Simultaneous Variable Selection

Berwin A Turlach[*] William N Venables[†] Stephen J Wright[‡]

September 16, 2004

Abstract

We propose a new method for selecting a common subset of explanatory variables where the aim is to model *several* response variables. The idea is a natural extension of the LASSO technique proposed by Tibshirani (1996) and is based on the (joint) residual sum of squares while constraining the parameter estimates to lie within a suitable polyhedral region. The properties of the resulting convex programming problem are analyzed for the special case of an orthonormal design. For the general case, we develop an efficient interior point algorithm. The method is illustrated on a data set with infra-red spectrometry measurements on 14 qualitatively different but correlated responses using 770 wavelengths. The aim is to select a subset of the wavelengths suitable to use as predictors for as many of the responses as possible.

KEY WORDS AND PHRASES. Constrained least squares problem, constrained regression, convex programming, infra-red spectrometry, interior-point algorithm, quadratic programming, subset selection, variable selection.

AMS (1991) SUBJECT CLASSIFICATION. Primary 62F07, Secondary 62J07, 90C51.

[*] School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

[†] CSIRO Mathematical and Information Sciences, PO Box 120, Cleveland Qld 4163, Australia

[‡] Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, USA

1 Introduction

Many practical (linear) regression problems are ill-conditioned. When the problem contains a large number of highly correlated predictors, the need to select predictors carefully, or otherwise regularize the problem, is well-known. Some traditional techniques used for this purpose are direct variable selection (Miller, 1990; Burnham and Anderson, 1998), ridge regression (see, among others, Hocking, 1996; Draper and Smith, 1998) and partial least squares (Wold, 1984; Martens and Naes, 1989; Brown, 1993). The latter is typically used if the number of explanatory variables is large relative to the number of observations. Newer techniques that have been proposed include the *non-negative garotte* by Breiman (1995) and the *least absolute shrinkage and selection operator* (LASSO) by Tibshirani (1996). Below we shall discuss the LASSO in more detail, and propose a method that extends the LASSO methodology in a natural way to the problem in which several related response variables are observed and the researchers desires, either because of available *a priori* information or for other reasons, to model these response variables using the same common subset of predictors.

Breiman and Friedman (1997) discuss various applications in which the aim is to model several related response variables using the same set of predictors, and propose a method that uses the relationship between the response variables to find a “simultaneous” model for each response variable. They show that such a simultaneous model can outperform an approach in which each response variable is modelled separately. However, their approach uses all available predictors to build simultaneously models for all of the response variables and they do not address the question of variable selection.

Using a Bayesian approach, Brown *et al.* (1998, 1999, 2002) address the question of variable selection in the setting where one has several related response variable and a (large) set of predictors to choose from. However, their methods require the use of quite sophisticated MCMC algorithms for which the choice of tuning parameters and the monitoring for convergence does not appear to be trivial. By way of contrast, the method that we propose for variable selection in this setting is based on a regularization approach inspired by the LASSO methodology. Although we are aware that many variable selection procedures that use a regularization approach, including the LASSO, can be explained via a Bayesian framework (see, for example, Leamer, 1978), we do not develop a Bayesian interpretation for our method in this paper.

The LASSO technique minimizes the residual sum of squares while bounding the L_1 norm of the coefficient vector by a specified value. Suppose that we observe data on a response variable y_i and p explanatory variables x_{il} ($i = 1, \dots, n$ and $l = 1, \dots, p$), and that the response variable is centered ($\sum_i y_i = 0$) and the explanatory variables are standardized ($\sum_i x_{il} = 0$ and $\sum_i x_{il}^2/n = 1$ for all $l = 1, \dots, p$). Then the LASSO estimates are given by the solution to the following optimization problem

$$\underset{b_1,\dots,b_p}{\text{minimize}} \qquad \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{p} x_{il}b_l\right)^2 \tag{1.1a}$$

$$\text{subject to} \qquad \sum_{l=1}^{p} |b_l| \le t \tag{1.1b}$$

Tibshirani (1996) shows that this approach has features in common with both ridge regression and variable selection. As in ridge regression, the solution $\hat{b}_i$ of (1.1) tends to shrink to zero as t goes to zero. On the other hand, the non smooth nature of the L_1 norm, which is is non differentiable when any components b_l are zero, tends to force some of the solution components $\hat{b}_i$'s to be zero. In this sense, the outcome is similar to variable selection.

In this paper we propose a method that extends the LASSO methodology in a natural way to the problem in which several related response variables are observed and the researcher's aim is to find predictors for all of them from a common subset of variables. The data set that motivated this research and which we will use to illustrate our methodology was kindly provided by Dr Bronwyn Harch of CMIS/CSIRO in Adelaide. In this data set, experimenters used 24 soil samples to take measurements on 14 quantities (EC, pH, pHCaCl2, CLeco, Org.C, NLeco, extP, Ca, Mg, Na, K, TotalCations, CEC and CaCO3) at 770 different wavelengths. The aim is to identify those wavelengths (explanatory variables) that are the most informative for detecting and quantifying the presence of a particular quantity, say organic C (Org.C).

If we apply the LASSO methodology to each response variable of this data set, choosing values for t between 0 and 1, we obtain the results shown in Figure 1.1. Here, the 770 different wavelengths, approximately equally spaced along a spectrum, were labeled X1 to X770 for simplicity. In each panel, the abscissa is the constraint bound t in (1.1b) and the ordinate is the coefficient value. (Since the predictors are scaled to zero mean and unit variance, the coefficients are also on comparable scales.) Note that for most of the response variables only a very few of the coefficients are non-zero for any given value of t, but the set of non-zero coefficients depends strongly on t. Typically a regressor enters the model (that is, has a non-zero coefficient) and drops out again to be replaced by a "nearby" regressor. For example, consider the panel for CaCO3. Initially, for small t, X116 is selected by the LASSO. As t increases to about 0.8, X116 is replaced by X117, which in turn is replaced by X119 as t approaches 1. These three regressors are highly correlated[1].

A feature in Figure 1.1 that is not so obvious is that at $t = 1$, the set of variables selected for most response variables includes only a few regressor variables. Moreover, the indices of these regressor variables are mostly in a few particular regions of the spectrum.

Figure 1.2 shows a pairwise scatterplot of the 14 response variables. We observe at least some correlation between most of the response variables; in particular, between EC, pH, pHCaCl2, Cleco, Org.C, Nleco, Ca, TotalCation, and CEC. With regard to NLeco we would also like to point out that one observation, namely the $10th$, appears to be an outlier.

The correlation between the response variables and the results of the separate LASSO analysis suggest that possibly a single set of regressor variables is sufficient to model all (or at least most) of the response variables. One may also expect that by using all response variables *simultaneously* to select a single set of regressor variables, it is possible to avoid overfitting, which is potentially a serious problem if we select a separate set of regressor variables for each response variable.

[1] Given this high correlation between regressors, another recently proposed modification of the LASSO, the "elastic net" (Zou and Hastie, 2003) might be more appropriate if one wants to model the response variables separately.

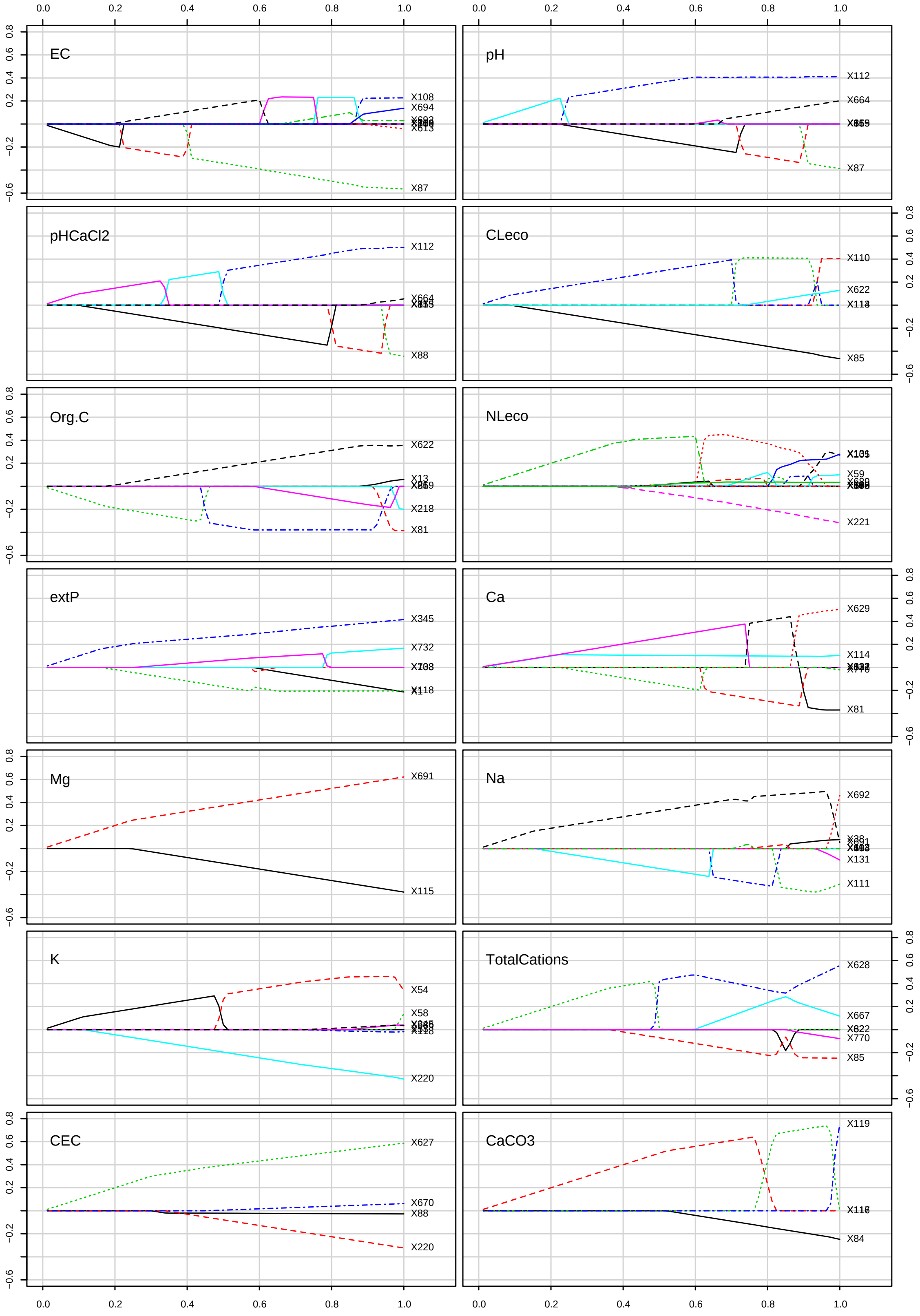

Figure 1.1: Applying the LASSO to each variable separately.

1.1 Extending the LASSO to multiple responses

We propose to extend the LASSO methodology to achieve simultaneous variable selection. To fix notation, suppose that we have n observations on k response variables y_{ij}, and p explana-

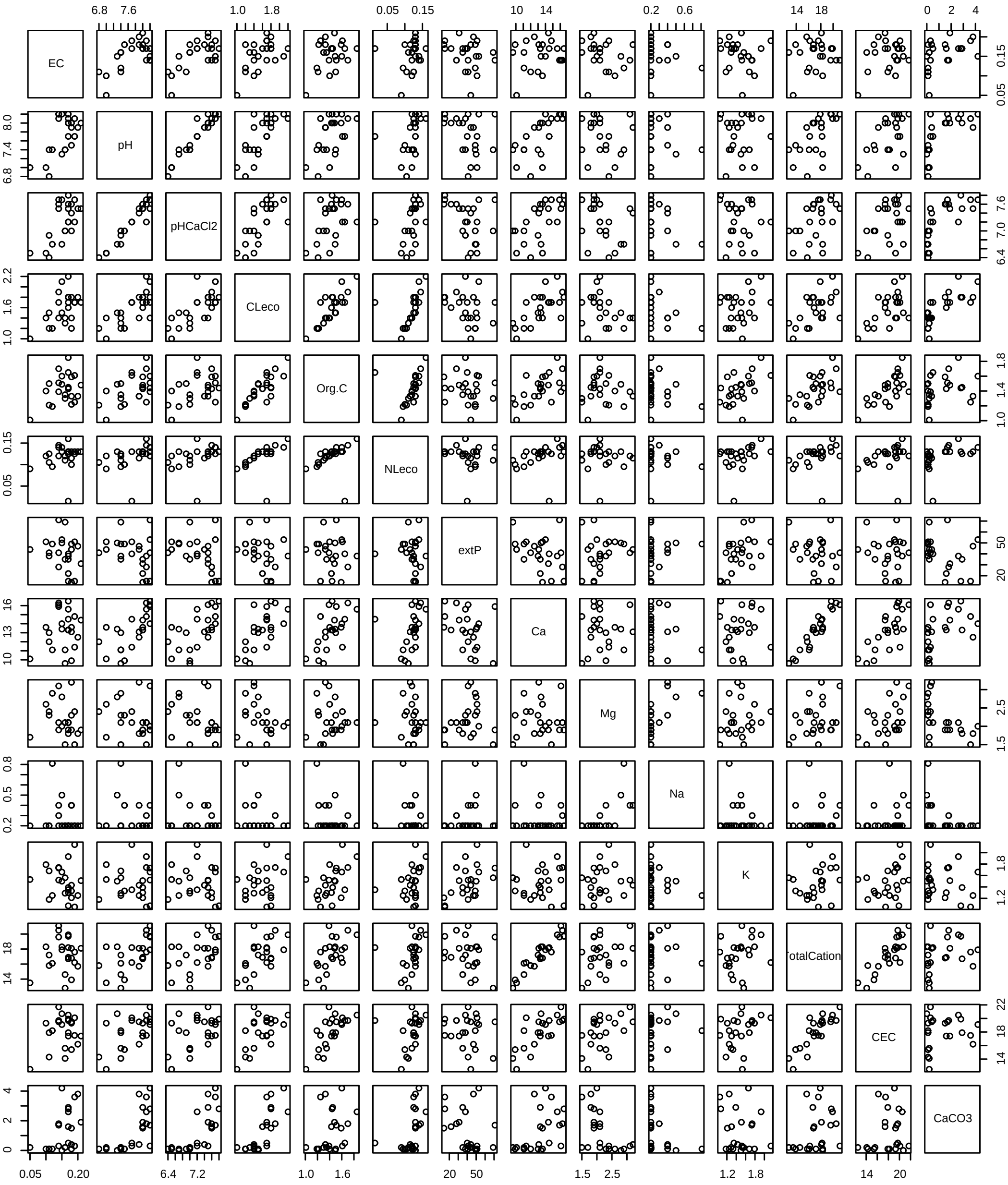

Figure 1.2: Pairwise plot of the 14 response variables

tory variables x_{il} ($i = 1, \ldots, n$, $j = 1, \ldots, k$ and $l = 1, \ldots, p$). We assume that not only are the explanatory variables standardized as described above, but that the response variables are as well; that is, $\sum_i y_{ij} = 0$ and $\sum_i y_{ij}^2/n = 1$ for all $j = 1, \ldots, k$. We might interpret the regression parameter b_{lj} as the "explanatory power" that the l-th regressor variable has on the j-th response variable. It seems natural to take $b_{\max}^l = \max(|b_{l1}|, \ldots, |b_{lk}|)$ as a measure of the "simultaneous explanatory power" of the l-th regressor on all k response variables. Following the approach of Tibshirani (1996) we may now impose a constraint on the sum of the $b_{\max}^l$, $l = 1, \ldots, p$, in order to identify the regressor variables that simultaneously best explain all response variables. Thus, we arrive at the following problem

$$\underset{b_{11},\ldots,b_{pk}}{\text{minimize}} \qquad \frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij}-\sum_{l=1}^{p}x_{il}b_{lj}\right)^2, \tag{1.2a}$$

$$\text{subject to} \qquad \sum_{l=1}^{p}\max(|b_{l1}|,\ldots,|b_{lk}|) \leq t. \tag{1.2b}$$

Note that if $k = 1$, (1.2) reduces to the LASSO (1.1). It should also be noted that we propose to use (1.2) as an exploratory tool to identify a suitable subset of regressor variables. Once this subset is identified, we suggest that its suitability for modelling most (or all) of the response variables is assessed further using standard statistical techniques, and that the selected regressor variables be used in *unconstrained* (linear) models. It is not clear to us that the actual parameter estimates at the solution of (1.2) have any inherent meaning or use.

While for the data set that we use to illustrate our methodology it was not crucial that the same set of predictors can be used to model all response variables, there may be situation where the ability to identify a set of predictors to model all response variables is crucial. For example[2], a manufacturer of high-frequency measurement devices produces an instrument that is designed to meet several different specifications (likely correlated) for all carrier frequencies in a given range. However, production engineers would not be able to afford to test every single frequency (e.g. 1MHz, ..., 500 MHz) to verify that the instrument coming off the production line passes all specifications at all frequencies. It would be to their advantage to find a small subset of frequencies that can be used to verify performance at *all* specifications. The engineers would be able to save enormous amounts of time by setting their signal generators to only a few frequencies, and to test the instrument response to all specifications as the frequency is changed from one setting to the next. We suggest that our methodology might be helpful in identifying such a subset of frequencies.

The rest of the paper is structured as follows. In Section 2 we provide some further motivation for the method that we propose and discuss how it relates to similar work by others. In Section 3 we give an exact characterization of the solutions to (1.2). We also describe a homotopy algorithm that calculates all solutions (as functions of t) in the case where the design matrix is orthonormal and develop an interior point algorithm for the general case. Using the latter algorithm, we reanalyze the infra–red spectrometry data in Section 4. Further discussion on how this method can be extended is given in Section 5 and some conclusions are offered in Section 6.

2 Some motivation and discussion of related work

The proposed methodology can be viewed as a way to select groups of regression estimates. That is, in a (potentially huge) regression problem with m parameters, $\beta_1,\ldots,\beta_m$, we partition the index set $\sigma = \{1,\ldots,m\}$ into p disjoint sets, σ_l, such that $\sigma = \cup_l\sigma_l$, and seek the "most significant" groups of parameters. We allow β_i to be nonzero only if i belongs to one of the selected groups σ_l.

In Section 3 we show that our problem can be viewed from this perspective. Another

[2] The authors would like to thank Prof. K. Kafadar for bringing this example to their attention.

problem that fits into this setting is variable selection in generalized additive models (Hastie and Tibshirani, 1990) as discussed by Bakin (1999). In this case, each nonparametric function in the generalized additive model is built from a B-spline basis, and the corresponding coefficients are collected into a group. By deciding which groups are "significant," Bakin (1999) essentially identifies those regressor variables that have a significant influence on the response variable.

How can such a groupwise selection of regression parameters be achieved? By generalizing other methods studied in the statistical literature (Leamer, 1978; Frank and Friedman, 1993; Tibshirani, 1996), one might consider imposing the following constraint onto the parameter estimates

$$\sum_{l=1}^{p} \left(\sum_{i \in \sigma_l} |\beta_i|^\alpha \right)^{\frac{1}{\alpha}} = \sum_{l=1}^{p} \|\beta_{\sigma_l}\|_\alpha \leq t, \tag{2.1}$$

where $t \geq 0$ is some constant, β_{σ_l} is the vector consisting of those β_is for which $i \in \sigma_l$, and $\| \cdot \|_\alpha$ is the L^α-norm. If $\alpha \geq 1$, the feasible region in (2.1) is a convex subset of $\mathbb{R}^m$. This property is advantageous from both numerical and theoretical points of view. It ensures that a solution exists if the parameter estimates are defined as the minimizer of a (strictly) convex function. In fact, for a strictly convex objective function the solution is unique. If the objective function is not strictly convex then one can ensure that the solution is unique under further regularity conditions, provided t is small enough; see for example the discussion in Osborne *et al.* (2000b) for the special case of the LASSO.

When t is small enough, the solution of the regression problem with constraint (2.1) lies on the boundary of the feasible set, that is, equality holds in (2.1). Thus, it is clear that imposing a constraint like (2.1) shrinks the parameter estimates towards zero as t goes to zero. The size of this shrinkage and the manner in which the parameter estimates are shrunk to zero, however, depends on the particular choice of α. For $\alpha = 1$ we have

$$\sum_{l=1}^{p} \|\beta_{\sigma_l}\|_\alpha = \sum_{l=1}^{p} \sum_{i \in \sigma_l} |\beta_i| = \sum_{l=1}^{m} |\beta_l| = \|\beta\|_1$$

and the constraint (2.1) reduces to the L^1-norm of the complete vector of parameter estimates, which is the constraint used by Tibshirani (1996) in his LASSO method. Given the behavior of the LASSO method (Tibshirani, 1996; Osborne *et al.*, 2000a,b), it is clear that this choice does not achieve the desired "simultaneous" variable selection, so the choice $\alpha = 1$ is not interesting in this context.

Arguably, the most obvious choices for $\alpha > 1$ would be $\alpha = 2$ and $\alpha = \infty$. The former choice was studied by Bakin (1999), while we study the latter choice $\alpha = \infty$ in this paper. Bakin (1999) notes that the use of $\alpha = 2$ can be interpreted as a hybrid between the LASSO (if $p = m$, i.e. each σ_l contains exactly one index) and ridge regression (if $p = 1$). Likewise, the use of $\alpha = \infty$ leads to a hybrid between the LASSO ($p = m$) and interval-restricted least squares ($p = 1$; Clark and Osborne, 1988). We note that the optimization problem with $\alpha = 2$ cannot be handled as effectively with currently available optimization techniques as can the problem with $\alpha = \infty$. The latter leads to a convex quadratic program, for which interior point methods can be devised that exploit its special structure, as we show in this paper. The former leads to a second-order cone program (see, for example, Lobo *et al.*, 1998). While software is now available for problems of this type (see, for example, Sturm, 1999), it is less able to take advantage of the structure of the problem and as a consequence will probably be less efficient in practice.

3 Numerical aspects of the estimator

We now return to (1.2). To avoid some cumbersome notation, we introduce the following matrix notation:

$$\underset{\sim}{y}_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix} \in \mathbb{R}^n, \quad j = 1, \ldots, k, \qquad \underset{\sim}{y} = \begin{pmatrix} \underset{\sim}{y}_1 \\ \vdots \\ \underset{\sim}{y}_k \end{pmatrix} \in \mathbb{R}^{nk},$$
$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}, \text{ and } \tilde{\mathbf{X}} = \mathbf{I}_k \otimes \mathbf{X} = \begin{pmatrix} \mathbf{X} & & \\ & \ddots & \\ & & \mathbf{X} \end{pmatrix} \in \mathbb{R}^{nk \times pk}, \tag{3.1}$$

where $\otimes$ denotes the Kronecker product. We arrange the regression parameters b_{lj} in a matrix to implicitly define vectors $\underset{\sim}{b}_1, \ldots, \underset{\sim}{b}_k \in \mathbb{R}^p$ and $\underset{\sim}{b}_{(1)}, \ldots, \underset{\sim}{b}_{(p)} \in \mathbb{R}^k$:

$$\begin{pmatrix} b_{11} & \ldots & b_{1k} \\ \vdots & & \vdots \\ b_{p1} & \ldots & b_{pk} \end{pmatrix} = \begin{pmatrix} | & & | \\ \underset{\sim}{b}_1 & \cdots & \underset{\sim}{b}_k \\ | & & | \end{pmatrix} = \begin{pmatrix} — \ \underset{\sim}{b}_{(1)}^{\mathsf{T}} \ — \\ \vdots \\ — \ \underset{\sim}{b}_{(p)}^{\mathsf{T}} \ — \end{pmatrix},$$

and then define the vector $\underset{\sim}{b}$ by

$$\underset{\sim}{b} = \begin{pmatrix} \underset{\sim}{b}_1 \\ \vdots \\ \underset{\sim}{b}_k \end{pmatrix} \in \mathbb{R}^{pk}.$$

Using this notation we write (1.2) as

$$\underset{\underset{\sim}{b} \in \mathbb{R}^{pk}}{\text{minimize}} \qquad f(\underset{\sim}{b}) = \tfrac{1}{2}\left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}\right)^{\mathsf{T}}\left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}\right) \tag{3.2a}$$

$$\text{subject to} \qquad g(\underset{\sim}{b}) = t - \sum_{l=1}^{p} \|\underset{\sim}{b}_{(l)}\|_\infty \geq 0. \tag{3.2b}$$

We also define the vector of residuals as follows:

$$r(\underset{\sim}{b}) = \underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}.$$

Remark 3.1*:* A referee pointed out that occasionally a covariance matrix for errors across responses typically is incorporated in multivariate regression problems. This amounts in changing the objective function in (3.2a) to $\frac{1}{2}\left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}\right)^{\mathsf{T}} \mathbf{W} \left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}\right)$ where $\mathbf{W}$ is a suitable (symmetric positive semidefinite) weight matrix. We can accommodate this generalization easily by pre-multiplying $\underset{\sim}{y}$ and $\tilde{\mathbf{X}}$ by $\mathbf{W}^{1/2}$. Since it is simple to incorporate a weight matrix $\mathbf{W}$ in our formulation, our method can be extended to other models in which the objective function is not the residual sum of squares but rather is obtained from an iteratively (re)weighted least squares procedure; for example, generalized linear models (McCullagh and Nelder, 1989).

Remark 3.2*:* As stated above, we generally assume that the explanatory variables and the response variables are centered and standardized to have (sample) mean zero and (sample) variance one. Given the previous remark about the possible incorporation of weights, such standardization may seem questionable. Obviously, just as for the LASSO and other statistical techniques that are in wide use, our proposed method is not invariant under rescaling of the variables (be they explanatory variables or response variables). Thus, we suggest that by default the variables should be centered and standardized unless the researcher has good reasons not to do so.

3.1 Characterization of solutions

We now use results from convex analysis (Rockafellar, 1970; Osborne, 1985; Clarke, 1990) to characterize solutions of (3.2). Introducing a Lagrange multiplier λ for the constraint (3.2b), we write the Lagrangian for (3.2) as follows:

$$\mathcal{L}(\underset{\sim}{b}, \lambda) = f(\underset{\sim}{b}) - \lambda g(\underset{\sim}{b}), \tag{3.3}$$

where $\lambda \geq 0$. If we fix $\lambda \geq 0$, then $\mathcal{L}(\underset{\sim}{b}, \lambda)$ is a convex function in $\underset{\sim}{b}$ and $\bar{\underset{\sim}{b}}$ minimizes $\mathcal{L}(\underset{\sim}{b}, \lambda)$ if, and only if, the pk-dimensional null-vector $\underset{\sim}{0}$ is an element of the subdifferential $\partial_{\underset{\sim}{b}}\mathcal{L}(\bar{\underset{\sim}{b}}, \lambda)$ (Osborne, 1985, p. 23). From (3.3), we have

$$\partial_{\underset{\sim}{b}}\mathcal{L}(\underset{\sim}{b}, \lambda) = -\tilde{\mathbf{X}}^{\mathsf{T}}\underset{\sim}{r} + \lambda\underset{\sim}{v},$$

where $\underset{\sim}{r} = r(\underset{\sim}{b}) = \underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}$ denotes the residual vector and $\underset{\sim}{v} = (v_1, \ldots, v_{pk})^{\mathsf{T}}$ has the following form:

- If $\|\underset{\sim}{b}_{(l)}\|_\infty > 0$, then $\underset{\sim}{v}_{(l)} = (v_{l1}, \ldots, v_{lk})^{\mathsf{T}}$ where $\sum_{j=1}^{k} |v_{lj}| = 1$ and, for $j = 1, \ldots, k$, we have $v_{lj} \geq 0$ if $b_{lj} = \|\underset{\sim}{b}_{(l)}\|_\infty$, $v_{lj} \leq 0$ if $b_{lj} = -\|\underset{\sim}{b}_{(l)}\|_\infty$ and $v_{lj} = 0$ if $|b_{lj}| \neq \|\underset{\sim}{b}_{(l)}\|_\infty$.
- If $\|\underset{\sim}{b}_{(l)}\|_\infty = 0$, then $\underset{\sim}{v}_{(l)} = (v_{l1}, \ldots, v_{lk})^{\mathsf{T}}$ where $\sum_{j=1}^{k} |v_{lj}| \leq 1$.

Thus if $\bar{\underset{\sim}{b}}$ minimizes $\mathcal{L}(\underset{\sim}{b}, \lambda)$ for a given value of λ,

$$\underset{\sim}{0} = -\tilde{\mathbf{X}}^{\mathsf{T}}\bar{\underset{\sim}{r}} + \lambda\bar{\underset{\sim}{v}}, \tag{3.4}$$

for some $\bar{\underset{\sim}{v}}$ of the form described above, and $\bar{\underset{\sim}{r}} = r(\bar{\underset{\sim}{b}}) = \underset{\sim}{y} - \tilde{\mathbf{X}}\bar{\underset{\sim}{b}}$. The properties of $\bar{\underset{\sim}{v}}$ imply that $\bar{\underset{\sim}{v}}^{\mathsf{T}}\bar{\underset{\sim}{b}} = \sum_{l=1}^{p} \|\bar{\underset{\sim}{b}}_{(l)}\|_\infty$, so it follows from (3.4) that

$$\lambda = \bar{\underset{\sim}{r}}^{\mathsf{T}}\tilde{\mathbf{X}}\bar{\underset{\sim}{b}} / \sum_{l=1}^{p} \|\bar{\underset{\sim}{b}}_{(l)}\|_\infty. \tag{3.5}$$

For $\bar{\underset{\sim}{b}}$ to be a solution of (3.2), we require not only that (3.4) holds for some vector $\bar{\underset{\sim}{v}}$ satisfying the properties above, but also that $\bar{\underset{\sim}{b}}$ satisfies the constraint (3.2b), that λ satisfying (3.5) has $\lambda \geq 0$, and that the following complementarity condition holds:

$$\lambda g(\underset{\sim}{b}) = \lambda \left(t - \sum_{l=1}^{p} \|\underset{\sim}{b}_{(l)}\|_\infty \right) = 0.$$

In the case of $\lambda = 0$, we have from (3.4) that $\tilde{\mathbf{X}}^\mathsf{T}\bar{\underset{\sim}{r}} = 0$, indicating that $\bar{\underset{\sim}{b}}$ is the unconstrained least-squares minimizer of (3.2a).

Although equation (3.4) gives a characterization of the solution for (3.2), the highly nonlinear way in which $\underset{\sim}{v}$ depends on $\underset{\sim}{b}$ makes it impossible to calculate the solution directly from the characterization just described. Some sort of iterative algorithm is needed. For general $\mathbf{X}$, an interior-point algorithm is developed in Section 3.3 below. The next section discusses the special case in which $\mathbf{X}$ is an orthonormal matrix.

3.2 The orthonormal design case

In this section we assume that $\mathbf{X}$ is orthonormal, and that $n > p$, so that $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_p$. For the LASSO, we can find explicit formulae for the LASSO estimate based on the unconstrained least squares estimate. Unfortunately, similar formulae do not seem to be available for the current problem. We can, however, develop a homotopy method, in which the constraint bound t becomes the homotopy parameter, and we can examine the behavior of the solution to (3.2) as t varies. A similar analysis is given by Osborne (1992) for the case of quantile regression, and by Osborne *et al.* (2000a) and Efron *et al.* (2004) for the LASSO. Specifically, the analysis shows that the solution $\underset{\sim}{b}$ of (1.2) is piecewise linear as a function of t and gives further insight into how our method selects variable simultaneously.

We start by noting that, since $\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} = \mathbf{I}_{pk}$, the unconstrained minimizer of (3.2a) is

$$\underset{\sim}{b}^0 = (b_1^0, \ldots, b_{pk}^0)^\mathsf{T} = \tilde{\mathbf{X}}^\mathsf{T}\underset{\sim}{y}.$$

Assuming that the b_{lj}s are, for some nonnegative quantities ρ_l, $l = 1, \ldots, p$, of the form

$$b_{lj} = \operatorname{sign}(b_{lj}^0) \times \min(|b_{lj}^0|, \rho_l), \qquad l = 1, \ldots, p, \quad j = 1, \ldots, k, \tag{3.6}$$

we now show, by specifying the dependence of these ρ_ls on t, that (3.6) indeed yields the solution of (3.2). Observe that, as long as $\rho_l \leq \|\underset{\sim}{b}_{(l)}^0\|_\infty$, for $l = 1, \ldots, p$, we have $\rho_l = \|\underset{\sim}{b}_{(l)}\|_\infty$. By using $\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} = \mathbf{I}_{pk}$ again, we rewrite $f(\underset{\sim}{b})$ as follows:

$$f(\underset{\sim}{b}) = \tfrac{1}{2}\left\{ \left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}^0\right)^\mathsf{T} \left(\underset{\sim}{y} - \tilde{\mathbf{X}}\underset{\sim}{b}^0\right) + \left(\underset{\sim}{b}^0 - \underset{\sim}{b}\right)^\mathsf{T} \left(\underset{\sim}{b}^0 - \underset{\sim}{b}\right) \right\}. \tag{3.7}$$

From (3.6), we have that $|b_{lj}^0 - b_{lj}| = (|b_{lj}^0| - \rho_l)_+$ where $(x)_+ = \max(0, x)$. By using this observation in conjunction with (3.7), we reformulate (3.2) as follows:

$$\underset{\rho_1,\ldots,\rho_l}{\text{minimize}} \qquad \tfrac{1}{2}\sum_{j=1}^{k}\sum_{l=1}^{p}(|b_{lj}^0| - \rho_l)_+^2 \tag{3.8a}$$

$$\text{subject to} \qquad \sum_{l=1}^{p}\rho_l = t. \tag{3.8b}$$

We now define $\sigma \subseteq \{1, \ldots, p\}$ such that if $l \notin \sigma$ then $\rho_l = 0$, that is, σ is the set of indices l for which ρ_l may be different from zero. Furthermore, for each $l = 1, \ldots, p$, let $\sigma_l \subseteq \{1, \ldots, k\}$ be the set of indices j such that $|b^0_{lj}| > \rho_l$ if, and only if, $j \in \sigma_l$. We then rewrite (3.8) as follows:

$$\begin{array}{ll} \underset{\rho_1,\cdots,\rho_l}{\text{minimize}} & \frac{1}{2}\sum_{l\in\sigma}\sum_{j\in\sigma_l}(|b^0_{lj}| - \rho_l)^2, \\ \text{subject to} & \sum_{l\in\sigma}\rho_l = t. \end{array}$$

Of course, we also require that $\rho_l \geq 0$ for $l \in \sigma$. By introducing a Lagrange multiplier μ for the constraint in this problem, we obtain from the optimality conditions that the solution must satisfy the following relations:

$$\mu = \sum_{j\in\sigma_l}(|b^0_{lj}| - \rho_l) = \left(\sum_{j\in\sigma_l}|b^0_{lj}|\right) - n_l\rho_l, \qquad l \in \sigma \text{ and } \rho_l > 0 \tag{3.9a}$$

$$t = \sum_{l\in\sigma}\rho_l, \tag{3.9b}$$

where $n_l = |\sigma_l|$ denotes the number of elements in the set σ_l, $l = 1, \ldots, p$.

We shall now use the Karush–Kuhn–Tucker (KKT) conditions (3.9) to show that the ρ_ls (and hence, by (3.6), the b_{lj}s) are piecewise linear functions of t. Specifically, we show that there is a sequence of "knots" $0 = t_0 < t_1 < t_2 < \cdots < t_m$ such that the ρ_ls are linear in t for $t \in [t_{i-1}, t_i]$ for $i = 1, 2, \ldots, m$. At each of the t_is either σ changes by adding one or more indices, one or more σ_ls change by dropping one or more indices, or both these events happen.

For $t_0 = 0$, we set $\sigma = \{l : \|\underset{\sim}{b}^0_{(l)}\|_1 = \max_{l=1,\ldots,p}\|\underset{\sim}{b}^0_{(l)}\|_1\}$ and $\rho_l(t_0) = 0$ for $l = 1, \ldots, p$. Clearly, this is the optimal solution for $t = 0$ and fulfills the KKT conditions (3.9). We now describe the homotopy algorithm iteratively as follows:

Assume that we are at t_i, and then define for each $l \in \sigma$

$$\rho_l(t) = \rho_l(t_i) + \frac{1}{\sum_{l\in\sigma}1/n_l}\frac{1}{n_l}(t - t_i).$$

Note that each ρ_l is a linear function of t and that if $\sum_{l\in\sigma}\rho_l(t_i) = t_i$, then $\sum_{l\in\sigma}\rho_l(t) = t$ for all $t > t_i$, provided the set σ does not change. Hence, these ρ_ls fulfill the KKT condition (3.9b) for t between t_i and t_{i+1}.

Furthermore, for each $l \in \sigma$, we define

$$\mu_l(t) = \left(\sum_{j\in\sigma_l}|b^0_{lj}|\right) - n_l\rho_l(t).$$

Because of the definition of the $\rho_l(t)$s, this definition ensures that the optimality condition (3.9a) holds for $t > t_i$ whenever it holds at t_i, as long as none of the sets σ_l change. That is, $\mu_{l_1}(t) = \mu_{l_2}(t) = \mu(t)$ for any $l_1, l_2 \in \sigma$.

We conclude that the $\rho_l(t)$ defined above are the solutions to (3.8) for all t between t_i and t_{i+1}. It remains to determine the next knot t_{i+1}. To find t_{i+1}, we calculate τ^*_l such that $\rho_l(\tau^*_l) = \min_{j\in\sigma_l}(|b^0_{lj}|)$, for each $l \in \sigma$. Let $\tau^* = \min_{l\in\sigma}\tau^*_l$ be the constraint bound t at which one or more of the σ_ls change by dropping one or more indices.

Furthermore, for some $l_0 \in \sigma$ let $\mu(t) = \mu_{l_0}(t)$. (As noted above, the $\mu_l(t)$, $l \in \sigma$ are all identical.) Then, provided that not all variables have yet entered the model (that is, $\sigma \neq \{1, \dots, p\}$), we calculate $\tau^\dagger$ such that $\mu(\tau^\dagger) = \max_{l \notin \sigma} \|\utilde{b}^0_{(l)}\|_1$. In other words, $\tau^\dagger$ is the constraint bound t at which σ changes by adding one or more indices. In the alternative case of $\sigma = \{1, \dots, p\}$, we calculate $\tau^\dagger$ such that $\mu(\tau^\dagger) = 0$. That is, $\tau^\dagger$ is the constraint bound at which we reach the unconstrained solution.

We then set $t_{i+1} = \min(\tau^*, \tau^\dagger) > t_i$. If $t_{i+1} = \tau^\dagger$ and $\mu(t_{i+1}) \neq 0$, we update σ to $\sigma = \sigma \cup \{l : \|\utilde{b}^0_{(l)}\|_1 = \mu(t_{i+1})\}$. If $\mu(t_{i+1}) = 0$, then we have reached the unconstrained solution and the algorithm stops; otherwise it continues as described above.

We conclude from this analysis that for $\mathbf{X}$ orthonormal, the solution vector $\utilde{b}$ of (1.2) is a (continuous) piecewise linear function of the constraint parameter t. This property of the solution vector $\utilde{b}$ also holds for general $\mathbf{X}$, if $\mathbf{X}$ has full column rank. The proof of the more general result is omitted but is available on request from the authors. We conjecture that the same statement holds for general $\mathbf{X}$, as is the case for the LASSO (Osborne *et al.*, 2000a; Efron *et al.*, 2004). However, proof of this conjecture for general $\mathbf{X}$ in the current setting appears to be a somewhat more difficult prospect.

This analysis also gives some insight into how our method selects variables. In the case of an orthonormal design it essentially orders the variables such that

$$\|\utilde{b}^0_{(l_1)}\|_1 \geq \|\utilde{b}^0_{(l_2)}\|_1 \geq \|\utilde{b}^0_{(l_3)}\|_1 \geq \cdots \geq \|\utilde{b}^0_{(l_{p-1})}\|_1 \geq \|\utilde{b}^0_{(l_p)}\|_1,$$

and then selects the variables $x_{il_1}, x_{il_2}, \dots, x_{il_m}$, where m depends on t, using this ordering. Note that the unconstrained coefficient estimates $\utilde{b}^0_{(l)}$ are sorted according to their L^1 norm. This shows that the constraint that we propose achieves its "simultaneous" variable selection by measuring the over all contribution of an explanatory variable by summing its (absolute) contribution over all of the k regressions. The variable that is best with respect to this measure is selected first, followed by the variable that is second best with respect to this measure, and so on.

3.3 The general case

In this section we develop an interior point algorithm for solving (1.2) for general $\tilde{\mathbf{X}}$. First, to express this problem as a convex quadratic program, we define

$$\mathbf{Q} = \tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}, \qquad c = -\tilde{\mathbf{X}}^\mathsf{T}\utilde{y} \in \mathbb{R}^{pk}, \qquad d = \tfrac{1}{2}\utilde{y}^\mathsf{T}\utilde{y},$$

and use $\utilde{u}_l$ to denote an l-dimensional vector with all entries equal to one. By introducing an auxiliary vector $\utilde{z} \in \mathbb{R}^p$, we now write (3.2) as follows:

$$\begin{aligned} &\underset{\utilde{b}}{\text{minimize}} && \tfrac{1}{2}\utilde{b}^\mathsf{T}\mathbf{Q}\utilde{b} + \utilde{c}^\mathsf{T}\utilde{b} + d && \text{(3.10a)} \\ &\text{subject to} && \utilde{u}_k \otimes \utilde{z} - \utilde{b} \geq \utilde{0} && \text{(3.10b)} \\ & && \utilde{u}_k \otimes \utilde{z} + \utilde{b} \geq \utilde{0} && \text{(3.10c)} \\ & && t - \utilde{u}_p^\mathsf{T}\utilde{z} \geq 0. && \text{(3.10d)} \end{aligned}$$

It is well known that convex quadratic programming problems can be solved efficiently using primal-dual infeasible-interior-point algorithms (Roos *et al.*, 1997; Wright, 1997; Ye, 1997). We now present a brief derivation of the interior-point approach, as applied to our specific problem (3.10).

Using the Lagrange multipliers $\utilde{\lambda}_l$, $\utilde{\lambda}_u \in \mathbb{R}^{kp}$ and $\tau \in \mathbb{R}$, the Lagrangian for (3.10) is

$$\mathcal{L}(\utilde{b}, \utilde{z}, \utilde{\lambda}_u, \utilde{\lambda}_l, \tau) = \tfrac{1}{2}\utilde{b}^{\mathsf{T}}\mathbf{Q}\utilde{b} + \utilde{c}^{\mathsf{T}}\utilde{b} + d - \utilde{\lambda}_u^{\mathsf{T}}(\utilde{u}_k \otimes \utilde{z} - \utilde{b}) - \utilde{\lambda}_l^{\mathsf{T}}(\utilde{u}_k \otimes \utilde{z} + \utilde{b}) - \tau(t - \utilde{u}_p^{\mathsf{T}}\utilde{z}).$$

The optimality conditions for $\utilde{b}$ to solve (3.10) are

$$\begin{aligned}
\mathbf{Q}\utilde{b} + \utilde{c} + \utilde{\lambda}_u - \utilde{\lambda}_l &= \utilde{0}, \\
-(\utilde{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\utilde{\lambda}_u - (\utilde{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\utilde{\lambda}_l + \tau\utilde{u}_p &= \utilde{0}, \\
\utilde{u}_k \otimes \utilde{z} - \utilde{b} &\geq \utilde{0}, \\
\utilde{u}_k \otimes \utilde{z} + \utilde{b} &\geq \utilde{0}, \\
t - \utilde{u}_p^{\mathsf{T}}\utilde{z} &\geq 0, \\
\utilde{\lambda}_u \geq \utilde{0}, \quad \utilde{\lambda}_l \geq \utilde{0}, \quad \tau &\geq 0, \\
\utilde{\lambda}_u^{\mathsf{T}}(\utilde{u}_k \otimes \utilde{z} - \utilde{b}) = \utilde{0}, \quad \utilde{\lambda}_l^{\mathsf{T}}(\utilde{u}_k \otimes \utilde{z} + \utilde{b}) = \utilde{0}, \quad \tau(t - \utilde{u}_p^{\mathsf{T}}\utilde{z}) &= 0.
\end{aligned}$$

Since the problem is a convex quadratic program, these conditions are sufficient as well as necessary. By introducing slack variables $\utilde{s}_u$ and $\utilde{s}_l$ in $\mathbb{R}^{kp}$, and $\zeta \in \mathbb{R}$, we restate these conditions in a form that is more convenient for development of the interior point approach:

$$\begin{aligned}
\mathbf{Q}\utilde{b} + \utilde{c} + \utilde{\lambda}_u - \utilde{\lambda}_l &= \utilde{0}, && \text{(3.11a)} \\
-(\utilde{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\utilde{\lambda}_u - (\utilde{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\utilde{\lambda}_l + \tau\utilde{u}_p &= \utilde{0}, && \text{(3.11b)} \\
\utilde{u}_k \otimes \utilde{z} - \utilde{b} - \utilde{s}_u &= \utilde{0}, && \text{(3.11c)} \\
\utilde{u}_k \otimes \utilde{z} + \utilde{b} - \utilde{s}_l &= \utilde{0}, && \text{(3.11d)} \\
t - \utilde{u}_p^{\mathsf{T}}\utilde{z} - \zeta &= 0, && \text{(3.11e)} \\
\mathbf{\Lambda}_u\mathbf{S}_u\utilde{u}_{pk} = \utilde{0}, \quad \mathbf{\Lambda}_l\mathbf{S}_l\utilde{u}_{pk} = \utilde{0}, \quad \tau\zeta &= 0, && \text{(3.11f)} \\
\utilde{\lambda}_u \geq \utilde{0}, \quad \utilde{\lambda}_l \geq \utilde{0}, \quad \tau \geq 0, \quad \utilde{s}_u \geq \utilde{0}, \quad \utilde{s}_l \geq \utilde{0}, \quad \zeta &\geq 0. && \text{(3.11g)}
\end{aligned}$$

(Here we use a standard notational convention from the interior point literature, namely, that if a lowercase and an uppercase letter are used at the same time, then the lowercase letter indicates a vector and the uppercase letter a diagonal matrix whose diagonal elements are the elements of the corresponding vector.) Primal-dual interior-point methods view (3.11) as a constrained system of nonlinear equations. They seek a root of the function F defined by the equality conditions in (3.11), that is,

$$F(\underset{\sim}{b}, \underset{\sim}{z}, \underset{\sim}{\lambda}_u, \underset{\sim}{\lambda}_l, \tau, \underset{\sim}{s}_u, \underset{\sim}{s}_l, \zeta) = \begin{pmatrix} \mathbf{Q}\underset{\sim}{b} + \underset{\sim}{c} + \underset{\sim}{\lambda}_u - \underset{\sim}{\lambda}_l \\ -(\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\underset{\sim}{\lambda}_u - (\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\underset{\sim}{\lambda}_l + \tau \underset{\sim}{u}_p \\ \underset{\sim}{u}_k \otimes \underset{\sim}{z} - \underset{\sim}{b} - \underset{\sim}{s}_u \\ \underset{\sim}{u}_k \otimes \underset{\sim}{z} + \underset{\sim}{b} - \underset{\sim}{s}_l \\ t - \underset{\sim}{u}_p^{\mathsf{T}} \underset{\sim}{z} - \zeta \\ \mathbf{\Lambda}_u \mathbf{S}_u \underset{\sim}{u}_{pk} \\ \mathbf{\Lambda}_l \mathbf{S}_l \underset{\sim}{u}_{pk} \\ \tau\zeta \end{pmatrix} = 0, \tag{3.12}$$

over the set defined by the inequalities listed in (3.11g). An important concept in primal-dual methods is the *central path*, which is defined as the solution of the following perturbed variant of (3.12), for some parameter $\mu > 0$,

$$F(\underset{\sim}{b}, \underset{\sim}{z}, \underset{\sim}{\lambda}_u, \underset{\sim}{\lambda}_l, \tau, \underset{\sim}{s}_u, \underset{\sim}{s}_l, \zeta) = \begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \\ 0 \\ \mu \underset{\sim}{u}_{pk} \\ \mu \underset{\sim}{u}_{pk} \\ \mu \end{pmatrix}, \tag{3.13}$$

over the strict interior of the feasible region defined by (3.11g), that is,

$$\underset{\sim}{\lambda}_u > \underset{\sim}{0}, \quad \underset{\sim}{\lambda}_l > \underset{\sim}{0}, \quad \tau > 0, \quad \underset{\sim}{s}_u > \underset{\sim}{0}, \quad \underset{\sim}{s}_l > \underset{\sim}{0}, \quad \zeta > 0. \tag{3.14}$$

Maintenance of strict positivity of these variables at each iteration is the origin of the term "interior-point."

Interior-point methods of the path-following type (such as the one we use here) find the solution of (3.11g), (3.12) by following the central path (3.13), (3.14) as μ decreases to zero. Rather than calculate the central path point exactly for each value of μ, path-following methods take a single Newton-like step toward a point on the central path that is, in a sense, closer to the solution than the current iterate. We define the central path point corresponding to the current iterate by defining μ as follows:

$$\mu = \frac{\underset{\sim}{\lambda}_l^{\mathsf{T}} \underset{\sim}{s}_l + \underset{\sim}{\lambda}_u^{\mathsf{T}} \underset{\sim}{s}_u + \tau\zeta}{2pk + 1}. \tag{3.15}$$

(Note that this μ is the average value of the pairwise products $\lambda_{u,i} s_{u,i}$ and $\lambda_{l,i} s_{l,i}$, $i = 1, 2, \ldots, pk$, and $\tau\zeta$.) We then choose a *centering parameter* $\sigma \in (0, 1)$, and apply a modified Newton step toward the central path point defined by (3.13), (3.14) in which μ is replaced by $\sigma\mu$. The modification (due to Mehrotra, 1992, and detailed below) enhances the approximation of (3.12) on which the Newton step is based, making it approach a second-order approximation rather than the usual first-order (linear) approximation. By requiring σ to be strictly less than 1, we ensure that the step aims at a point further along the central path than the point corresponding to the current iterate. A heuristic for choosing σ is also described in Mehrotra (1992) and detailed below. Practical variants of this algorithm may contain other important features, such as techniques for determining the distance to move along the calculated step, a method for calculating the starting point, and possibly third- and higher-order modifications to the search direction.

We now focus on the system of equations obtained from the modified Newton step for (3.13). By defining the residuals at the current point from (3.12), with the appropriate adjustments for the central path perturbation (the terms involving σ) and for higher-order enhancement (the terms $\hat{\underset{\sim}{r}}_{hi}$, $\hat{\underset{\sim}{r}}_{lo}$, and $\hat{r}_{\tau\zeta}$), we obtain

$$\begin{pmatrix} \underset{\sim}{r}_b \\ \underset{\sim}{r}_z \\ \underset{\sim}{r}_u \\ \underset{\sim}{r}_l \\ r_t \\ \underset{\sim}{r}_{hi} \\ \underset{\sim}{r}_{lo} \\ r_{\tau\zeta} \end{pmatrix} := \begin{pmatrix} \mathbf{Q}\underset{\sim}{b} + \underset{\sim}{c} + \underset{\sim}{\lambda}_u - \underset{\sim}{\lambda}_l \\ -(\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\underset{\sim}{\lambda}_u - (\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p)\underset{\sim}{\lambda}_l + \tau \underset{\sim}{u}_p \\ \underset{\sim}{u}_k \otimes \underset{\sim}{z} - \underset{\sim}{b} - \underset{\sim}{s}_u \\ \underset{\sim}{u}_k \otimes \underset{\sim}{z} + \underset{\sim}{b} - \underset{\sim}{s}_l \\ t - \underset{\sim}{u}_p^{\mathsf{T}}\underset{\sim}{z} - \zeta \\ \mathbf{\Lambda}_u \mathbf{S}_u \underset{\sim}{u}_{pk} - \sigma\mu\underset{\sim}{u}_{pk} + \hat{\underset{\sim}{r}}_{hi} \\ \mathbf{\Lambda}_l \mathbf{S}_l \underset{\sim}{u}_{pk} - \sigma\mu\underset{\sim}{u}_{pk} + \hat{\underset{\sim}{r}}_{lo} \\ \tau\zeta - \sigma\mu + \hat{r}_{\tau\zeta} \end{pmatrix}. \tag{3.16}$$

The modified Newton system is then

$$\begin{pmatrix} \mathbf{Q} & & \mathbf{I}_{pk} & -\mathbf{I}_{pk} & & & & \\ & & -\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p & -\underset{\sim}{u}_k^{\mathsf{T}} \otimes \mathbf{I}_p & \underset{\sim}{u}_p & & & \\ -\mathbf{I}_{pk} & \underset{\sim}{u}_k \otimes \mathbf{I}_p & & & & -\mathbf{I}_{pk} & & \\ \mathbf{I}_{pk} & \underset{\sim}{u}_k \otimes \mathbf{I}_p & & & & & -\mathbf{I}_{pk} & \\ & -\underset{\sim}{u}_p^{\mathsf{T}} & & & & & & -1 \\ & & \mathbf{S}_u & & & \mathbf{\Lambda}_u & & \\ & & & \mathbf{S}_l & & & \mathbf{\Lambda}_l & \\ & & & & \zeta & & & \tau \end{pmatrix} \begin{pmatrix} \Delta\underset{\sim}{b} \\ \Delta\underset{\sim}{z} \\ \Delta\underset{\sim}{\lambda}_u \\ \Delta\underset{\sim}{\lambda}_l \\ \Delta\tau \\ \Delta\underset{\sim}{s}_u \\ \Delta\underset{\sim}{s}_l \\ \Delta\zeta \end{pmatrix} = - \begin{pmatrix} \underset{\sim}{r}_b \\ \underset{\sim}{r}_z \\ \underset{\sim}{r}_u \\ \underset{\sim}{r}_l \\ r_t \\ \underset{\sim}{r}_{hi} \\ \underset{\sim}{r}_{lo} \\ r_{\tau\zeta} \end{pmatrix}. \tag{3.17}$$

Thus, at each iteration of our interior point algorithm we have to solve systems of equations of the form (3.17). Although this system of equations is very large, it is also highly structured, and by performing some block elimination we can reduce its dimension greatly. Details of these algebraic manipulations are given in the technical report, available from the authors on request, on which this paper is based.

We continue with some details of our implementation of the Mehrotra algorithm, which actually solves two systems of the form (3.17) at each iteration, with the same coefficient matrices but different right-hand sides (3.16). In the first of these systems, we set $\sigma = 0$ and $\hat{\underset{\sim}{r}}_{hi} = 0$, $\hat{\underset{\sim}{r}}_{lo} = 0$, and $\hat{r}_{\tau\zeta} = 0$, to obtain the *affine-scaling direction.* This direction, which we denote by

$$(\Delta\underset{\sim}{b}^{\text{aff}}, \Delta\underset{\sim}{z}^{\text{aff}}, \Delta\underset{\sim}{\lambda}_u^{\text{aff}}, \Delta\underset{\sim}{\lambda}_l^{\text{aff}}, \Delta\tau^{\text{aff}}, \Delta\underset{\sim}{s}_u^{\text{aff}}, \Delta\underset{\sim}{s}_l^{\text{aff}}, \Delta\zeta^{\text{aff}}) \tag{3.18}$$

is simply the pure Newton direction for the system of equations F given in (3.12). We then find the largest step length $\alpha_{\text{aff}} \in (0, 1]$ such that a step of length α_{aff} along this direction from the current iterate satisfies the conditions (3.11g). We then calculate the value μ_{aff} from (3.15) that would occur if we actually took this step, and set

$$\sigma = \left(\frac{\mu_{\text{aff}}}{\mu}\right)^3, \tag{3.19}$$

where μ is calculated using the current iterate. We use this value of σ to form the right-hand side for the second system, and also use the components of (3.18) to define the residual modifications:

$$\hat{\underset{\sim}{r}}_{hi} = \Delta\mathbf{\Lambda}_u^{\text{aff}}\Delta\mathbf{S}_u^{\text{aff}}\underset{\sim}{u}_{pk}, \qquad \hat{\underset{\sim}{r}}_{lo} = \Delta\mathbf{\Lambda}_l^{\text{aff}}\Delta\mathbf{S}_l^{\text{aff}}\underset{\sim}{u}_{pk}, \qquad \hat{r}_{\tau\zeta} = \Delta\tau^{\text{aff}}\Delta\zeta^{\text{aff}}. \tag{3.20}$$

We then solve (3.17) with the new right-hand side to obtain the actual search direction. The heuristic for σ in (3.19) yields a value close to 0 when the pure Newton direction appears to be a profitable search direction. Thus the calculated step will not be much different from the pure Newton direction, and will move quite aggressively to reduce the value of μ on this iteration. When the affine-scaling direction does not make much progress in reducing μ, the heuristic yields a conservative choice of σ, closer to 1.

The step length along the search direction is chosen by means of a heuristic due to Mehrotra (1992, Section 6) and described in Wright (1997, p. 205). The heuristic is modified in an obvious way to account for the fact that our objective function is quadratic rather than linear. This choice of step ensures that the *strict* inequalities (3.14) are satisfied by the new iterate.

We terminate the algorithm when μ and the residuals in (3.16) become sufficiently small. (In our code, we apply the simple test $\mu < 10^{-8}$.) We obtain a starting point by simply setting $\underset{\sim}{z} = 0$ and $\underset{\sim}{b} = 0$, while the components of $\underset{\sim}{\lambda}_u$, $\underset{\sim}{\lambda}_l$, $\underset{\sim}{s}_u$, $\underset{\sim}{s}_l$, τ, and ζ are all set to some value, in our code 10^5.

An outline of the overall algorithm is as follows:

1. Choose a starting point.

2. Calculate μ from (3.15). If μ is small enough, STOP.

3. Calculate the affine-scaling direction (3.18) by solving the system (3.16), (3.17) with $\sigma = 0$ and zero residual modifications.

4. Use the affine-scaling direction to calculate σ according to (3.19) and the residual modifications according to (3.20).

5. Solve (3.16), (3.17) with the new RHS to obtain the actual step.

6. Calculate the step length and take the step.

7. Return to 2 and iterate.

When the algorithm terminates, the final iterate is usually close to the solution of (1.2), but has all its components non-zero. We use a heuristic to determine which of these components represent indices of the variables that should be in the model. Specifically, we set

$$\mathcal{I} = \{l : \|\underset{\sim}{b}_{(l)}\|_\infty > t\,10^{-4},\ l = 1, \ldots, p\}$$

	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	1.0
X81							—	—	—	—	—	—
X82						—	—	—	—	—	—	
X84	—	—	—	—	—	—	—	—	—			
X85		—	—	—	—	—	—	—	—	—	—	—
X86								—	—	—	—	—
X112								—	—	—	—	—
X113				—	—	—	—	—	—	—	—	—
X114	—	—	—	—	—	—	—	—	—	—	—	—
X118											—	—
X220												—
X622		—	—	—	—	—	—	—	—	—	—	
X623										—	—	—
X662					—	—						
X667								—	—	—	—	—
X672							—	—	—			
X673							—	—	—	—	—	—
X691												—

Table 4.1: Selected variables using the complete data set

and all b_{lj} with $l \notin \mathcal{I}$ are set to zero. For the case $k = 1$, there is some evidence that this heuristic is too liberal, in the sense that coefficients that are zero at the exact solution are some distance from zero at the final interior-point iterate. However, this occurs only for some values of t, namely some of those at which variables enter or drop out of the model. This behavior is, thus, not of great concern since, as argued earlier, we regard the methodology as exploratory only.

4 The infra–red spectrometry data revisited

We implemented the algorithm described in Section 3.3 in C and applied it to the infra–red spectrometry data discussed in Section 1. Our hope is that, by using all response variables simultaneously to select a single set of regressor variables, we can avoid problems of overfitting and high variability.

As remarked earlier, Figure 1.2 indicates that one observation in NLeco, namely the $10th$, is suspicious and possibly an outlier. Hence, we ran our analysis twice, once using all observations and once with the $10th$ observation removed from *all* the variables. In this way we also hoped to get some insight into the robustness of the proposed methodology with respect to outliers.

We used several values for the tuning parameter t, namely $t = 0.1$, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9 and 1.0. The results are summarized in Table 4.1 for the complete data set and in Table 4.2 for the data set with the $10th$ observation missing. Each table lists all chosen regressor variables for various values of t. The horizontal lines link the values of t for which each coefficient remains non-zero in the solution.

Note that the tables are quite similar, showing essentially the same group of variables selected over the range of t values. Only X691 from Table 4.1 is replaced by X690 in Table 4.2 and X87 is added. Thus, at least for this extreme example, the method appears to be fairly

	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	1.0
X81							■	■	■	■	■	■
X82								■	■	■		
X84	■	■	■	■	■	■	■					
X85		■	■	■	■	■	■	■	■	■	■	■
X86										■	■	■
X87												■
X112								■	■	■	■	■
X113		■	■	■	■	■	■	■	■	■	■	■
X114	■	■	■	■	■	■	■	■	■	■	■	■
X118											■	■
X220												■
X622			■	■	■	■	■	■	■	■	■	■
X623												■
X662					■	■	■					
X667								■	■	■	■	■
X672							■	■	■	■		
X673							■	■	■	■	■	■
X690												■

Table 4.2: Selected variables without the $10th$ observation

robust with respect to the outlier in the $10th$ observation of NLeco.

Although there still appears to be some variation, and several variables are identified as being non-zero for only a few values of t, the method consistently picks regressor variables from only three separate ranges of the spectrum. Roughly speaking, these ranges are the $81st$–$87th$, the $112th$–$118th$ and the $622nd$–$691st$ frequencies. Within each of these ranges the regressor variables are highly correlated. The minimum correlation in the group X81, X82, X84, X85, X86, and X87 is larger than 0.9989, while the minimum correlation in the group X112, X113, X114, and X118 is larger than 0.9993. For the last group X622, X623, X662, X667, X672, X673, X690, and X691, the minimum correlation is larger than 0.9431. Since this group spans a wider range of frequencies, it is not surprising that its correlation is slightly smaller than the others.

Given the high correlations, there are essentially two ways how one could proceed. Either one selects one regressor variable from each group or one averages over the variables in each group. Here, for illustrative purposes we use the first method and choose those variables that are selected for most values of t, i.e. X85, X114 and X622. Both Table 4.1 and Table 4.2 suggest using this set of explanatory variables.

To investigate how well this selection of variables performs, we fit linear regression models to each of the response variables using these three regressor variables. The resulting fits are shown in Figure 4.1, for the case in which all observations are used; and in Figure 4.4 for the case in which the $10th$ observation is removed. Figures 4.2 and 4.5 show the corresponding plots of the jackknifed residuals, while Figures 4.3 and 4.6 show the normal quantile plots based on these jackknifed residuals.

From these figures it seems that a linear regression model using these three predictors is satisfactory, at least in most cases. Of course, in the figures that are produced from the complete data set, the outlier in NLeco is clearly visible. The residual plot for Na shows a

	All observations				Without 10th observation			
	Intercept	X85	X114	X622	Intercept	X85	X114	X622
EC	***	*			***	*		
pH	***	***	*	**	***	***	*	*
pHCaCl2	***	***	**		***	***	**	
CLeco	***	***	**	*	***	***	**	*
Org.C	***	***		*	***	**		*
NLeco	***				***	***		*
extP	***				***			
Ca	***	***		***	***	**		***
Mg	***		***	***	***		***	***
Na	***		*	*	***		*	*
K	***				***			
TotalCations	***	***		***	***	**		***
CEC	***	***	***	***	***	***	***	***
CaCO3	***	**	***		***	**	***	

Table 4.3: Significance of selected variables in each model

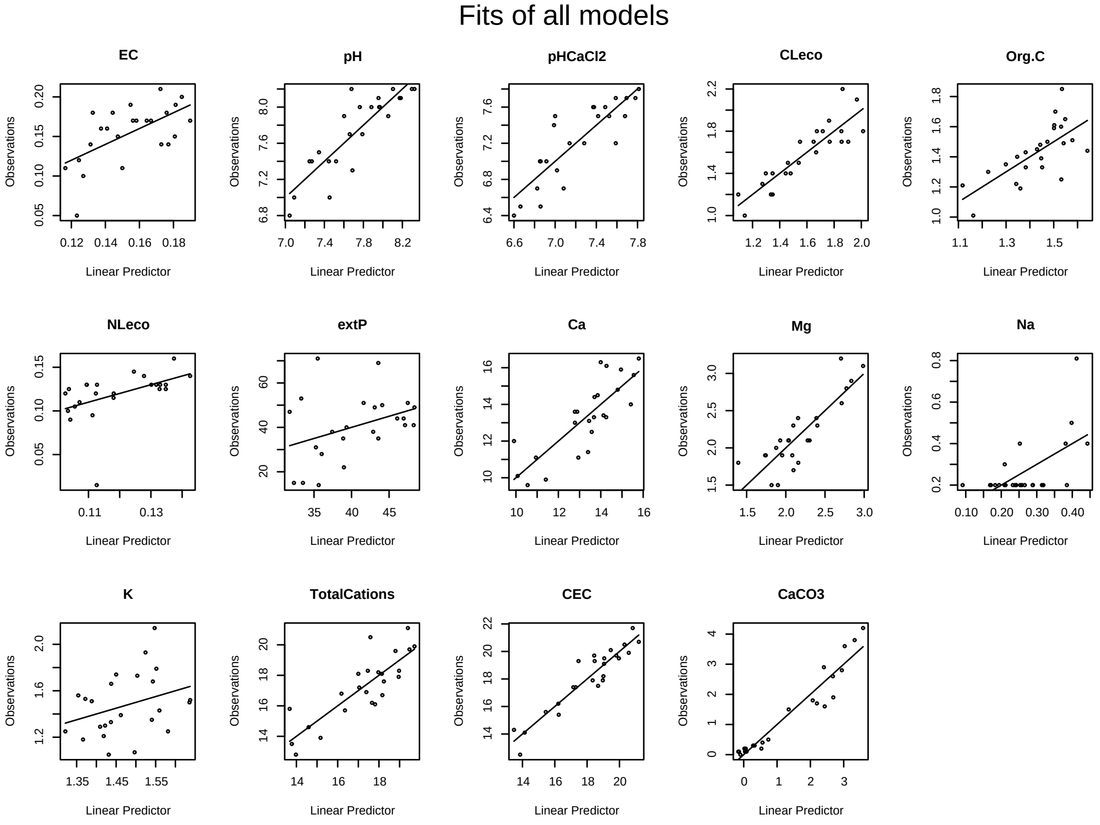

Figure 4.1: Linear fits using 3 explanatory variables (all observations)

lot of structure, but this is due to the granularity of this response variable. Modelling of Na clearly is a difficult task, since this response variable takes only five distinct values, with eighteen replications of the smallest value and three replications of the median.

The results of significance tests for each parameter in each of the linear models is summarized in Table 4.3. An entry of "***" in this table means that in the linear model for

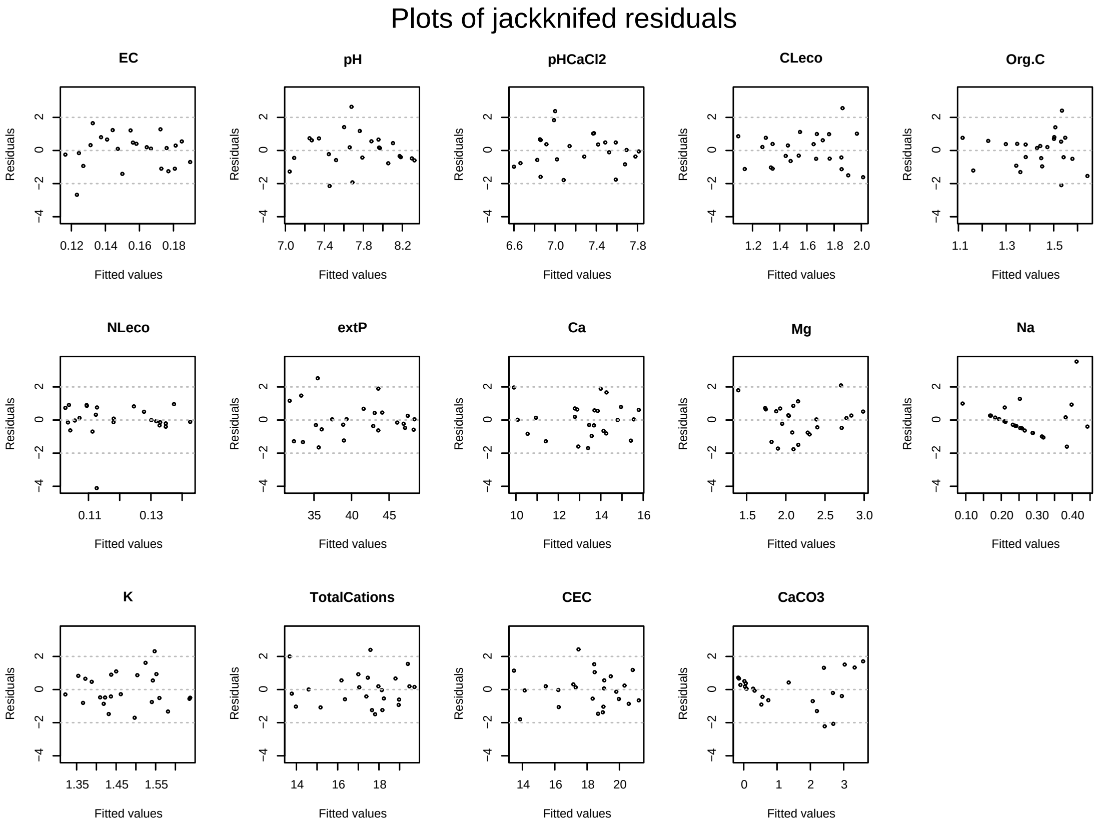

Figure 4.2: Residual plots (all observations)

the response variable (given in the left-most column), the p-value for the t-statistics of the parameter estimate for the regressor variable (given in the top row) was below 0.1%. If the p-values is between 0.1% and 1%, the entry is "**" and a "*" denotes that the p-value was between 1% and 5%.

The results in Table 4.3 show that, if all observations are used, then none of the three regressor is significant for NLeco. This is hardly surprising, as the outlying observation grossly inflates the residual sum of squares. Table 4.3 also shows that, except for EC, extP, and K, at least two of the three regressor variables that we have chosen are significant for each response variable. In the case of K, this observation is not surprising, given the results of the LASSO summarized in Figure 1.1. When the LASSO method is applied to each response variable separately, the regressor variables chosen for K, with $t = 1$ are essentially X54 and X220. For EC, the selected regressor variables with $t = 1$ are X87, X108, and X694; whereas for extP the LASSO selects X1, X118, X345, and X732. It is surprising that X114 is not significant (if used together with X85 and X622) for either EC or extP in Table 4.3.

For all other response variables, Figure 1.1 shows that for $t = 1$, the LASSO selects regressor variables that are close to the set X85, X114, and X662. However, from these individual results it is much harder to select a single set of regressor variables.

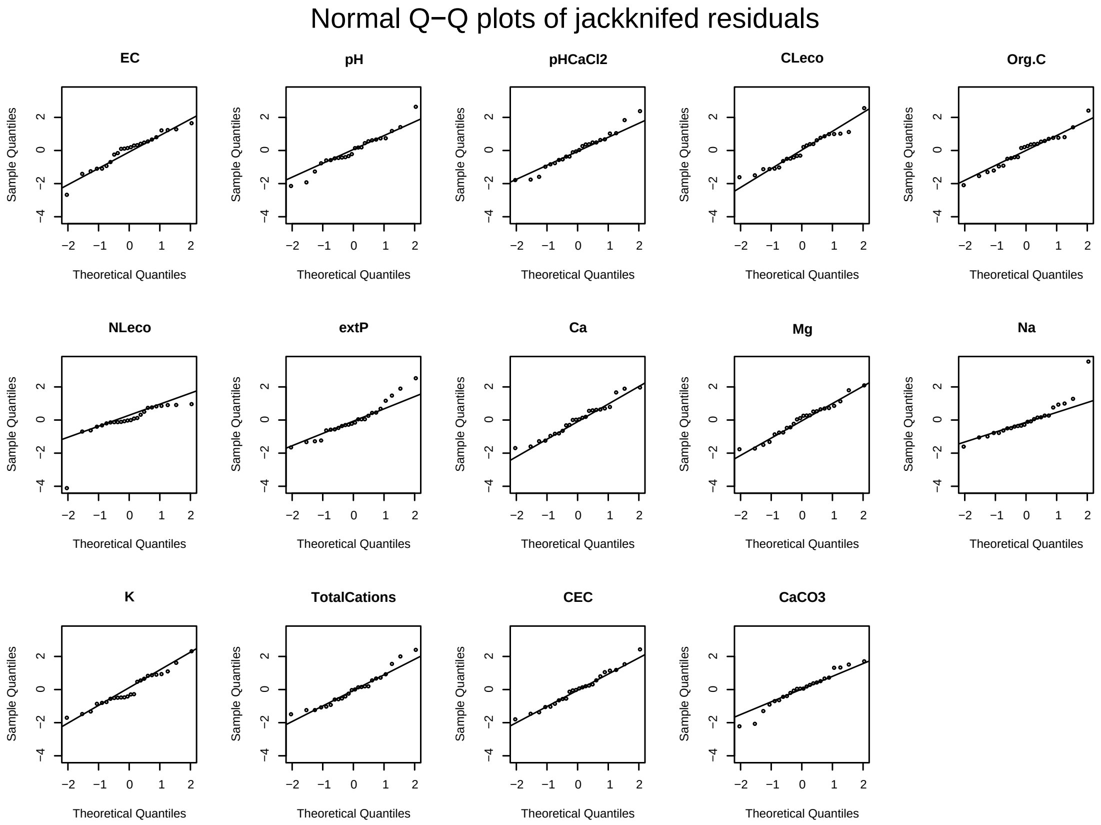

Figure 4.3: Normal quantiles plot for the residuals (all observations)

5 Possible extension

In Section 4 the infra–red spectrometry data were analyzed twice—with and without the $10th$ observation. As shown in Figure 1.2, this observation appears to be an outlier for the response variable NLeco. With respect to the other variables, however, it does not seem suspicious. Even though the result of the analysis in Section 4 does not seem to be influenced by the outlier, it may be preferable in other situations to mark such observations as missing values. This would allow us to use the observations for the other response variables in the analysis.

Theoretically, the methodology proposed here can be modified easily to take missing values in the response variables into account. One way to modify the procedure would be to remove the corresponding rows from $\underset{\sim}{y}$ and $\tilde{\mathbf{X}}$ in (3.1). We plan to investigate this modification in a future project. However, given the similarity of the results in Section 4, regardless on whether the outlying observation was deleted, we believe that application of this modification on the data set studied here would not be of high interest.

The implementation described in this paper has put some strain on our computational resources. A rough analysis shows that the algorithm, as described in Section 3.3, needs memory of the order $O(p^2 + np + kn^2)$ and the number of operations per iteration is roughly $O(p^3 + kn^3 + kpn^2)$. A detailed description of the implementation of our algorithm can be found in the technical report, available from the authors on request, on which this paper is based.

Implementation of the modification with missing rows could be performed without increasing the computational demands appreciably, though the complexity of the implementation

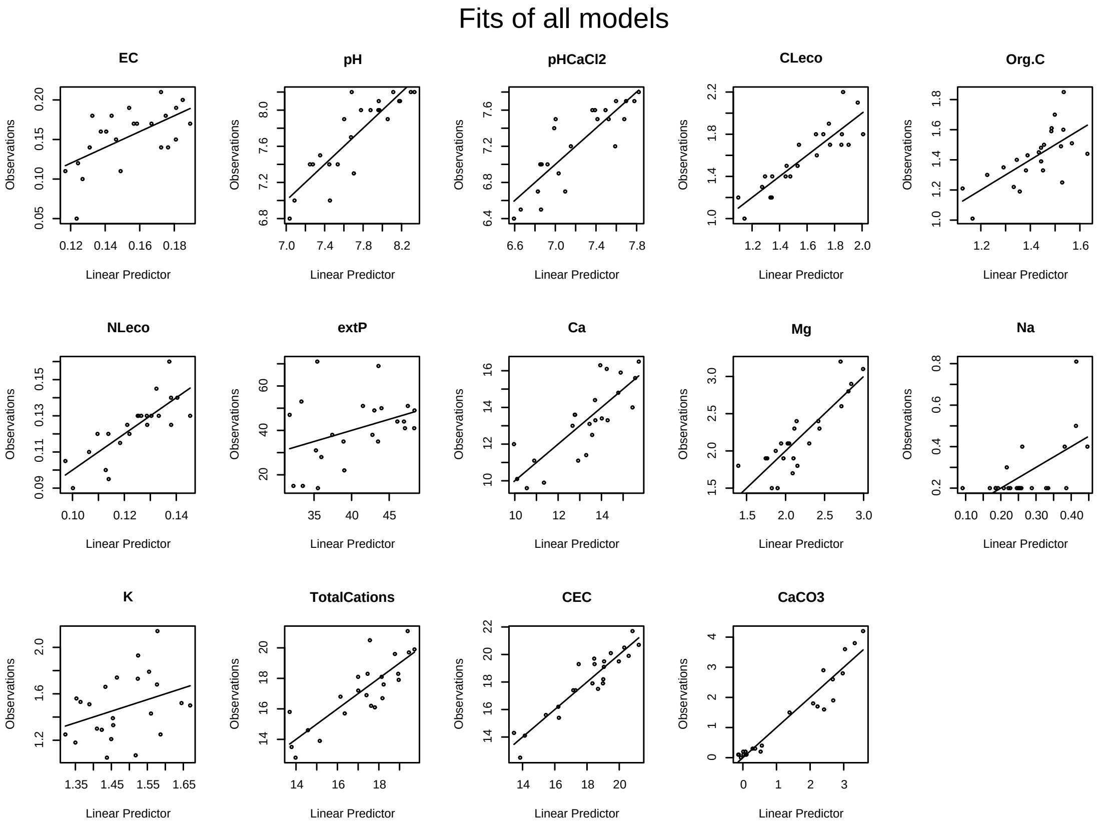

Figure 4.4: Linear fits using 3 explanatory variables (without $10th$ observation)

would increase. We would still need to store only a single copy of the matrix $\mathbf{X}$, together with a list of missing values for each observation.

For the infra-red spectrometry data, the dimensions are $k = 14$ and $p = 770$. All results shown in this manuscript were calculated on a 450MHz Pentium PC with 128MB of RAM running Linux. Because of the size of p for this data set, each iteration of the interior point algorithm used roughly 10.4 seconds and, depending on the value of t, we needed between 2 and 5 minutes to calculate the solution of (1.2). By way of contrast, the results shown in Figure 1.1 were calculated using the algorithm described in Osborne *et al.* (2000b). That algorithm is an active set algorithm specifically designed to calculate the LASSO estimate fast and efficiently. Each panel in Figure 1.1 is based on 80 equispaced values for t between 0 and 1. For a single response variable, the time to calculate the solutions of (1.1) for all 80 values of t was roughly 3.2 seconds. Unfortunately, the active set algorithm of Osborne *et al.* (2000b) can not be adapted readily to the more general problem (1.2). Given the large difference in performance, however, we believe that it would be worthwhile to develop an active set method to solve (1.2).

6 Conclusions

Tibshirani (1996) showed that restricting parameter estimates to a polyhedral region while minimizing the residual sum of squares yields a method that successfully combines elements of ridge regression and subset selection. In this paper we extended his idea to the situation

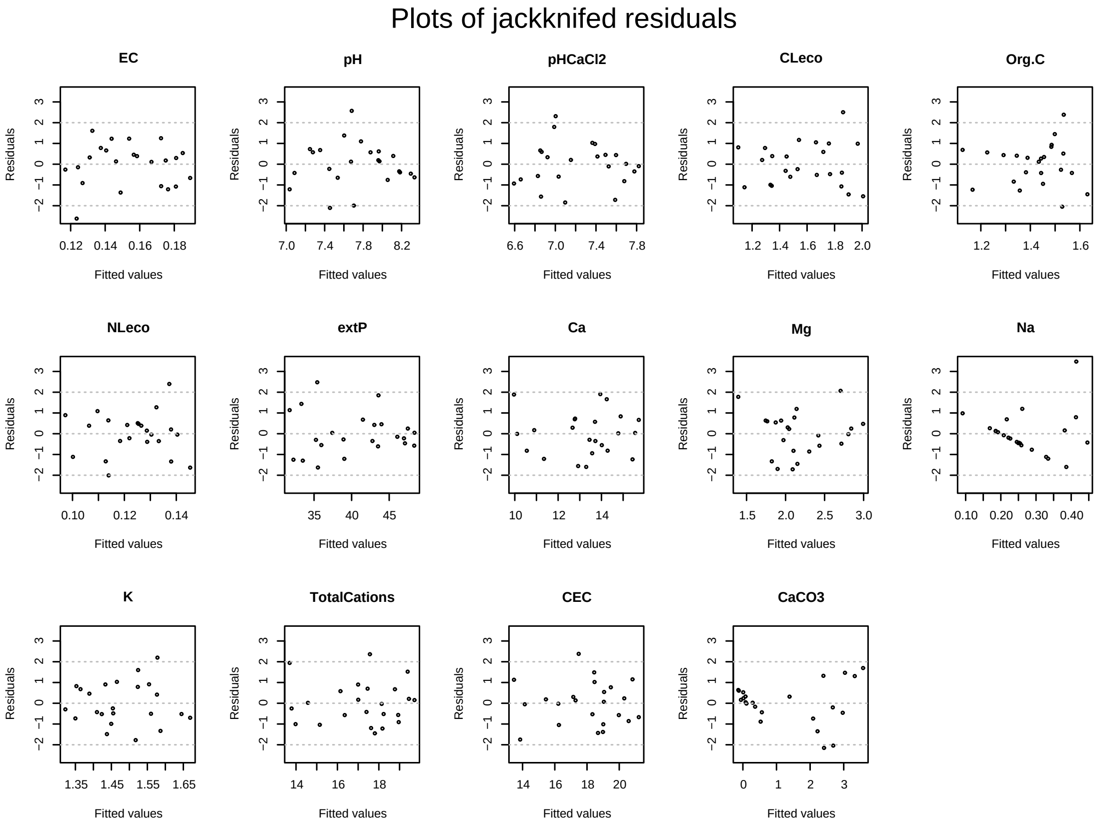

Figure 4.5: Residual plots (without $10th$ observation)

in which we seek a subset of regressor variables that is useful for several response variables simultaneously. This extension leads again to a convex programming problem; we describe an interior point algorithm for solving this problem.

Application of this method to the infra-red spectrometry data shows that this method can be quite useful in identifying a single subset for simultaneous modelling purposes. The example chosen to motivate and illustrate much of what we have done here was extreme in the sense that the number of regressors is huge, the number of responses is moderate and the number of observations is almost unrealistically small. We contend that in less extreme cases the computational load will be more manageable and the methodology will be just as useful. Testing the method on more data sets would be enlightening.

Acknowledgements

The authors would like to thank the Editor and three referees for their comments which helped to improve this paper. Research of the third author was supported by the U.S. National Science Foundation under Grants 9726385, 0196485, 0296033, 0127857, and 0113051.

References

Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems*, PhD thesis, Australian National University, Canberra ACT 0200, Australia.

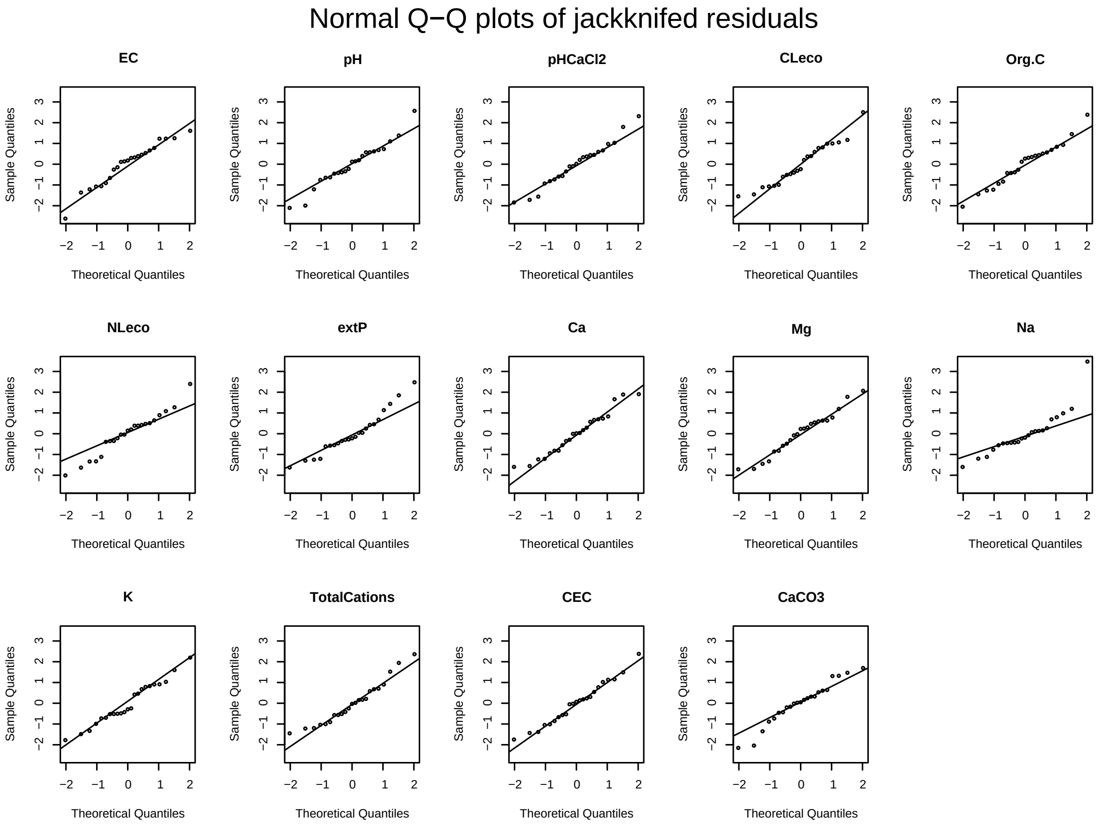

Figure 4.6: Normal quantiles plot for the residuals (without 10*th* observation)

Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics* **37**(4): 373–384.

Breiman, L. and Friedman, J.H. (1997). Predicting multivariate responses in multiple linear regression (with discussion), *Journal of the Royal Statistical Society, Series B* **59**(1): 3–54.

Brown, P.J. (1993). *Measurement, regression, and calibration*, Clarendon Press, Oxford.

Brown, P.J., Fearn, T. and Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate bayesian decision approach, *Biometrika* **86**(3): 635–648.

Brown, P.J., Vannucci, M. and Fearn, T. (1998). Multivariate bayesian variable selection and prediction, *Journal of the Royal Statistical Society, Series B* **60**(3): 627–641.

Brown, P.J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors, *Journal of the Royal Statistical Society, Series B* **64**(3): 519–536.

Burnham, K.P. and Anderson, D.A. (1998). *Model Selection and Inference: A Practical Information Theoretic Approach*, New York, Springer-Verlag.

Clark, D.I. and Osborne, M.R. (1988). On linear restricted and interval least-squares problems, *IMA Journal of Numerical Analysis* **8**: 23–36.

Clarke, F.H. (1990). *Optimization and Nonsmooth Analysis*, Vol. 5 of *Classics in Applied Mathematics*, SIAM, Philadelphia. Originally published by John Wiley & Sons (1983).

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, Wiley Series in Probability and Statistics, 3 edn, John Wiley & Sons, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion), *Annals of Statistics* **32**(2): 407–499.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**: 109–148.

Golub, G.H. and van Loan, C.F. (1996). *Matrix Computations*, 3 edn, John Hopkins University Press, Baltimore.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hocking, R.R. (1996). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Leamer, E.E. (1978). Regression selection strategies and revealed priors, *Journal of the American Statistical Association* **73**(363): 580–587.

Lobo, M.S., Vandenberghe, L., Boyd, S. and Lebret, H. (1998). Applications of second-order cone programming, *Linear Algebra and its Applications* **284**(1–3): 193–228.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*, John Wiley & Sons, Chichester.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Mehrotra, S. (1992). On the implementation of a primal-dual interior point method, *SIAM Journal on Optimization* **2**: 575–601.

Miller, A.J. (1990). *Subset Selection in Regression*, Vol. 40 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Osborne, M.R. (1985). *Finite Algorithms in Optimization and Data Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Chichester.

Osborne, M.R. (1992). An effective method for computing regression quantiles, *IMA Journal of Numerical Analysis* **12**: 151–166.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis* **20**(3): 389–403.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000b). On the LASSO and its dual, *Journal of Computational and Graphical Statistics* **9**(2): 319–337.

Rockafellar, R.T. (1970). *Convex Analysis*, Vol. 28 of *Princeton Mathematical Series*, Princeton University Press, Princeton, New Jersey.

Roos, C., Terlaky, T. and Vial, J.P. (1997). *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley & Sons, New York.

Sturm, J.F. (1999). Using SEDUMI 1.02: A Matlab Toolbox for optimization over symmetric cones, *Optimization Methods and Software* **11–12**: 625–653.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.

Wold, H. (1984). PLS regression, *in* N.L. Johnson and S. Kotz (eds), *Encyclopedia of Statistical Sciences*, Vol. 6, John Wiley & Sons, New York, pp. 581–591.

Wright, S.J. (1997). *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia.

Ye, Y. (1997). *Interior Point Algorithms: Theory and Analysis*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, New York.

Zou, H. and Hastie, T. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays, *unpublished manuscript*, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.
URL: *http://www-stat.stanford.edu/˜hastie/Papers/elasticnet.pdf*