Publication V

Timo Similä and Jarkko Tikka (2006). Common subset selection of inputs in multiresponse regression, *Proceedings of the IEEE International Joint Conference on Neural Networks - IJCNN 2006*, pp. 1908–1915.

V

2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Common Subset Selection of Inputs in Multiresponse Regression

Timo Similä and Jarkko Tikka

Abstract— **We propose the Multiresponse Sparse Regression algorithm, an input selection method for the purpose of estimating several response variables. It is a forward selection procedure for linearly parameterized models, which updates with carefully chosen step lengths. The step length rule extends the correlation criterion of the Least Angle Regression algorithm for many responses. We present a general concept and explicit formulas for three different variants of the algorithm. Based on experiments with simulated data, the proposed method competes favorably with other methods when many correlated inputs are available for model construction. We also study the performance with several real data sets.**

I. Introduction

Many practical regression tasks have several input variables available for model construction. However, the data analyst may not know how the inputs are related to the response variables. Our focus is on linear models (linear basis expansions) that have many response variables with the single response problem as a special case. A small number of observations compared to the number of inputs causes the problem of *overfitting*: the model fits well on training data but poorly on all other ones [1]. Highly correlated inputs cause the problem of *collinearity*: model interpretation is misleading as variation in an input can be compensated by variation in another one [2]. We offer techniques to help the data analyst to solve these problems.

A popular approach is to build a separate model for each response variable when various procedures are available, see for instance [3]. The three main techniques are pure input selection [1], regularization or shrinking [4], and subspace methods [5]. Shrinking means constraining the regression coefficients such that the unimportant inputs tend to have small coefficient values. In the subspace approach the data are projected onto a smaller subspace in which the model is fitted. Input selection differs from the two other techniques as some of the inputs are completely left out of the model.

Combinations of shrinkage and input selection have attracted attention recently [6], [7]. These methods enjoy practical benefits of input selection including aid in model interpretation, economic efficiency if measured inputs have costs, and computational efficiency due to simplicity of the model. Input selection alone may not solve the problem of overfitting if an unconstrained model is used. Shrinkage helps in this and also makes the solution easier to obtain. Stepwise subset selection without shrinkage may fail to recognize important combinations of inputs, especially when collinearities are present [2]. Subset selection is a hard combinatorial problem in general. On the contrary, subset selection coupled with shrinkage can be formulated as a convex optimization problem [6], or the whole solution path may be computed as a function of the shrinking parameter by an efficient forward selection algorithm with varying step lengths [7], [8], [9].

Regression models using the same set of inputs to estimate several responses are popular for instance in chemistry [10]. Multiresponse models have advantages over separate models when the responses are correlated [11], [12]. Common subset selection of inputs can be motivated by the computational burden of repeating input selection for each response variable separately [13]. Moreover, sometimes the individual responses are not relevant and the aim is to assess the importance of inputs in the estimation of all the responses together. Commonly used criteria for input selection are tests of statistical significance [14], information criteria [15], and prediction error [13]. However, they only rank combinations of inputs, and some stepwise method is usually applied to find promising combinations. In [16], a stochastic Bayesian input selection method is proposed, but it involves several tuning parameters and requires a strategy for monitoring convergence to a stationary distribution. A more explicit approach is Simultaneous Variable Selection (SVS) [17]. It does shrinking and input selection, and it is formulated as a convex optimization problem. A drawback of SVS is that it is more like an exploratory tool as it is suggested in [17] to use the method only to select inputs. A separate model is constructed using the selected subset.

We proposed the Multiresponse Sparse Regression (MRSR) algorithm in [18] for input selection and shrinkage. It equals to the Least Angle Regression (LARS) algorithm [8] when a single response is estimated, but unlike LARS, it is applicable with several responses as well. MRSR uses the same inputs in the estimation of all the responses like SVS does, but it differs from SVS in the sense that the model is useful from the start. The formulation of MRSR using a 1-norm correlation criterion, as presented in [18], scales badly with the number of responses. In this article, we introduce variants of MRSR for the 1-norm, 2-norm, and ∞-norm criteria, which are all fast to compute even when both input and response data are high-dimensional.

The rest of the article is organized as follows. In Section II, we present the general MRSR algorithm and give formulas for the step length in three special cases corresponding to different choices of the norm. In Section III, we consider related methods, which we compare with MRSR later in Section IV. The comparisons are structured as follows. Firstly,

Timo Similä is with the Laboratory of Computer and Information Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland (phone: +358 9 451 3360; fax: +358 9 451 3277; email: timo.simila@hut.fi)

Jarkko Tikka is with the Laboratory of Computer and Information Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland (email: tikka@cis.hut.fi)

0-7803-9490-9/06/$20.00/©2006 IEEE

simulation experiments are carried out to explore the effect of collinearity among the input variables. Secondly, three real world data sets are analyzed. Thirdly, image reconstruction is presented as an example of a high-dimensional problem. Section V concludes the article.

II. Multiresponse Sparse Regression Algorithm

Suppose that we have n observations of both q response variables and m input variables. To fix notation, denote the response data by an $n \times q$ matrix $\boldsymbol{T} = [\boldsymbol{t}_1 \cdots \boldsymbol{t}_q]$ and the input data by an $n \times m$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_m]$. We assume that all variables have zero mean and their scales are comparable. A linear model

$$\boldsymbol{Y}_k = \boldsymbol{X}\boldsymbol{W}_k \tag{1}$$

is used in the estimation of the responses $\boldsymbol{T}$, where the $m \times q$ matrix $\boldsymbol{W}_k$ denotes the regression coefficients. The MRSR algorithm adds inputs sequentially to the model. In the beginning, all entries of $\boldsymbol{W}_0$ are zero. Each step of the algorithm $k = 0, 1, \ldots$ introduces a new nonzero row to $\boldsymbol{W}_k$ and thus a new active input. If we proceed to the end of the algorithm with the step $k = m-1$, we finally have all m inputs in the model and $\boldsymbol{Y}_k$ equals to the ordinary least squares (OLS) estimate for $\boldsymbol{T}$.

Denote the correlation between the residuals and the jth input in the beginning of the step k by

$$c_{k,j} = \|(\boldsymbol{T} - \boldsymbol{Y}_k)^T \boldsymbol{x}_j\|_p, \tag{2}$$

where $p \geq 1$ fixes a norm[1]. A high value of $c_{k,j}$ suggests including $\boldsymbol{x}_j$ in the model, since it correlates with the part of $\boldsymbol{T}$, which is currently unexplained. A variational interpretation for $c_{k,j}$ can be given in terms of the gradient of the error sum of squares with respect to the regression coefficients associated with the jth input[2]

$$\boldsymbol{g}_j(\boldsymbol{W}) = \tfrac{\partial}{\partial \boldsymbol{w}_{(j)}} \tfrac{1}{2} \|\boldsymbol{T} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2,$$

where the jth row of $\boldsymbol{W}$ is $\boldsymbol{w}_{(j)}^T$. Now the criterion (2) measures sensitivity of the error to changes in $\boldsymbol{w}_{(j)}$ in the p-norm sense, $c_{k,j} = \|\boldsymbol{g}_j(\boldsymbol{W}_k)\|_p$. The higher the value of $c_{k,j}$ is, the more efficient $\boldsymbol{x}_j$ is in the reduction of the current error.

Let the maximum correlation be denoted by $\hat{c}_k$ and the group of inputs that satisfy the maximum by $\mathcal{A}_{k+1}$, i.e.

$$\hat{c}_k = \max_{1 \leq j \leq m} c_{k,j}, \quad \mathcal{A}_{k+1} = \{j : c_{k,j} = \hat{c}_k\}. \tag{3}$$

Collect the inputs that belong to the active set $\mathcal{A}_{k+1}$ as an $n \times |\mathcal{A}_{k+1}|$ matrix $\boldsymbol{X}_{k+1} = [\cdots \boldsymbol{x}_j \cdots]_{j \in \mathcal{A}_{k+1}}$. By using only the active inputs, the least squares estimate $\widehat{\boldsymbol{Y}}_{k+1}$ for the responses and the least squares estimate $\widehat{\boldsymbol{W}}_{k+1}$ for the regression coefficients can be computed

$$\widehat{\boldsymbol{Y}}_{k+1} = \boldsymbol{X}_{k+1}\widehat{\boldsymbol{W}}_{k+1} \tag{4}$$

$$\begin{aligned} \widehat{\boldsymbol{W}}_{k+1} &= \operatorname{argmin} \tfrac{1}{2}\|\boldsymbol{T} - \boldsymbol{X}_{k+1}\boldsymbol{W}\|_{\mathrm{F}}^2 \\ &= (\boldsymbol{X}_{k+1}^T \boldsymbol{X}_{k+1})^{-1} \boldsymbol{X}_{k+1}^T \boldsymbol{T}. \end{aligned} \tag{5}$$

[1]The p-norm of a vector $\boldsymbol{x}$ is $\|\boldsymbol{x}\|_p = (\sum_j |x_j|^p)^{\frac{1}{p}}$ and in the limit, as $p \to \infty$, the norm is $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$.

[2]The Frobenius norm of a matrix $\boldsymbol{X}$ is $\|\boldsymbol{X}\|_{\mathrm{F}} = (\sum_{ij} x_{ij}^2)^{\frac{1}{2}}$.

The MRSR estimate $\boldsymbol{Y}_{k+1}$ for the responses and the MRSR estimate $\boldsymbol{W}_{k+1}$ for the regression coefficients are updated

$$\boldsymbol{Y}_{k+1} = (1 - \gamma_k)\boldsymbol{Y}_k + \gamma_k \widehat{\boldsymbol{Y}}_{k+1} \tag{6}$$

$$\boldsymbol{W}_{k+1} = (1 - \gamma_k)\boldsymbol{W}_k + \gamma_k \widehat{\boldsymbol{W}}_{k+1}^*, \tag{7}$$

where $\widehat{\boldsymbol{W}}_{k+1}^*$ is an $m \times q$ row sparse matrix whose nonzero rows, indexed by $\mathcal{A}_{k+1}$, are filled with the corresponding rows of $\widehat{\boldsymbol{W}}_{k+1}$.

If we would always choose the step length $\gamma_k = 1$, we had a greedy forward selection algorithm that moves from a least squares solution to another. On the other hand, the step length should be positive in order to improve the model fitting. Thus, $\gamma_k \in (0, 1]$ acts as a shrinking parameter for the regression coefficients of the active inputs, while the coefficients of the nonactive inputs are constrained to zero. A very intuitive way to choose γ_k is proposed in [8], and is further generalized for multiresponse regression (with $p = 1$) in [18]: Move the current estimate $\boldsymbol{Y}_k$ toward the least squares estimate $\widehat{\boldsymbol{Y}}_{k+1}$ until any of the nonactive inputs has the same correlation as the active inputs according to (2). This makes γ_k the smallest positive value such that some new index joins the active set. If the active set already contains all the indices, we simply take a full step $\gamma_k = 1$ to the OLS solution.

According to (4)–(5) we have $\boldsymbol{X}_{k+1}^T \widehat{\boldsymbol{Y}}_{k+1} = \boldsymbol{X}_{k+1}^T \boldsymbol{T}$. By using this and by substituting (6) into (2), the correlations can be written as a function of γ in the next step

$$c_{k+1,j}(\gamma) = |1 - \gamma|\hat{c}_k \quad \text{for } j \in \mathcal{A}_{k+1} \tag{8}$$

$$c_{k+1,j}(\gamma) = \|\boldsymbol{u}_{k,j} - \gamma \boldsymbol{v}_{k,j}\|_p \quad \text{for } j \notin \mathcal{A}_{k+1}, \tag{9}$$

where $\boldsymbol{u}_{k,j} = (\boldsymbol{T} - \boldsymbol{Y}_k)^T \boldsymbol{x}_j$ and $\boldsymbol{v}_{k,j} = (\widehat{\boldsymbol{Y}}_{k+1} - \boldsymbol{Y}_k)^T \boldsymbol{x}_j$. A new input with index $j \notin \mathcal{A}_{k+1}$ enters the model when (8) and (9) are equal. The correct γ_k is the one that has the smallest positive value of such step lengths. The following theorem says that we always find a step length candidate for all $j \notin \mathcal{A}_{k+1}$. Proof is given in the appendix.

Theorem 1: The common correlation curve of active inputs (8) and the correlation curve of a single nonactive input (9) intersect at a unique point on interval $(0, 1]$.

A. Step Length for the 1-Norm Algorithm

Consider the case $p = 1$. The point $\gamma_{k,j}$ in which (8) and (9) intersect on interval $(0, 1]$ can be computed

$$\begin{aligned} \gamma_{k,j} &= \max_{\gamma} \{\gamma : \|\boldsymbol{u}_{k,j} - \gamma \boldsymbol{v}_{k,j}\|_1 \leq (1 - \gamma)\hat{c}_k\} \\ &= \max_{\gamma} \left\{ \gamma : \sum_{i=1}^{q} s_i(u_{k,ji} - \gamma v_{k,ji}) \leq (1 - \gamma)\hat{c}_k \right\} \\ &= \min^+ \left\{ \frac{\hat{c}_k - \sum_{i=1}^q s_i u_{k,ji}}{\hat{c}_k - \sum_{i=1}^q s_i v_{k,ji}} \right\}, \end{aligned} \tag{10}$$

where each $s_i = \pm 1$, so there are total 2^q terms and "$\min^+$" indicates that the minimum is taken only over positive terms. In order to explain the last equality, we note that

$$\hat{c}_k - \textstyle\sum_i s_i u_{k,ji} \geq \hat{c}_k - \|\boldsymbol{u}_{k,j}\|_1 = \hat{c}_k - c_{k,j} > 0$$

holds according to (2)–(3). So whenever $\hat{c}_k - \sum_i s_i v_{k,ji} < 0$ holds, we have a negative lower bound for γ. On the other hand, $\hat{c}_k - \sum_i s_i v_{k,ji} > 0$ implies a positive upper bound for γ. The solution $\gamma_{k,j}$ equals to the smallest of these upper bounds. This is the step length needed for a nonactive input $\boldsymbol{x}_j$ to enter the model. The correct input is the one that enters with the smallest step length

$$\gamma_k = \min_{j \notin \mathcal{A}_{k+1}} \gamma_{k,j}. \tag{11}$$

The step length calculation (10) is used in [18]. However, this approach is practicable only when there are a few responses, because the computation of (10) scales as $\mathcal{O}(2^q)$ given $\boldsymbol{u}_{k,j}$, $\boldsymbol{v}_{k,j}$, and $\hat{c}_k$. Another way is to define an auxiliary function as (8) subtracted from (9), and $\gamma_{k,j}$ is the sole zero of this function on interval $(0, 1]$ according to Theorem 1. Any line search method can be used in finding the zero efficiently.

B. Step Length for the 2-Norm Algorithm

In the case $p = 2$, the correlations (8) and (9) are equal on interval $(0, 1]$ at point

$$\begin{aligned}\gamma_{k,j} &= \min_{\gamma > 0}\{\gamma : \|\boldsymbol{u}_{k,j} - \gamma \boldsymbol{v}_{k,j}\|_2^2 = (1-\gamma)^2 \hat{c}_k^2\} \\ &= \min{}^+ \left\{ \frac{b \pm \sqrt{b^2 - ac}}{a} : \begin{array}{l} a = \hat{c}_k^2 - \|\boldsymbol{v}_{k,j}\|_2^2 \\ b = \hat{c}_k^2 - \boldsymbol{u}_{k,j}^T \boldsymbol{v}_{k,j} \\ c = \hat{c}_k^2 - \|\boldsymbol{u}_{k,j}\|_2^2 \end{array} \right\} \end{aligned} \tag{12}$$

and the computation of (12) scales linearly with the number of outputs $\mathcal{O}(q)$ given $\boldsymbol{u}_{k,j}$, $\boldsymbol{v}_{k,j}$, and $\hat{c}_k$. The 2-norm MRSR algorithm takes the smallest step according to (11).

C. Step Length for the ∞-Norm Algorithm

Finally, consider the $p = \infty$-norm and observe that the correlation curves (8) and (9) intersect on interval $(0, 1]$ at

$$\begin{aligned}\gamma_{k,j} &= \max_{\gamma}\{\gamma : \|\boldsymbol{u}_{k,j} - \gamma \boldsymbol{v}_{k,j}\|_\infty \leq (1-\gamma)\hat{c}_k\} \\ &= \max_{\gamma}\{\gamma : \pm(u_{k,ji} - \gamma v_{k,ji}) \leq (1-\gamma)\hat{c}_k,\ 1 \leq i \leq q\} \\ &= \min_{1 \leq i \leq q}{}^+ \left\{ \frac{\hat{c}_k + u_{k,ji}}{\hat{c}_k + v_{k,ji}}, \frac{\hat{c}_k - u_{k,ji}}{\hat{c}_k - v_{k,ji}} \right\}. \end{aligned} \tag{13}$$

The last equality follows from the fact that

$$\hat{c}_k \pm u_{k,ji} \geq \hat{c}_k - \|\boldsymbol{u}_{k,j}\|_\infty = \hat{c}_k - c_{k,j} > 0$$

holds according to (2)–(3). Thus, if we have $\hat{c}_k \pm v_{k,ji} < 0$, then γ has a negative lower bound. On the other hand, γ has a positive upper bound in the case $\hat{c}_k \pm v_{k,ji} > 0$. The solution $\gamma_{k,j}$ is the most stringent upper bound. The computation of (13) scales as $\mathcal{O}(q)$ given $\boldsymbol{u}_{k,j}$, $\boldsymbol{v}_{k,j}$, and $\hat{c}_k$. Again, Eq. (11) gives the step length for the ∞-norm MRSR algorithm.

III. Related Methods

We consider two related methods, which we compare with the MRSR algorithm in the experiments section: the greedy forward selection (FS) algorithm and Simultaneous Variable Selection (SVS) [17]. The FS algorithm proceeds in a quite similar way as MRSR. New nonzero rows are added to $\boldsymbol{W}_k$ in the model (1) according to the criterion (2). The difference is that the step length is always $\gamma_k = 1$, so $\boldsymbol{W}_k$ equals to the row sparse OLS solution $\widehat{\boldsymbol{W}}_k^*$. This changes potentially the order in which the inputs enter the model.

SVS is in turn characterized by an optimization problem

$$\min \tfrac{1}{2}\|\boldsymbol{T} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 \quad \text{s.t.} \sum_{j=1}^{m} \|\boldsymbol{w}_{(j)}\|_\infty \leq \tau, \tag{14}$$

where $\boldsymbol{w}_{(j)}^T$ denotes the jth row of $\boldsymbol{W}$. The constraint imposes row sparsity, which is controlled by the parameter $\tau \geq 0$. The SVS problem (14) can be rewritten as a linearly constrained quadratic optimization problem and thereby it can be solved by standard techniques. However, this formulation is practicable only for relatively low-dimensional problems. A recent work [19] focuses in the design of an efficient algorithm that follows the solution path as a function of τ. It is suggested in [17] to apply the solution of (14) only to identify a suitable subset of inputs. In the experiments, we compute the OLS estimates using the inputs selected by SVS.

Interestingly, the MRSR algorithm is also related to single response regression techniques that select groups of regression coefficients. Suppose that we stack the columns of $\boldsymbol{T}$ into a $qn \times 1$ vector and the columns of $\boldsymbol{W}$ into a $qm \times 1$ vector, and copy $\boldsymbol{X}$ into the diagonal blocks of a new $qn \times qm$ input data matrix. If we use this parameterization and form m groups, each of which contains the q regression coefficients associated with one of the inputs, we observe that the Group LARS algorithm [20] is exactly the 2-norm MRSR algorithm. Given this connection and assuming $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, the 2-norm MRSR algorithm follows the solution path of the problem

$$\min \tfrac{1}{2}\|\boldsymbol{T} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 \quad \text{s.t.} \sum_{j=1}^{m} \|\boldsymbol{w}_{(j)}\|_2 \leq \tau. \tag{15}$$

For a general $\boldsymbol{X}$ and $q \geq 2$ the path is piecewise nonlinear and the 2-norm MRSR algorithm is only an approximation.

IV. Experiments

A. Experiments with Simulated Data

MRSR, FS, and SVS are compared to each other according to prediction accuracy and correctness of input selection using simulated data. The collinearity of input variables has a strong effect on linear models and this experiment illustrates the effect. Data are generated from the setting $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, where $\boldsymbol{E}$ denotes the error term. The number of observations, responses, and inputs are $n = 50$, $q = 5$, and $m = 100$, respectively.

The input data are generated according to an m-dimensional normal distribution with zero mean and covariance $\boldsymbol{\Sigma}_x$,

$$\boldsymbol{x}_{(i)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_x), \quad \text{where} \quad [\boldsymbol{\Sigma}_x]_{ij} = \sigma_x^{|i-j|}.$$

The parameter σ_x controls the covariance and we consider the values $\sigma_x = 0$, $\sigma_x = 0.5$, and $\sigma_x = 0.9$. The correlation between all the inputs is zero with $\sigma_x = 0$. A few inputs have medium correlation with $\sigma_x = 0.5$ and the average is

0.02. The average correlation is 0.16 with $\sigma_x = 0.9$, but then there are some strong correlations among the inputs. Errors are distributed according to a q-dimensional normal distribution with zero mean and covariance $\boldsymbol{\Sigma}_e$,

$$\boldsymbol{e}_{(i)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_e), \quad \text{where} \quad [\boldsymbol{\Sigma}_e]_{ij} = 0.2^2 \cdot 0.6^{|i-j|}.$$

The errors have an equal standard deviation 0.2 and there are correlations between the errors of different responses.

The actual matrix of regression coefficients $\boldsymbol{B}$ is set to have a row sparse structure by selecting 20 rows out of the total number of 100 rows randomly. The values of the coefficients in the selected rows are independently normally distributed with zero mean and unit variance. The other 80 rows are filled with zeros. The columns $\boldsymbol{b}_i$ of the matrix $\boldsymbol{B}$ are scaled after sampling as follows

$$\boldsymbol{b}_i \leftarrow \boldsymbol{b}_i \Big/ \sqrt{\boldsymbol{b}_i^T \boldsymbol{\Sigma}_x \boldsymbol{b}_i} \quad \text{for } i = 1, \ldots, q.$$

This scaling ensures that the responses $\boldsymbol{t}_i$ are comparable, so we can use the mean of the individual mean squared errors

$$\mathrm{MSE} = \frac{1}{q} \sum_{i=1}^{q} (\boldsymbol{b}_i - \boldsymbol{w}_i)^T \boldsymbol{\Sigma}_x (\boldsymbol{b}_i - \boldsymbol{w}_i).$$

as an accuracy measure for different methods [11].

For each value of σ_x the generation of data is replicated 500 times. The maximum number of active inputs is 50 in the MRSR and FS algorithms, because we have only 50 observations. SVS is evaluated at 150 linearly equally spaced points of τ in the range $[0, 6.5]$. For those values of τ, where SVS has selected at most 50 inputs, we also compute the subset OLS solutions. This range of τ varies slightly in the replicated data sets, but it is roughly $[0, 3]$.

The two leftmost columns in Fig. 1 show the average results obtained using MRSR and FS. In the MRSR models, the minimum MSE is achieved using approximately 40 inputs and the minimum is roughly the same with each value of σ_x. This indicates that MRSR is not sensitive to the degree of collinearity among the inputs. The number of selected inputs is larger than the correct number of 20 inputs, but overfitting is avoided due to shrinking of the regression coefficients. The minimum MSEs of the FS methods are similar to the ones of the MRSR methods in the cases of $\sigma_x = 0$ and $\sigma_x = 0.5$. They are achieved using only 20 inputs, which are nearly all correctly selected. On the other hand, the advantage of MRSR over FS can be evidently seen when $\sigma_x = 0.9$. The minimum MSE of MRSR is smaller than the minimum MSE of FS. In addition, MRSR does better in input selection. The both algorithms choose inputs equally correctly up to the models with 20 inputs. After that, MRSR performs better.

The two rightmost columns in Fig. 1 encapsulate the average results for SVS. The MSE of the SVS model decreases slowly as a function of τ with each value of σ_x. The minimum MSEs are achieved using approximately 80 inputs. The correct inputs are included, but there are clearly too many false ones. The calculation of the OLS estimates using the inputs selected by SVS (SVS+OLS) decreases the MSEs in the cases of $\sigma_x = 0$ and $\sigma_x = 0.5$. However, the minimum MSEs are still larger than in the MRSR models. With each value of σ_x, the most accurate SVS+OLS model includes about 30 inputs. The proportion of correctly selected inputs decreases when the degree of collinearity increases. Since shrinkage is not applied in the final model, false inputs are more harmful, and SVS+OLS is less accurate than MRSR.

The first and third columns in Fig. 2 show the standard deviations (STD) of MSEs. The minimum MSE MRSR models are more stable than the most accurate FS and SVS+OLS models. These MRSR models have also smaller deviations than the minimum MSE SVS models, excluding the case $\sigma_x = 0.9$. In the MRSR models, the deviations of MSEs are roughly the same regardless of the value of σ_x. Interestingly, the deviations of the other methods decrease as the value of σ_x increases.

The second and fourth columns in Fig. 2 show the STD of the number of correct inputs. There is no major difference between MRSR and FS when $\sigma_x = 0$, but otherwise MRSR performs better. In addition, the minimum MSE SVS+OLS models have higher deviations than the most accurate MRSR models. The minimum MSE SVS models have the smallest deviations. It is however quite natural that the deviation is small when the number of selected inputs is large.

B. Experiments with Real Data

In this experiment MRSR, FS, and SVS are applied to three real data sets. The first data set is Chemometrics data, taken from [21]. The data are from a simulation of a low density tubular polyethylene reactor. The set includes $n = 56$ observations of $m = 22$ inputs and $q = 6$ responses. Following [11], we log-transformed the responses, because they are skewed to the right.

The second data set is Macroeconomic data, which is taken from [22]. It is a 10-dimensional time-series from the United Kingdom with quarterly measurements. The data contain $n = 36$ observations of $q = 5$ responses and five inputs. A quadratic model with all the terms x_j, x_j^2, and $x_i x_j$ is fitted, which increases the total number of inputs to $m = 20$. The time-dependency of the observations is ignored.

The last data set is called Chemical reaction data, and it is obtained from [14]. There are $n = 19$ observations of $q = 3$ responses and three inputs from a planned chemical reaction experiment. The quadratic model is also used in this case, which gives the total number of inputs $m = 9$.

All the responses and inputs were normalized to zero mean and unit variance before the analysis to make the results more comparable. The average absolute correlation between the inputs is 0.44, 0.89, and 0.45 in Chemometrics, Macroeconomic, and Chemical reaction data, respectively. The actual models are not known, so accuracy is estimated using the average squared leave-one-out cross-validation (LOO-CV) error.

The MRSR and FS algorithms are evaluated at each breakpoint, where a new input variable enters the model. SVS is evaluated at 500 values of τ, which are logarithmically equally spaced in the range $[0.01, \ \tau_{\mathrm{OLS}}]$. The value τ_{OLS}

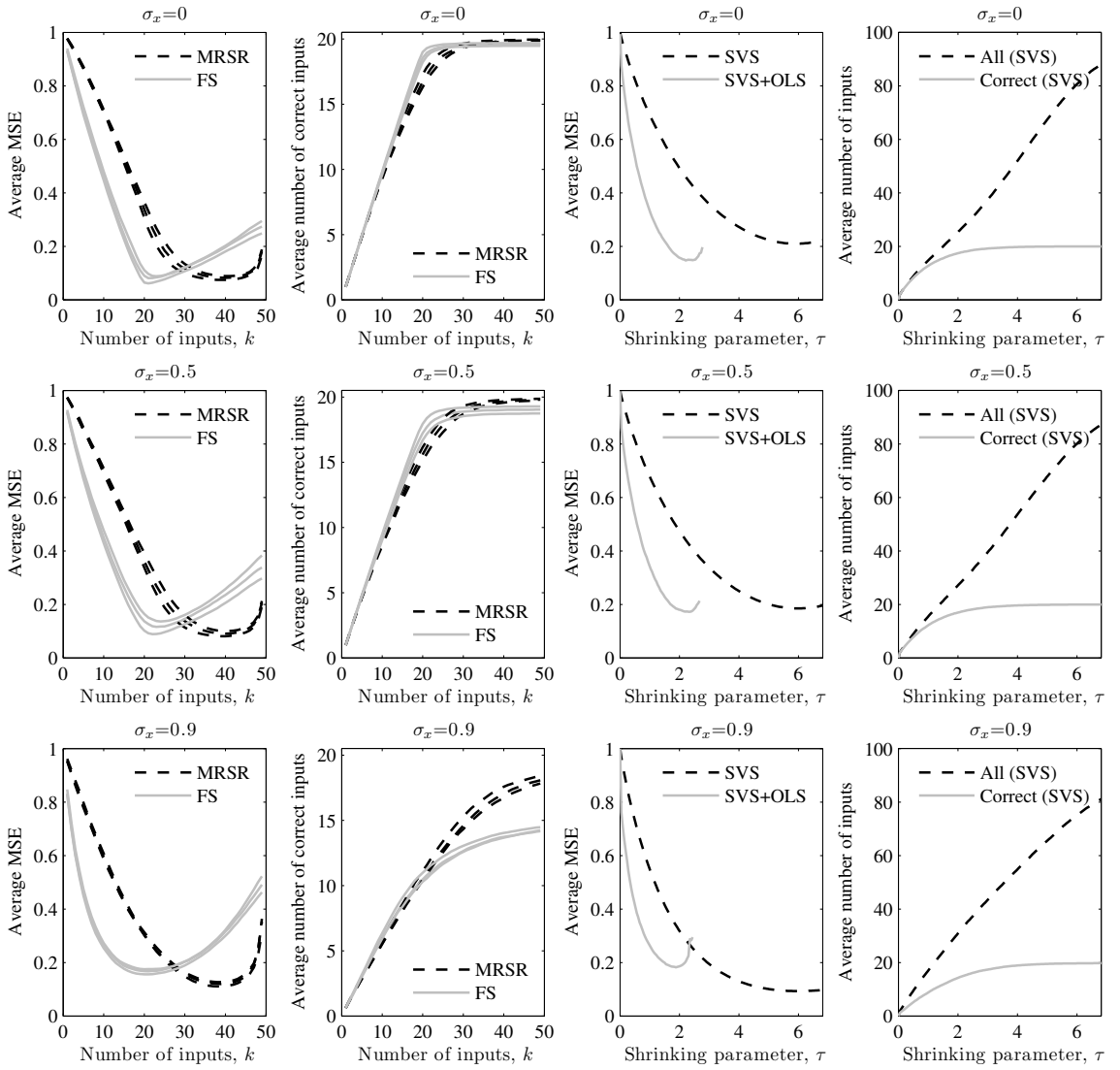

Fig. 1. Averages calculated from 500 replicates of simulated data. In the two leftmost columns, the three dashed black lines and three solid gray lines correspond to different norms, $p = 1,\ 2,\ \infty$, in MRSR and FS, respectively.

corresponds to the sum of norms in (14), evaluated at the full OLS solution. The number of inputs in the SVS models can vary during the cross-validation for a fixed value of τ.

The results are summarized in Table I. All the methods perform equally well according to the LOO-CV error with all the three data sets. The minimum errors are approximately the same and they are clearly within the standard deviations, which are notably large. The errors are adequate for Chemometrics and Macroeconomic data but worse for Chemical reaction data. Observe that all the subset models are more accurate than the full OLS solution. However, it is hard to evaluate the correctness of input selection with our knowledge of the data sets. A couple of comments can still be made. Firstly, the 2-norm FS algorithm selects fewer inputs than the other methods for Chemometrics data. Secondly, the SVS+OLS model has fewer inputs than the SVS model for Macroeconomic data. Thirdly, the number of inputs selected by MRSR is higher than by the other methods for Chemical reaction data.

C. Reconstruction of Image Data

This experiment studies images of handwritten digits $0, \ldots, 9$ with 28×28 pixel grid and the data are taken from

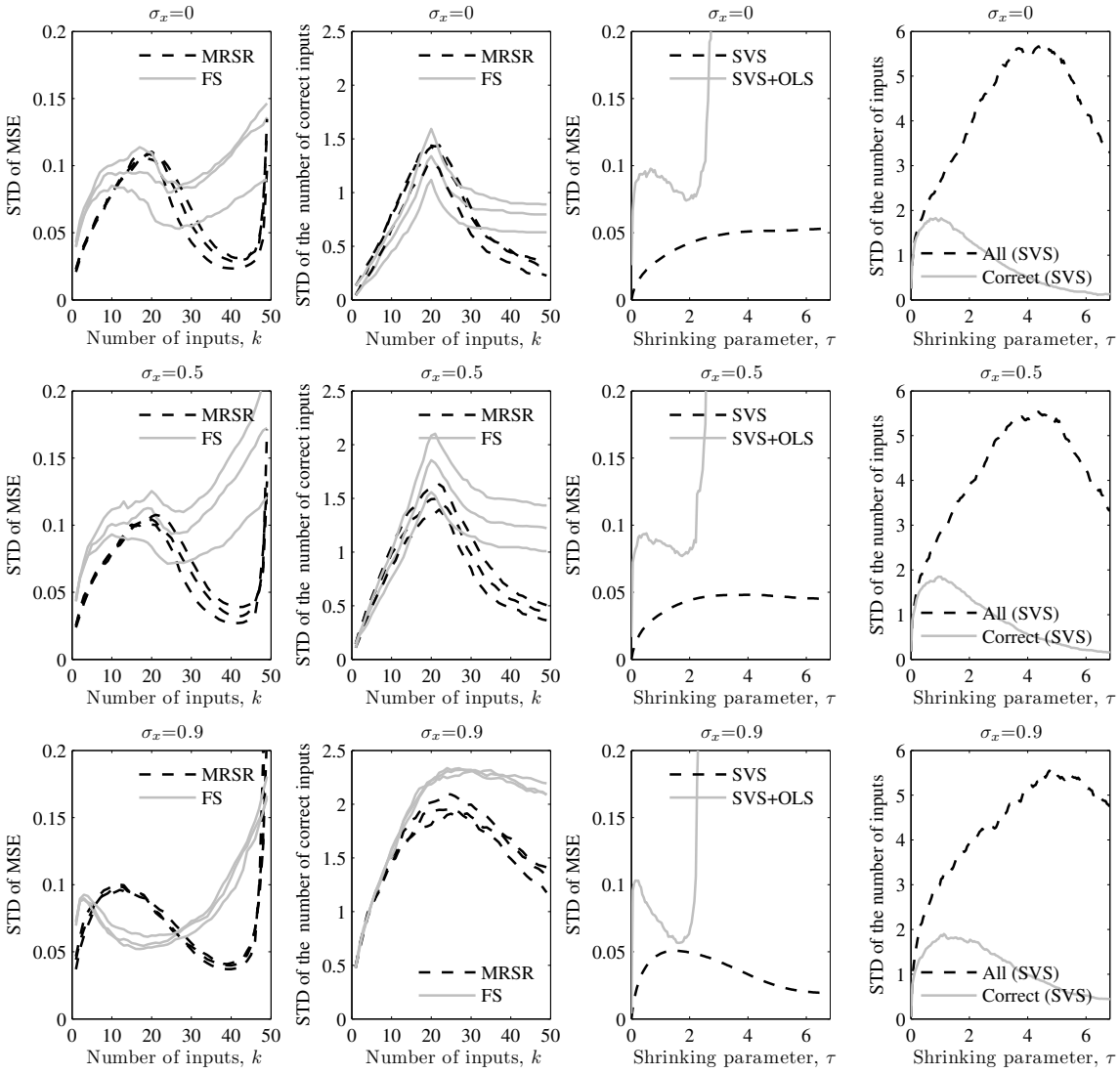

Fig. 2. Standard deviations (STD) calculated from 500 replicates of simulated data. In the two leftmost columns, the three dashed black lines and three solid gray lines correspond to different norms, $p = 1, 2, \infty$, in MRSR and FS, respectively.

TABLE I

RESULTS FOR REAL DATA SETS. (UPPER VALUE) THE LOO-CV ERROR WITH STANDARD DEVIATION. (LOWER VALUE) THE NUMBER OF SELECTED INPUTS. AVERAGE NUMBERS WITH STANDARD DEVIATIONS ARE REPORTED FOR SVS AND SVS+OLS.

	L1 MRSR	L2 MRSR	L∞ MRSR	L1 FS	L2 FS	L∞ FS	SVS	SVS+OLS	Full OLS
Chemo–metrics	0.24 (0.47) 20	0.25 (0.55) 17	0.24 (0.52) 16	0.21 (0.40) 16	0.20 (0.31) 9	0.22 (0.46) 16	0.20 (0.35) 17.4(0.7)	0.22 (0.46) 17.0(0.2)	0.42 (1.02) 22
Macro–economic	0.21 (0.18) 11	0.20 (0.17) 11	0.22 (0.23) 11	0.23 (0.18) 11	0.22 (0.16) 11	0.20 (0.17) 10	0.22 (0.18) 18.2(1.0)	0.19 (0.18) 7.0(0.3)	0.36 (0.33) 20
Chemical reaction	0.38 (0.42) 8	0.40 (0.43) 8	0.41 (0.44) 8	0.40 (0.40) 6	0.39 (0.40) 6	0.33 (0.36) 4	0.35 (0.38) 6.3(0.6)	0.33 (0.39) 5.6(0.8)	0.70 (1.22) 9

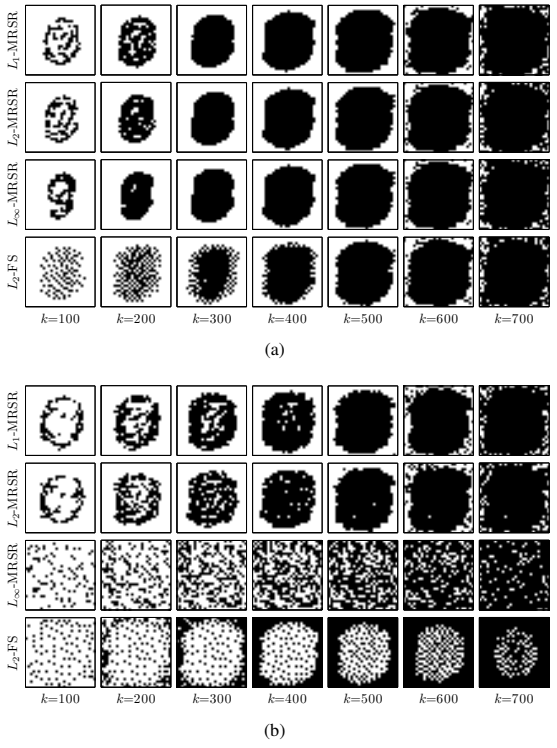

Fig. 3. Selected inputs, marked with black pixels, in the reconstruction. Results for models fitted with (a) unscaled and (b) scaled training data using the FS and MRSR algorithms. L_p denotes the p-norm correlation and k is the number of selected inputs. L_1-FS and L_∞-FS look similar to L_2-FS.

the MNIST database[3]. An image is represented as a vector with $m = 784$ elements containing the grayscale values of pixels in the image, scaled between zero and one. We constructed two distinct data sets randomly: a training set with 100 observations per digit (total $n = 1000$ images) and a test set with 150 observations per digit (total $n = 1500$ images). Then we added independently normally distributed noise with zero mean and standard deviation 0.0448 to the data sets. The value 0.0448 amounts 10% of the maximum standard deviation over all pixels in the unnoisy training data. The aim is to reconstruct the original images from the noisy ones by using the model (1), where both $\boldsymbol{X}$ and $\boldsymbol{T}$ contain the same set of data. Since $\boldsymbol{W}_k$ is row sparse, only some of the pixels are used in the reconstruction $\boldsymbol{Y}_k$.

The first set of results corresponds using training data, where the mean values of the variables are all zero, but the scaling is left untouched. Fig. 3(a) shows that FS and MRSR are both able to select reasonable inputs for all choices of the norm $p = 1, 2, \infty$ in the criterion (2), because relevant information is in the middle of the images. Fig. 4(a) shows the mean squared error between $\boldsymbol{Y}_k$ and the original unnoisy images. All algorithms perform better as the number of active

[3]http://yann.lecun.com/exdb/mnist/

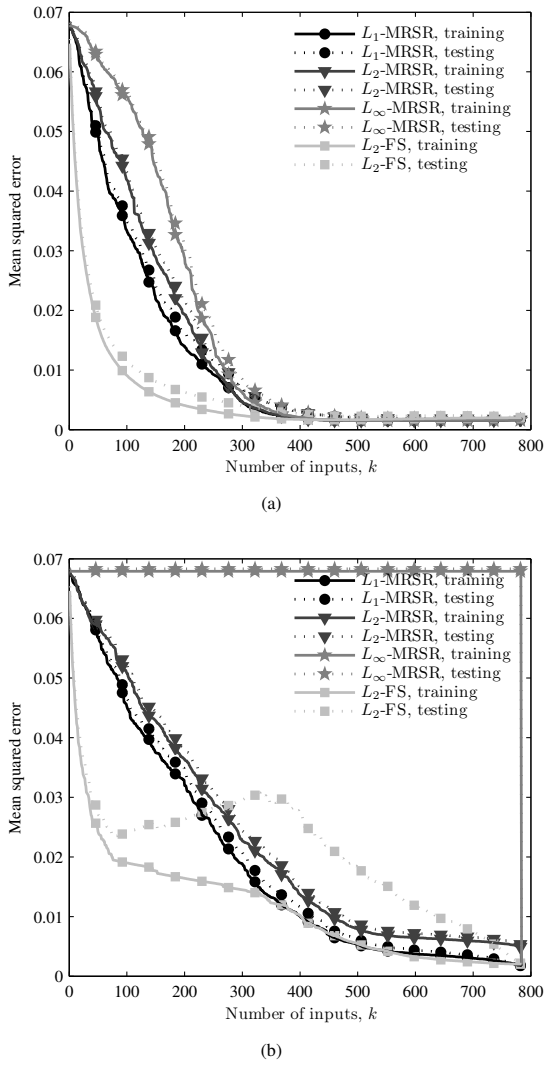

Fig. 4. Mean squared errors for training and test sets in the reconstruction. Results for models fitted with (a) unscaled and (b) scaled training data as a function of the number of selected inputs k. L_1-FS performs similar to L_2-FS and L_∞-FS is worse than L_2-FS.

inputs increases in terms of both training and testing errors. The error of the full model is essentially the variance of the noise added to the images. The message is that the training error of FS decreases most rapidly. As an example, the error of FS with 100 inputs is similar to the errors of all the MRSR variants with more than 250 inputs.

The second set of results corresponds using training data with the variables scaled to zero mean and unit variance. Now the problem is a bit harder as the unimportant inputs are not discarded simply by their low variance. Fig. 3(b) shows that only MRSR with the 1-norm and 2-norm correlation criteria are able to select reasonable inputs. Because the inputs and

responses are equal and scaled, we have $\|\boldsymbol{T}^T\boldsymbol{x}_j\|_\infty$ constant for all $\boldsymbol{x}_j$. Thus, the ∞-norm MRSR algorithm selects inputs randomly, but it makes the first positive update only after all of them have been selected and it goes straight to the least squares solution, see Fig. 4(b). FS reduces the training error very rapidly, but the testing error starts to rise after 100 active inputs and then it falls again after 350 active inputs. Based on Fig. 3(b), FS focuses mostly on modeling the noise after it has selected 100 inputs.

V. Conclusions

We presented the MRSR algorithm for input selection in multiresponse linear regression. It is a forward selection procedure that extends the LARS algorithm for several response variables. The selection criterion is the p-norm of the vector of correlations, measured between an input variable and residuals corresponding to all the response variables. MRSR adds inputs sequentially to the model such that the added input correlates most with the current residuals. The order in which the inputs enter the model reflects their importance. The MRSR algorithm updates always toward the current least squares solution but does not reach it until in the final step. The method is thus less greedy than a pure forward selection.

We considered the 1-norm, 2-norm, and ∞-norm criteria in detail. Based on experiments with simulated and real data, all the three choices of the norm result a quite similar performance. Computational simplicity and connections to the problem (15) under an orthonormality assumption on the inputs may favor the 2-norm algorithm. The simulation experiments also showed that MRSR selects too many inputs, but it is not prone to overfitting due to shrinking. The SVS method suffers from the two-stage estimation approach. If the inputs are not correctly selected in the first stage, overfitting can happen in the second stage. In addition, MRSR is also computationally more efficient. MRSR is better than the greedy forward selection, in terms of prediction accuracy and correctness of selection, when the input variables are correlated. Similar conclusions were drawn from experiments with highly correlated image data. These results are well in line with the fact that collinearities may confuse greedy stepwise methods [2]. All the subset methods were equally accurate and better than the full OLS solution with the three real world data sets that we analyzed.

Appendix

Proof of Theorem 1

(Existence) Define $f_1(\gamma) = (1-\gamma)\hat{c}_k$, which coincides to (8) for $\gamma \leq 1$, and denote (9) by $f_2(\gamma)$ for some fixed $j \notin \mathcal{A}_{k+1}$. According to (2)–(3), we have $f_2(0) = c_{k,j} < \hat{c}_k = f_1(0)$. We also know that $f_2(1) \geq 0 = f_1(1)$ due to nonnegativity of a norm. If $f_2(1) = 0$, the existence is proved by $\hat{\gamma} = 1$. On the other hand, assume that $f_2(1) > 0$ and denote $f(\gamma) = f_1(\gamma) - f_2(\gamma)$, which is continuous on a closed interval $\gamma \in [0,1]$ such that $f(0) > 0$ and $f(1) < 0$. According to Bolzano Theorem there exists a number $\hat{\gamma} \in (0,1)$ with $f(\hat{\gamma}) = 0$, which proves the existence of a solution.

(Uniqueness) Assume, to the contrary, that there exist two points $0 < \hat{\gamma}_1 < \hat{\gamma}_2 \leq 1$ such that $f_1(\hat{\gamma}_1) = f_2(\hat{\gamma}_1)$ and $f_1(\hat{\gamma}_2) = f_2(\hat{\gamma}_2)$. Observe that $f_2(\gamma)$ is a convex function, which can be proved by triangle inequality. Thus, there exists an affine function $g(\gamma) = a\gamma + b$ such that $g(\hat{\gamma}_1) = f_2(\hat{\gamma}_1)$ and $g(\gamma) \leq f_2(\gamma)$ for $\gamma \in \mathbb{R}$. Combining this to the assumption gives $g(\hat{\gamma}_1) = f_1(\hat{\gamma}_1)$ and $g(\hat{\gamma}_2) \leq f_1(\hat{\gamma}_2)$, which means $b \geq \hat{c}_k$. However, this leads to the contradiction $c_{k,j} = f_2(0) \geq g(0) = b \geq \hat{c}_k$, which proves the uniqueness of the solution. ■

References

[1] A. J. Miller, *Subset Selection in Regression*, London: Chapman & Hall, 1990.

[2] R. R. Hocking, "Developments in linear regression methodology: 1959–1982," *Technometrics*, vol. 25, pp. 219–249, Aug. 1983.

[3] M. Sulkava, J. Tikka, and J. Hollmén, "Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees," *Ecological Modelling*, vol. 191, pp. 118–130, Jan. 2006.

[4] J. Copas, "Regression, pediction and shrinkage," *Journal of the Royal Statistical Society*, series B, vol. 45, pp. 311–354, 1983.

[5] B. Abraham and G. Merola, "Dimensionality reduction approach to multivariate prediction," *Computational Statistics & Data Analysis*, vol. 48, pp. 5–16, Jan. 2005.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, series B, vol. 58, pp. 267–288, 1996.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, series B, vol. 67, pp. 301–320, Apr. 2005.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, Apr. 2004.

[9] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, pp. 389–404, July 2000.

[10] A. J. Burnham, J. F. MacGregor, and R. Viveros, "Latent variable multivariate regression modeling," *Chemometrics and Intelligent Laboratory Systems*, vol. 48, pp. 167–180, Aug. 1999.

[11] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multivariate regression," *Journal of the Royal Statistical Society*, series B, vol. 59, pp. 3–54, 1997.

[12] M. S. Srivastava and T. K. S. Solanky, "Predicting multivariate response in linear regression model," *Communications in Statistics – Simulation and Computation*, vol. 32, pp. 389–409, May 2003.

[13] B. E. Barrett and J. B. Gray, "A computational framework for variable selection in multivariate regression," *Statistics and Computing*, vol. 4, pp. 203–212, 1994.

[14] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd Edition, New York: John Wiley & Sons, 2002.

[15] E. J. Bedrick and C-L. Tsai, "Model Selection for multivariate regression in small samples," *Biometrics*, vol. 50, pp. 226–231, Mar. 1994.

[16] P. J. Brown, M. Vanucci, and T. Fearn, "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society*, series B, vol. 64, pp. 519–536, Aug. 2002.

[17] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, pp. 349–363, Aug. 2005.

[18] T. Similä and J. Tikka, "Multiresponse sparse regression with application to multidimensional scaling," *Proc. International Conference on Artificial Neural Networks (ICANN)*, Warsaw, Poland, Sept. 2005, pp. 97–102.

[19] B. A. Turlach, "On homotopy algorithms in statistics," keynote talk during the *Symposium on Optimisation and Data Analysis* in honour of Mike Osborne's 70th birthday, Canberra, ACT, Australia, Sept. 2005.

[20] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, series B, vol. 68, pp. 49–67, 2006.

[21] M. S. Srivastava, *Methods of Multivariate Statistics*, New York: John Wiley & Sons, 2002.

[22] G. C. Reinsel and R. P. Velu, *Multivariate Reduced-Rank Regression, Theory and Applications*, New York: Springer-Verlag, 1998.