 95-110, Lecture Notes in Control and Information Sciences, Springer, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[7] Yuan, J et al. (2007). Integrating relevance vector machines and genetic algorithms for optimization of seed-separating process. *Engineering Applications of Artificial Intelligence*, **20**, pp. 970-979.

[8] Chen, S et al. (2001). The Relevance Vector Machine Technique for Channel Equalization Application. *IEEE Transactions on Neural Networks*, Vol **20**, No. 6, pp. 1529-1532.

[9] D'Souza, A et al. (2004). The Bayesian Backfitting Relevance Vector Machine. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[10] Tripathi, S and Govindaraju, R. S. (2007). On Selection of kernel parameters in relevance vector machines for hydrologic applications. *Stoch Environ Res Risk Assess*, **21**, pp. 747-764.

[11] Wipf, D and Nagarajan, Srikantan (2007). Beamforming using the Relevance Vector Machine. *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, (USA), 2007.

[12] Yinhai, L et al. (2006). Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine. *Genome Research*, **online**, 2006.

[13] Silva, C and Ribeiro, B. (2008). Towards Expanding Relevance Vector Machines to Large Scale Datasets. *International Journal of Neural Systems*, Vol **18**, No 1, pp. 45-58, World Scientific Publishig Company, February 2008.

[14] Silva, C and Ribeiro, B. (2007). Combining Active Learning and Relevance Vector Machines for Text Classification. *Proceedings of the IEEE International Conference on Machine Learning Applications*, pp. 130-135.

[15] Silva, C and Ribeiro, B. (2007). RVM Ensemble for Text Classification. *International Journal of Computational Intelligence Research*, Vol **3**, No 1, pp. 31-35, January 2007.

[16] Thayananthan, A et al. (2008). Multivariate Relevance Vector Machines for Tracking. *Technical Report.* Cambridge University.

[17] Yeung, P. F. et al. (2005). Relevance Vector Machine for content-based retrieval of 3D head models. *Proceedings of the Ninth International Conference on Information Visualization*, 6-8 July 2005, pp. 425-429.

[18] Dasgupta, N et al. (2007). Relevance Vector Machine Quantization and Density Function Estimation: Application to HMM-Based Multi-Aspect Text Classification *Technical Report.* Duke University.

[19] Camps-Valls, G. et al. (2006). Retrieval of Oceanic Chlorophyll Concentration with Relevance Vector Machines. *Remote Sensing of Environment*, Vol **105**, No. 1, pp. 23-33.

[20] Camps-Valls, G. et al. (2005). Relevance Vector Machines for Sparse Learning of Biophysical Parameters. *Proceedings of SPIE, the International Society of Optical*

Engineering. Image and Signal Processing for Remote Sensing, Vol **5982**, pp. 59820Z.1-59820Z.12.

[21] Wei, L et al. (2005). A Relevance Vector Machine Technique for the automatic detecttion of clustered microcalcifications. *Medical Imaging 2005: Image Processing. Edited by Fitzpatrick, J. Michael; Reinhardt, Joseph, Proceedings of the SPIE*, Vol **5747**, pp. 831-839.

[22] Weiss, R. J. et Ellis, D. P. W.(2005). Estimating Single Channel Source Separation Masks: Relevance Vector Machine Classifiers vs Pitch-Based Masking. *Technical Report*, Dept. of Elec. Eng, Columbia University, New York, NY 10027, USA.

[23] Masataro, O. and Hiroyuki, M.(2005). Short-term load Forecasting with Relevance Vector Machine. *Papers of Technical Meeting on Power Engineering, IEE Japan*, Vol **PE-05**, No. 20-23.25-27, pp. 1-6.

[24] Wong, W. S. et al. (2005). Using a Sparse Learning Relevance Vector Machine in Facial Expression Recognition. *Technical Report*, Man-Machine Interaction Group, Delft University of Technology, The Netherlands, eMail:L.J.M.Rothkrantz@ewi.tudelft.nl