1-3-2011

Efficient techniques for simultaneous variable selection and sensor selection via convex selection inducing penalties

Ernest Fokoue

Efficient Techniques for Simultaneous Variable Selection and Sensor Selection via Convex Selection Inducing Penalties

Ernest Fokoué
Center for Quality and Applied Statistics
Rochester Institute of Technology
Rochester, NY 14624, USA
eMail: epfeqa@rit.edu

January 3, 2011

Abstract

This paper extends results from the traditional D-optimality machinery to derive an efficient technique for simultaneous variable selection and sensor selection. An important advantage of the proposed technique is the convexity of the formulated optimization task along with a byproduct of straightforward sparsity. The theoretical foundation of the proposed method is explored at great length, and a variety of examples are provided to demonstrated the effectiveness of our technique. Comparisons with existing techniques are offered that provide evidence as the superiority of our technique on a variety of indicators.

Keywords: *Variable Selection, Convex optimization, Bayesian analysis, Optimal Experimental Design, Sensor selection, Sparsity, D-Optimality.*

1 Introduction

Let $\mathbf{x}_j^\top \equiv (\mathrm{x}_{j1}, \mathrm{x}_{j2}, \cdots, \mathrm{x}_{jp})$ 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 $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$. 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

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

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. (Joshi and Boyd 2009) 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

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

Let $\mathrm{w}_k \in (0, 1)$ denote the importance (relevance) of variable x_k. Consider a diagonal matrix

$$\boldsymbol{W} = \begin{bmatrix} \mathrm{w}_1 & & & \\ & \mathrm{w}_2 & & \\ & & \ddots & \\ & & & \mathrm{w}_p \end{bmatrix} = \mathsf{diag}(\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_p)$$

Instead of using the input vector $(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_p)$, consider using $\tilde{\mathbf{x}}_j = \boldsymbol{W}\mathbf{x}_j$, so that the response Y_j is now $Y_j = \mathbf{x}_j^\top \boldsymbol{W}^\top \beta + \epsilon_j, \quad j = 1, \cdots, n$. Now, let

$$\mathsf{Pen}(\boldsymbol{W}) = \left[\sum_{k=1}^{p}\log(\mathrm{w}_k) + \sum_{k=1}^{p}\log(1-\mathrm{w}_k)\right] \quad \text{and} \quad \boldsymbol{X}^\top\boldsymbol{X} = \left[\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j^\top\right]$$

Variable selection can be achieved by solving the problem formulated in equation (4).

$$\mathcal{E}(\boldsymbol{W}) = \log\det\left(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}\right) + \nu\mathsf{Pen}(\boldsymbol{W}) \tag{3}$$

The corresponding optimization problem is

$$\begin{bmatrix} \mathsf{Maximize} & \mathcal{E}(\boldsymbol{W}) \\ \mathsf{Subject\ to} & 0 \leq \mathrm{w}_k \leq 1, \qquad k = 1, \cdots, p \\ & \displaystyle\sum_{k=1}^{p}\mathrm{w}_k = \omega. \end{bmatrix} \tag{4}$$

where ω is a fraction of p indicating the extend of parsimony desired. It is worth noting that that second part of the constraint need not be specified in practice during the optimization procedure. Typically,

one may want to use w_k's that are indicator variables, so as to perform straightforward selection. However, it turns out that using real numbers between 0 and 1 has many advantages as discussed by (Joshi and Boyd 2009) and (Fokoue and Goel 2009). For instance, values of ν less than 0.5 force the resulting w_k's to be close to 0 or 1, thereby providing a strong determination of the importance (relevance) of the corresponding variable. Another important advantage is both computational and numerical: to see, let's assume that $p = 2$ and $n = 2$. Then

$$g(\boldsymbol{W}) = \boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{W} = \begin{bmatrix} \mathrm{w}_1^2 \sum_{i=1}^{3} \mathrm{x}_{1i}\mathrm{x}_{i1} & \mathrm{w}_1\mathrm{w}_2 \sum_{i=1}^{3} \mathrm{x}_{i1}\mathrm{x}_{i2} \\ \mathrm{w}_1\mathrm{w}_2 \sum_{i=1}^{3} \mathrm{x}_{i2}\mathrm{x}_{i1} & \mathrm{w}_2^2 \sum_{i=1}^{3} \mathrm{x}_{i2}\mathrm{x}_{2i} \end{bmatrix}$$

Throughout the optimization procedure, it is crucial that this matrix remain well conditioned. Now, if the w_k's are binary indicators, then the matrix cannot stay full rank. For instance, if $\mathrm{w}_2 = 0$, then the matrix collapses to a scalar and the whole computation cannot continue. However, with $\mathrm{w}_k \in (0, 1)$, the matrix remains well conditioned throughout. This advantage is crucial to obtaining the desired variable selection solution, but also makes the procedure doable throughout traditional methods like Newton's method.

Theorem 1 *If $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{p\times p}$ are two symmetric matrices, then*

$$\frac{\partial \log\det A^\top BA}{\partial A} = \frac{\partial \log\det A^\top BA}{\partial A^\top BA}\,\frac{\partial A^\top BA}{\partial A} = (A^\top BA)^{-\top}\frac{\partial A^\top BA}{\partial A} = 2(A^\top BA)^{-\top}BA.$$

Therefore,

$$\frac{\partial g(\boldsymbol{W})}{\partial \boldsymbol{W}} = 2(\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{W})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{W}$$

Let $\mathbf{D} = \mathbf{X}^\top\mathbf{X}$. Then, we need to derive

$$\begin{aligned} g''(\mathbf{W}) &= \frac{\partial(\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}\mathbf{D}\boldsymbol{W}}{\partial \boldsymbol{W}} \\ &= \frac{\partial(\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}}{\partial \boldsymbol{W}}\mathbf{D}\boldsymbol{W} + (\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}\frac{\partial \mathbf{D}\boldsymbol{W}}{\partial \boldsymbol{W}} \\ &= \frac{\partial(\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}}{\partial \boldsymbol{W}^\top\mathbf{D}\boldsymbol{W}}\frac{\partial \boldsymbol{W}^\top\mathbf{D}\boldsymbol{W}}{\partial \boldsymbol{W}}\mathbf{D}\boldsymbol{W} + (\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}\mathbf{D} \\ &= -(\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-\top} \otimes (\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}2\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{W} + (\boldsymbol{W}^\top\mathbf{D}\boldsymbol{W})^{-1}\mathbf{D} \end{aligned}$$

Perhaps the most importance advantage of our proposed scheme lies in the fact that we have a convex optimization problem, with the guarantee of a unique solution.

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \mathcal{E}(\boldsymbol{W})$$

This is even a better advantage because once could even think of simultaneously achieving variable selection and the corresponding D-optimal design.

$$\mathcal{E}(\boldsymbol{W}, \boldsymbol{Z}) = \log\det\left(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}\boldsymbol{W}\right) + \nu\mathsf{Pen}(\boldsymbol{W}) + \kappa\mathsf{Pen}(\boldsymbol{Z}) \tag{5}$$

where

$$\mathsf{Pen}(\boldsymbol{Z}) = \left[\sum_{j=1}^{n}\log(\mathrm{z}_j) + \sum_{j=1}^{n}\log(1-\mathrm{z}_j)\right] \tag{6}$$

with $\boldsymbol{Z} = \mathsf{diag}(\mathrm{z}_1, \mathrm{z}_2, \cdots, \mathrm{z}_n)$ is the diagonal matrix whose entries act on the sensors. The convex optimization problem to solve in order to achieve that is

$$\left[\begin{array}{lll} \mathsf{Maximize} & \mathcal{E}(\boldsymbol{W}, \boldsymbol{Z}) & \\ \mathsf{Subject\ to} & 0 \leq \mathrm{w}_k \leq 1, \qquad k = 1, \cdots, p \quad \text{and} & \displaystyle\sum_{k=1}^{p} \mathrm{w}_k = \omega. \\ & 0 \leq \mathrm{z}_j \leq 1, \qquad j = 1, \cdots, n \quad \text{and} & \displaystyle\sum_{j=1}^{n} \mathrm{z}_j = \zeta. \end{array}\right. \tag{7}$$

The solution

$$(\hat{\boldsymbol{Z}}) = \underset{\boldsymbol{W}, \boldsymbol{Z}}{\mathsf{arg\,max}}\ \mathcal{E}(\boldsymbol{W}, \boldsymbol{Z})$$

provides a total $n + p$ estimates, all in $(0, 1)$, with the $\hat{\mathrm{w}}_k$'s allowing variable selection while the $\hat{\mathrm{z}}_j$'s allow sensor selection. It is worth emphasizing once again that we have the great advantage of convex optimization.

Theorem 2 *If $A \in \mathbb{R}^{p \times p}$ and $B \in \mathbb{R}^{p \times p}$ are two symmetric matrices, then*

$$\frac{\partial \log\det A^\top B A}{\partial B} = \frac{\partial \log\det A^\top B A}{\partial A^\top B A}\,\frac{\partial A^\top B A}{\partial B} = (A^\top B A)^{-\top}\frac{\partial A^\top B A}{\partial B} = A(A^\top B A)^{-\top} A^\top.$$

From this proposition, the derivative of $h(\boldsymbol{Z}) = \log\det\left(\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{X} \boldsymbol{W}\right)$ is given by

$$\frac{\partial h(\boldsymbol{Z})}{\partial \boldsymbol{Z}} = \boldsymbol{X}\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{X} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{X}^\top.$$

Obviously, the case $\boldsymbol{W} = \boldsymbol{I}_p$ corresponds to the traditional D-optimality.

1.1 Traditional approximate D-optimality

Let $\boldsymbol{Z} = \mathsf{diag}(\mathrm{z}_1, \mathrm{z}_2, \cdots, \mathrm{z}_n)$ be the diagonal matrix whose entries act on the sensors.

$$\mathcal{E}(\boldsymbol{Z}) = \quad \log\det\left(\tau \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{X} + \mathbf{Q}^{-1}\right) + \kappa \left[\sum_{j=1}^{n} \log(\mathrm{z}_j) + \sum_{j=1}^{n} \log(1 - \mathrm{z}_j)\right] \tag{8}$$

The convex optimization problem to solve in order to achieve that is

$$\left[\begin{array}{lll} \mathsf{Maximize} & \mathcal{E}(\boldsymbol{Z}) & \\ \mathsf{Subject\ to} & 0 \leq \mathrm{w}_k \leq 1, \qquad k = 1, \cdots, p \quad \text{and} & \displaystyle\sum_{k=1}^{p} \mathrm{w}_k = \omega. \\ & 0 \leq \mathrm{z}_j \leq 1, \qquad j = 1, \cdots, n \quad \text{and} & \displaystyle\sum_{j=1}^{n} \mathrm{z}_j = \zeta. \end{array}\right. \tag{9}$$

The solution

$$(\hat{\boldsymbol{W}}, \hat{\boldsymbol{Z}}) = \underset{\boldsymbol{W}, \boldsymbol{Z}}{\mathsf{arg\,max}}\ \mathcal{E}(\boldsymbol{W}, \boldsymbol{Z})$$

provides a total $n + p$ estimates, all in $(0, 1)$, with the $\hat{\mathrm{w}}_k$'s allowing variable selection while the $\hat{\mathrm{z}}_j$'s allow sensor selection. It is worth emphasizing once again that we have the great advantage of convex optimization.

Theorem 3 *If $A \in I\!R^{p\times p}$ and $B \in I\!R^{p\times p}$ are two symmetric matrices, then*

$$\frac{\partial \log\det A^\top BA}{\partial B} = \frac{\partial \log\det A^\top BA}{\partial A^\top BA}\,\frac{\partial A^\top BA}{\partial B} = (A^\top BA)^{-\top}\frac{\partial A^\top BA}{\partial B} = A(A^\top BA)^{-\top}A^\top.$$

From this proposition, the derivative of $h(\boldsymbol{Z}) = \log\det\left(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}\boldsymbol{W}\right)$ is given by

$$\frac{\partial h(\boldsymbol{Z})}{\partial \boldsymbol{Z}} = \boldsymbol{X}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{X}^\top.$$

Obviously, the case $\boldsymbol{W} = \boldsymbol{I}_p$ corresponds to the traditional D-optimality.
Let $\mathbf{Z} = \mathsf{diag}(\mathrm{z}_1, \mathrm{z}_2, \cdots, \mathrm{z}_n)$ denote the diagonal matrix of sensor pointers (weights).

$$f(\mathbf{Z}) = -\log(\det(\tau\mathbf{X}^\top\mathbf{Z}\mathbf{X} + \mathbf{Q}^{-1})) - \kappa\sum_{j=1}^{n}\{\log\mathrm{z}_j + \log(1-\mathrm{z}_j)\}$$

$$\mathbf{W} = \left[\mathbf{X}^\top\mathbf{Z}\mathbf{X}\right]^{-1} \quad \text{and} \quad \mathbf{V} = \mathbf{X}\mathbf{W}\mathbf{X}^\top$$

Clearly,

$$\begin{aligned}\frac{\partial \log(\det(\tau\mathbf{X}^\top\mathbf{Z}\mathbf{X} + \mathbf{Q}^{-1}))}{\partial \mathbf{Z}} &= \frac{\partial \log(\det(\tau\mathbf{X}^\top\mathbf{Z}\mathbf{X} + \mathbf{Q}^{-1}))}{\partial \mathbf{X}^\top\mathbf{Z}\mathbf{X}}\frac{\partial \mathbf{X}^\top\mathbf{Z}\mathbf{X}}{\partial \mathbf{Z}}\\ &= -\mathsf{vec}((\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-\top})^\top\mathbf{X}^\top\otimes\mathbf{X}^\top\\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\end{aligned}$$

As a result, the Jacobian would then be

$$\nabla f = -\mathsf{diag}(\mathbf{V}) - \kappa\left(\frac{1}{\mathrm{z}_j} - \frac{1}{1-\mathrm{z}_j}\right)$$

Now,

$$\begin{aligned}\frac{\partial \mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top}{\partial \mathbf{Z}} &= \frac{\partial \mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top}{\partial (\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}}\frac{\partial (\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}}{\partial \mathbf{Z}}\\ &= \mathbf{X}\otimes\mathbf{X}\frac{\partial (\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}}{\partial \mathbf{X}^\top\mathbf{Z}\mathbf{X}}\frac{\partial \mathbf{X}^\top\mathbf{Z}\mathbf{X}}{\partial \mathbf{Z}}\\ &= \mathbf{X}\otimes\mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-\top}\otimes(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\otimes\mathbf{X}^\top\\ &= [\mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top]\otimes[\mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top]\end{aligned}$$

Therefore, the Hessian in this case is given by

$$\mathbf{H} = \nabla\nabla f = \mathbf{V}\otimes\mathbf{V} + \kappa\ \mathsf{diag}\left(\frac{1}{\mathrm{z}_j^2} + \frac{1}{(1-\mathrm{z}_j)^2}\right)$$

2 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.

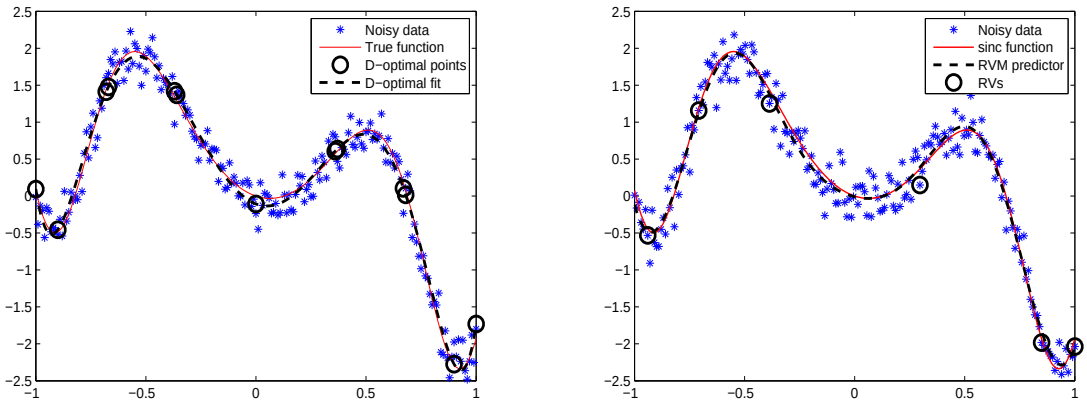

Figure 1: (left) D-optimal support points; (right) RVM relevant vectors.

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 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 (2) 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 (**??**). We solved this using `CVX`, a package for specifying and solving convex programs (Joshi and Boyd 2009). Once the D-optimality criterion is enriched with the selection inducing Beta as in equation (**??**), 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 (Tipping 2001).

3 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

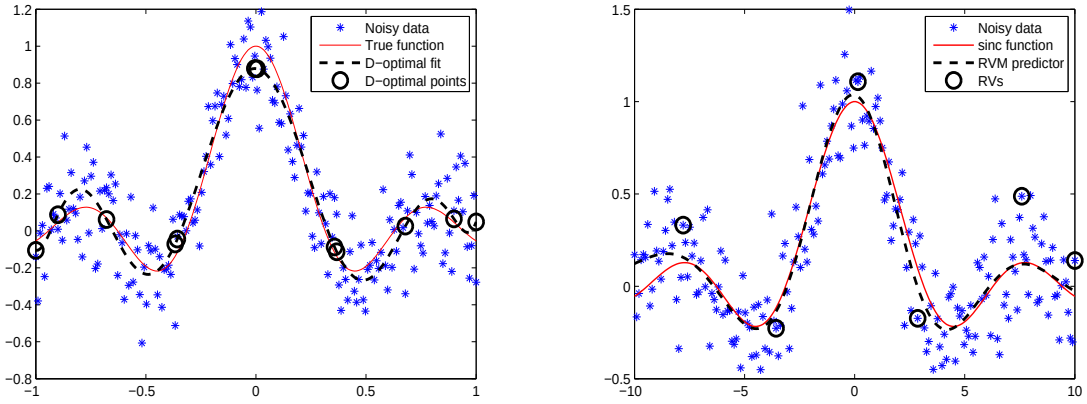

Figure 2: (left) D-optimal support points; (right) RVM relevant vectors.

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 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

Fokoue, E. and P. Goel (2009, December). An Optimal Experimental Design Perspective on the Relevance Vector Machine. Technical Report EPF-09-7-1, Rochester Institute of Technology, Rochester, New York, Center for Quality and Applied Statistics, 98 Lomb Memorial Drive, Rochester, NY 14623.

Joshi, S. and S. Boyd (2009). Sensor selection via convex optimization. *IEEE Trans. Signal Processing 57*(2), 451–462.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *J. Mach. Learning Res. 1*, 211–244.