Using regression function information in determining the same

Tobias Adolfsson

April 30, 2002

Abstract

There exists a number of methods for regression in the case when we have a poorly conditioned design matrix. Most of these methods use various regularisations on the design matrix. We will construct a regression method that uses information in the response as well as the design matrix. The method allows the design matrix to be ill conditioned. The method presented is shown to be theoretically simple compared with other methods using response information. Procedures for dimensionality testing are derived and predictive properties are thoroughly examined. Variance estimation in the case of effect sparsity is asymptotically examined and the asymptotics are studied by simulations.

1 Introduction

Many new and interesting regression methods have been born from the field of chemometrics. These methods are often subspace regression methods, i.e instead of solving the complete problem they solve an easier problem with lower dimension that approximates the original problem. These methods are used since

a the dimension is often very large and

b there is a wish to reduce the noise.

The underlying models and theory are quite easy to understand but to check its correctness is a rather difficult problem. Since one is working with subspace regression it is of crucial importance to get the correct dimension of the subspace. If a wrong decision is made one can make two basic mistakes: too small dimension or too large dimension. In the former case the regression function will be inadequate in the sense that information is lost. If the dimension is too large the variance of the estimates will be too large. As always there is a trade off between variance and unbiasedness. By increasing the bias one could reduce the variance and hence increase the precision of the estimates. If, on the other hand, the bias is too big the answer will be useless. The methods mainly used for assessing the dimension is cross validation, Aikake's Cp, various bootstrap methods etc, see Denham (2000) for a view of the methods used in PLS. One of the problems of estimating the dimension is that the methods are often constructed in such a way that there is no natural method to choose dimension. Our regression method is constructed to use the information gained in the experiment in an efficient way as well as being able to draw conclusions about the dimensionality of the subspaces. The idea is to let the size of the regression function on the normed explanatory variables decide the subspaces. When this space is drained of information, using a method based on the Singular Value Decomposition (SVD), we can construct a test for that the derived subspace really includes the effects we want. For more information on SVD and related topics see Horn and Johnson (1985, 1991). The rational of this method can be viewed from a perspective that might not be the usual in statistics. When solving a regression problem in statistics we usually write it as finding β that minimises $||Y - X\beta||_2$, this is however equivalent to minimising $||\varepsilon||_2$. The drawback with this method is that we force a minimisation on the hole of ε, this might not be necessary or even correct. It might well be the case that we should rather minimise ε projected onto another space, i.e. $||P\varepsilon||_2$ would be the correct quantity to minimise. If this is the case we would actually allow the error to become large in the directions where we cannot control it, then simply ignore it and make it small in the directions that we can control, and finally include these into our model. The principal idea in this article is to let the regression function tell us where it does make a difference and in this way help us to minimise the error where we actually can do this. This keeps us faithful to the old ideas of least squares but on a space where the regression function helps us to understand that it is meaningful. We will show that the criteria for which directions to include will be very simple.

There is a method based on a testing procedure for Principal Components Regression (PCR) called Significance Regression (SR) which will yeild similar estimations, see Faber (2001). SR does not control the multiple level of significance as our method will. The method of derivation is also different. I am not aware of any multivariate eqvivalent of SR whereas our method can be directly transfered to the multivariate case.

2 The model and basic methods

Let us begin with the usual univariate regression model

$$Y = X\beta + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^{p\times 1}$ and obviously $Y \in \mathbb{R}^{n\times 1}$. We consider the design matrix X as fixed but not controlled i.e. we have not really designed X but rather it contains far more columns than needed. The last note about not controlled is not necessary but if X contains highly structured data we can use better methods for analysis than the one we are going to construct.

When an experiment has been performed we know X and Y. Any analysis that we perform can only be based on this matrix and vector. As X is thought of as containing the means of which we control the outcomes in Y (disregarding random errors) we see that every meaningful estimate of β must lie in the column space of X. This of course extends to the usual definition of estimability. We hence believe that we from X can get Y apart from a random error. The Ordinary Least Squares (OLS) solution is obtained by finding $\beta \in \mathbb{R}^{p\times 1}$ such that $||Y - X\beta||_2$ is minimised. In some applications X might not have full rank or at least X has a very poor condition number. If this is the case it is well known that OLS will, at least on average, give poor estimates. There are different methods for getting around this problem. One way is to regularize X as in, for instance ridge regression, and hence create a new problem which is easier to solve. We could also use a smaller part of X, i.e. a subspace of the column space of X, as an approximation of X and solve this smaller problem. In both these approaches we make regularizations or put restrictions on X while the problem is about Y or β depending on what we want as end result. In Principal Component Regression (PCR) we use the singular value decomposition (SVD) to give us the dominate space in which we solve our problem. Using a rank k truncated SVD to estimate a matrix

gives the best approximation in Euclidean norm but once again I stress that we are not trying to solve any problem for X! It might well be that not all of X:s column space is needed to solve the problem but should we not let Y help decide what parts of X to use? There are methods that incorporate Y into the subspace selection process. One such example is Partial Least Squares (PLS) which stepwise chooses the maximum covariance between X and Y when looking for the correct subspace. But PLS has proven to be rather complicated when it comes to assessing properties as bias or variance of the estimates. See Helland (1988, 1990); Helland and Almøy (1994), von Rosen (1994), Frank and Friedman (1993) and Adolfsson (1999) for more information on PLS. Some effort has lately been put into testing procedure to determine the number of factors in PLS see Aziz and Cléroux (2001) and Denham (2000).

3 A new way of choosing the subspace for regression

The idea is once again choose to a subspace on which we solve the problem. This time however we use the regression function to help us pick out the right space. Using the model (1) we make a SVD of X, that is

$$X = U\Lambda V', \quad U'U = I_n, \quad V'V = I_p.$$

where

$$\Lambda = \begin{pmatrix} D & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n\times p}, \quad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_r),$$

$r \leq \min(n, p)$ and $\lambda_1 \geq \ldots \geq \lambda_r > 0$. For notational convenience we adopt the the notation

$$\Lambda^{-1} = \begin{pmatrix} D^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{p\times n}.$$

The matrix U_rU_r', where U_r is the matrix formed by the r first columns, is the orthogonal projection onto the column space of X which is the space where the estimate of β must lie. The OLS estimate of β is $\hat{\beta} = (X'X)^{-1}X'Y$. We now want to study the behaviour of this estimate on the level curves of $X'X$ or rather the normalised level curves. From the SVD of X we have that $X'X = V\Lambda^2V'$. We now introduce an artificial variable $Z = XV\Lambda^{-1}$ and solve the equation

$$Y = Z\gamma + \varepsilon$$

instead. This equation has the solution

$$\hat{\gamma} = Z'Y = \Lambda^{-1}V'X'Y = \Lambda^{-1}\Lambda U'Y = U'Y.$$

This vector describes the direction in the column space of X on which the estimated regression function is maximal. This however only gives us one direction which in general is far to small a space for solving the problem. We will however use this general idea to get the space on which we will solve the problem. Instead of choosing only the maximal direction in $S = col(X)$ we will choose the K-dimensional orthogonal base of a subspace R of S that best approximates the maximum normed regression function in a $||\cdot||_2$ sense.

We set $U_r = [u_1, \ldots, u_r]$ and project Y onto these vectors one at a time, hence $r_i = u_i u_i'Y, \quad i = 1, \ldots, r$. Breaking this down into parts we examine the scalar $u_i'Y$ more carefully. According to (1) we have

$$\begin{aligned} u_i'Y &= u_i'(X\beta + \varepsilon) = u_i'(U\Lambda V'\beta + \varepsilon) = e_i'\Lambda V'\beta + u_i'\varepsilon = \lambda_i e_i'V'\beta + u_i'\varepsilon \\ &= \lambda_i v_i'\beta + u_i'\varepsilon = f_i(\beta) + u_i'\varepsilon. \end{aligned} \tag{2}$$

We see that $\lambda_i v_i'\beta$ is a scalar describing how great effect the regression function has along the vector u_i and that $u_i'\varepsilon \sim N(0, \sigma^2 \sum_j u_{ij}^2) = N(0, \sigma^2)$ since u_i has length 1. Further more, letting $\tilde{\varepsilon}_i = u_i'\varepsilon$, we have that $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j$ are uncorrelated (and hence independent) for $i \neq j$ since $u_i \perp u_j$ for $i \neq j$. The vectors $r_1, \ldots, r_r$ now span S but they also carry information about the validness of the space in the sense that there length is proportional to the size of the estimated regression function in the space. We form a new matrix by $R = [r_1, \ldots, r_r]$ and do a SVD on this matrix. In matrix analysis one normally says that a square matrix T is orthogonal if $T'T = I$, I personally think this is a orthonormal matrix rather than an orthogonal. In this nonstandard notation we see that R is an orthogonal matrix, this makes the SVD simple since we can do it by inspection:

$$R = U_R D P_n, \quad D = \text{diag}(d_{(i)}).$$

The matrix U_R has the same columns as U but not necessarily in the same order and the d_i:s are the absolute values of $f_i(\beta) + \tilde{\varepsilon}_i$ ordered according to their size and the matrix P_n is a permutation matrix. This representation is useful since this tells us that if we use the TSVD with dimension k of R we can not do better in $||\cdot||_2$-sense when estimating the space containing the biggest estimated regression function. Another big advantage of this

approach is that the distributional properties are easy to handle since we have independence. In a setting where n is moderately large and we from prior knowledge know that large portions of X does not effect Y we can easily derive a testing procedure, for choosing the correct k, which maintains the correct multiple level of significance (see Adolfsson (2002b)). Assuming that we have now fixed a subspace R on which to solve the problem we now just project the equation onto this subspace and solve the problem there.

4 Multivariate regression

With some minor modification the same method and arguments goes through for multivariate regression. We now assume that we have a model

$$Y = XB + E, \quad E \sim N_{r,n}(0, \sigma^2 I_n, I_r)$$

where $Y = (Y_1, \ldots, Y_r) \in \mathbb{R}^{n \times r}$, $B = (\beta_1, \ldots, \beta_r) \in \mathbb{R}^{p \times r}$ and
$E = (\varepsilon_1, \ldots, \varepsilon_r)$. Using the same notation as in the previous section we have

$$\sum_{j=i}^{r} (u_i' Y_j)^2 = \sum_{j=1}^{r} \left(\lambda_i v_i' \beta_j + u_i' \varepsilon_j\right)^2 = \sum_{j=1}^{r} \left(\lambda_i v_i' \beta_j + \tilde{\varepsilon}_i\right)^2 \tag{3}$$

if we once more set $X = U \Lambda V'$. Let us now assume that there are no true regression effects. We will then have that each of the $\sum_{j=i}^{r} (u_i' Y_j)^2 t$ is χ_r^2 and independent. If we now use the k (say) smallest of these variables to estimate the standard deviation σ we will have the following likelihood function

$$L(\sigma) = \binom{n}{k} \prod_{i=1}^{k} \left(\frac{1}{\sigma^2} f_{\chi_r^2}\left(\frac{x_i}{\sigma^2}\right)\right) \left(1 - F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)\right)^{n-k}.$$

The log-likelihood function, disregarding constants, will then be

$$\begin{aligned} l(\sigma) &= -2k \ln(\sigma) + \sum_{i=1}^{k} \left(\frac{k}{2} - 1\right) \ln\left(\frac{x_i}{\sigma^2}\right) - \sum_{i=1}^{k} \frac{x_i}{2\sigma^2} \\ &+ (n-k) \ln\left(1 - F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)\right). \end{aligned}$$

Setting $\frac{\partial l}{\partial \sigma} = 0$ we get

$$\sigma^2 = \frac{k}{2(n-k)} (r\sigma^2 - S_k^2) \frac{1 - F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}{\frac{x_r}{\sigma^2} f_{\chi_r^2}\left(\frac{x_r}{\sigma^2}\right)}$$

where

$$S_k^2 = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

Using lemma 1 and lemma 2 in Adolfsson (2002b) and noting that the failure rate of the gamma distribution is non decreasing for the parameter values of interest we can construct a test for true regression effects that maintains the correct multiple level of significance.

4.1 Efficiency of the variance estimate for multiple testing

If we have a regression experiment where we know that a large portion of X is non informative the estimate of σ above will probably be rather good, but how large portion do we need? One way of answering this question would be to examine the asymptotic variance of the estimate. It is well known that, under certain regularity conditions, we asymptotically achieve the Cramér-Rao limit for ml-estimates. We will now calculate this for the estimation method given above. We have that

$$-\frac{\partial^2 l}{\partial \sigma^2} = \frac{3}{\sigma^2}\sum_{i=1}^{k}\frac{x_i}{\sigma^2} - \frac{kr}{\sigma^2} - \frac{2}{\sigma^2}\frac{\frac{x_k}{\sigma^2}f_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}{1-F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}\left(\frac{x_k}{\sigma^2} - r - 1 - \frac{\frac{x_k}{\sigma^2}f_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}{1-F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}\right).$$

We now assume that $k/n = p \in (0,1)$ and let $k, n \to \infty$. By the law of large numbers we will then have that $x_k/\sigma^2 \to z_\infty = F_{\chi_r^2}^{-1}(p)$, $\frac{1}{n}\sum_{i=1}^{k} X_i/\sigma^2 \to \mathbb{E}[Z|Z < z_\infty]$, $(Z \sim \chi_r^2)$ and hence, $F_{\chi_r^2}(x_k/\sigma^2) \to p$. We now have the following asymptotic expression

$$\begin{aligned} -\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} &= \frac{1}{\sigma^2}\Bigg(3\mathbb{E}[Z|Z<z_\infty] - pr \\ &+ z_\infty f_{\chi_r^2}(z_\infty)\left(2\frac{z_\infty f_{\chi_r^2}(z_\infty)}{1-F_{\chi_r^2}(z_\infty)} + r + 1 - z_\infty\right)\Bigg) \\ &= \frac{1}{\sigma^2}f(p). \end{aligned}$$

It is easy to see that $-\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} \to 0$ as $p \to 0$ but the limit $p \to 1$ might not be as obvious. We have however that

$$\frac{f_{\chi_r^2}(z_\infty)}{1-F_{\chi_r^2}(z_\infty)} = \frac{c(z_\infty)^{\alpha-1}e^{-\frac{z_\infty}{2}}}{\int_{z_\infty}^{\infty} cu^{\alpha-1}e^{-\frac{u}{2}}du}$$

$$
\begin{aligned}
&= \frac{1}{\int_0^\infty \left(1+\frac{v}{z_\infty}\right)^{\alpha-1} e^{-\frac{v}{2}} dv} \\
&\leq \frac{1}{\int_0^\infty e^{-\frac{v}{2}} dv} = 2
\end{aligned}
$$

since $\alpha \geq 1$ and $z_\infty \geq 0$ and hence it is clear that $-\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} \to \frac{2r}{\sigma^2}$ as $p \to 1$. Since the ordinary estimate of the variance using all observations has the same variance as our estimate when $p = 1$ we do not seem too aim to much off target. A plot of the variance factor $f(p)$ for different degrees of freedom can be found in figure 1.

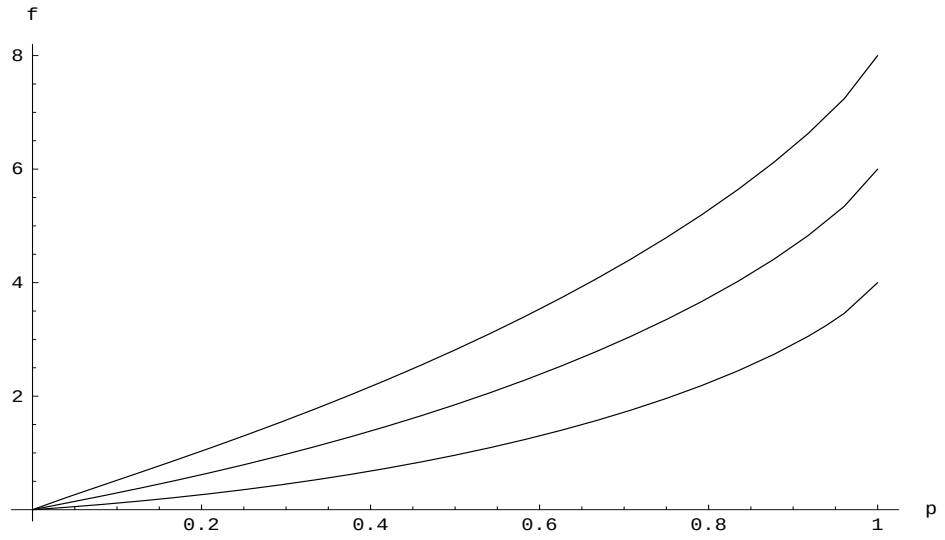

Figure 1: Asymptotic variance factor for 2, 3 and 4 degrees of freedom.

5 Prediction properties

In applications such as, for instance calibration, the predictive ability and properties are of great importance. We will now examine the Mean Square Error of Prediction (MSEP). The MSEP is defined as

$$
\text{MSEP} = \mathbb{E}\left[||Y_0 - \hat{Y}_0||_2^2 \right].
$$

Let us assume that the new observations that we want to predict are new observations in the same sample points as used for the estimation. In the case of calibration this would make sense since we would calibrate where we want to measure. Let $\beta_{\mathcal{I}} = \mathbb{E}\left[\hat{\beta}_{\mathcal{I}}\right]$ where $\hat{\beta}_{\mathcal{I}}$ is the estimate using $u_i : i \in \mathcal{I}$, $|\mathcal{I}| = k$ where $\mathcal{I}$ is a (fixed) set of indexes $1 \leq |\mathcal{I}| \leq n$. We then get

$$\text{MSEP} \propto \mathbb{E}\left[(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})' X' X (\hat{\beta}_{\mathcal{I}} - \beta_I)\right] + \mathbb{E}[(\beta - \beta_{\mathcal{I}})' X' X (\beta - \beta_{\mathcal{I}})]. \quad (4)$$

We now rewrite the estimate $\beta_{\mathcal{I}}$ using the following observation: If $X = U\Lambda V'$ we get the OLS solution as $\beta = X^+ Y$ where X^+ is the Moore-Penrose generalised inverse. This inverse is given by $X^+ = V\Lambda^{-1}U'$ (see Horn and Johnson (1985) p.421) and hence we have that

$$\hat{\beta} = \sum_{i=1}^{n} \frac{u_i' Y}{\lambda_i} v_i.$$

We then have that

$$\hat{\beta}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \frac{u_i' Y}{\lambda_i} v_i = \sum_{i \in \mathcal{I}} \frac{\lambda_i v_i' \beta + \tilde{\varepsilon}_i}{\lambda_i} v_i = \sum_{i \in \mathcal{I}} v_i' \beta v_i + \sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i.$$

One can now immediately see that

$$\mathbb{E}\left[\hat{\beta}_I\right] = \sum_{i \in \mathcal{I}} v_i' \beta v_i \text{ and } \mathbb{E}\left[\beta - \hat{\beta}_{\mathcal{I}}\right] = \sum_{i \notin \mathcal{I}} v_i' \beta v_i.$$

Noting that $X'X = V\Lambda^2 V'$ we have that

$$\begin{aligned} \mathbb{E}\left[(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})' X' X (\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})\right] &= \mathbb{E}\left[\left(\sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i\right)' V\Lambda^2 V' \left(\sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i\right)\right] \\ &= \sum_{i \in \mathcal{I}} \mathbb{E}\left[\tilde{\varepsilon}_i^2\right] = |\mathcal{I}|\sigma^2, \end{aligned} \quad (5)$$

and

$$\begin{aligned} (\beta - \beta_{\mathcal{I}})' X' X (\beta - \beta_{\mathcal{I}}) &= \left(\sum_{i \notin \mathcal{I}} v_i' \beta v_i\right)' V\Lambda^2 V' \left(\sum_{i \notin \mathcal{I}} v_i' \beta v_i\right) \\ &= \sum_{i \notin \mathcal{I}} (\lambda_i v_i' \beta)^2. \end{aligned} \quad (6)$$

Recalling that $u_i Y = \lambda_i v_i' \beta + \tilde{\varepsilon}_i$ the relationship between effect and variance, with respect to prediction, is now clear. If we include a new direction u_l we will decrease the squared bias with a factor $(\lambda_l v_l' \beta)^2$, according to (6), but at the same time we will increase the variance with a factor σ^2 due to (5). These observations together with (4) gives our decision rule for prediction: if the absolute value of an effect is less then the standard deviation it should be excluded, and if the effect exceeds the standard deviation it should be included! This is obvious since if we include a direction that includes an effect less then σ the increase in variance is σ^2 but the decrease in bias is less then σ^2 and vice versa. This also gives the implication that when the testing procedure is used the power should be great for alternatives where the effect exceeds the standard deviation for other alternatives we should not include the effect.

6 With a stochastic X

Let X be a stochastic matrix with a distribution that is independent of ε. There exists a SVD of $X = U\Lambda V'$ where U, Λ and V are independent of ε. In the case when $X\beta = 0$ eq. (2) gives

$$\frac{1}{\sigma^2}(u_i' Y)^2 = \frac{1}{\sigma^2}(u_i'\varepsilon)(\varepsilon' u_i) = \frac{1}{\sigma^2}\frac{u_i'\varepsilon\varepsilon' u_i}{u_i' u_i} \sim \chi_1^2$$

since $\varepsilon\varepsilon'$ has a Pseudo-Wishart distribution see, for instance, Kshirsagar (1972) for a proof. Note that this result is valid for any stochastic orthogonal set of vectors $u_i, \ldots, u_n$ which is independent of ε.

Thus the null case is clear. When there are true regression effects present it is not clear what happens. Since the u_i:s are stochastic the resulting distribution will be of the type $\chi_r^2(\delta)$ but δ is stochastic. In simulations this has show to have small effect on the method. The result in section 5 are however no longer valid.

7 A spectroscopic example

The data set used in this example consists of 440 measurements of organic solvent residue in penicillium samples. The concentration of four different solvents where measured in each sample. The different samples where then

analysed in a NIR-spectrometer and the absorbation at 1050 wavelengths where recorded for each sample. Since the the absorbation follows Beer's law we have the setup

$$Y = X\beta, Y \in \mathbb{R}^{440\times 4}, X \in \mathbb{R}^{440\times 1050}.$$

The goal is to estimate β, i.e. the to calibrate the spectrometer. The condition number of a matrix is the quotient of the largest and the smallest singular values of the matrix. The larger the condition number the more ill-condition is the matrix. The condition number of X is $6.0 \cdot 10^5$ which means that X is rather ill-conditioned. The singular values $\lambda_3, \ldots, \lambda_{440}$ of X is displayed in figure 2. Using eq. (3) in section 4 to form the information

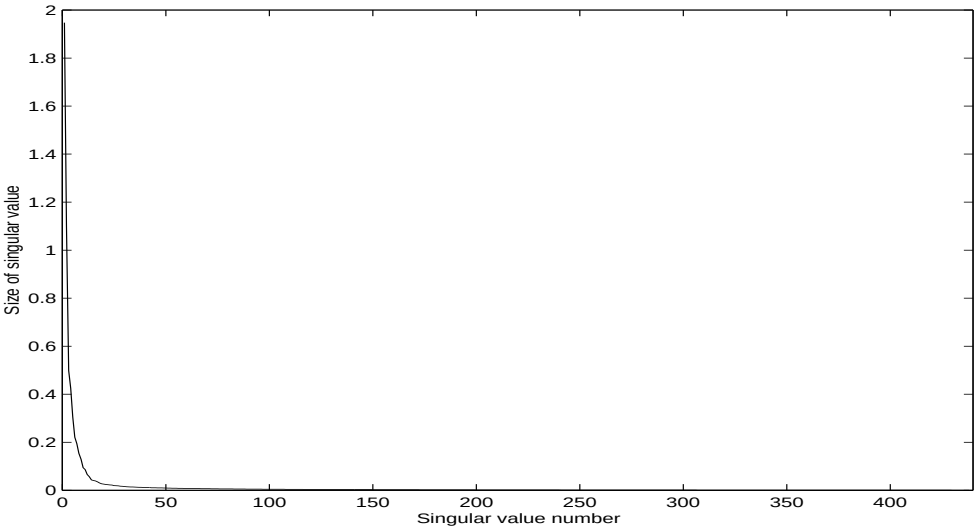

Figure 2: The singular values of X.

in each of the singular vectors u_i we can study the size of the singular values when they are reordered by the information. In figure 3 we see the singular values ordered by the information criteria and vice versa. The figure should be viewed column vise. The top right plot displays $\lambda_3, \ldots, \lambda_{40}$ and the bottom right plot shows the u_iY:s ordered according to the λ:s. The top left plot shows the λ:s ordered according to the size of the u_iY plotted in the bottom leftmost plot. The two largest effects are excluded in all plots since the coincide and are so large that they make smaller observations difficult to

plot. In these plots one for instance see that one of the largest singular values occurs, when sorted by the $u_i Y$:s, on 23:th place. Large singular values is no guarantee that there is information present.

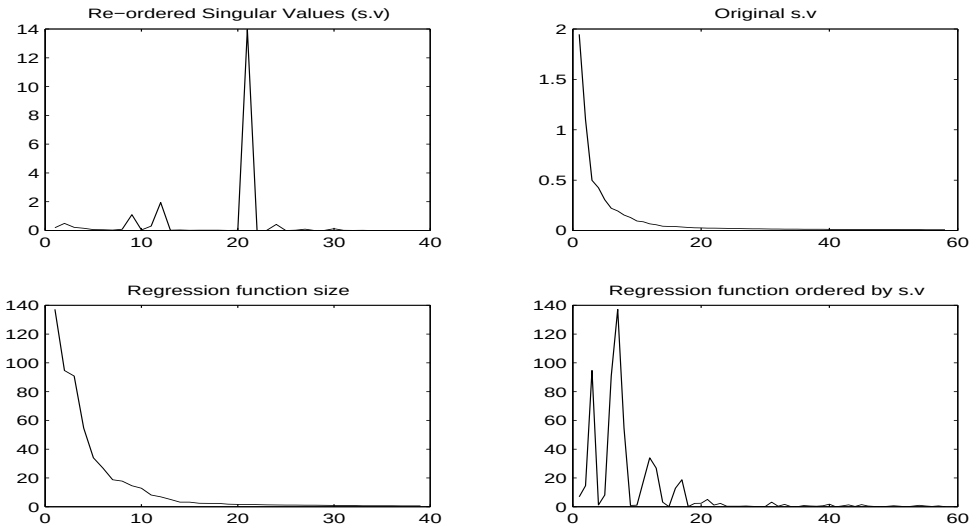

Figure 3: Reordering of the singular values.

Comparing this regression method to that of PCR we see that the residual sum of squares is smaller for the same number of base vectors (components). Using the step-down test procedure with the smallest 250 observations of the statistics for variance estimation we get that we can declare the first 71 base vectors as significant. Using the test procedure in Adolfsson (2002a) which is used for determining the number of components to use in PCR we get 27.

8 Conclusions

The regression method purposed in this paper shows great potential both theoretically and practically. In the case of a non-stochastic X matrix the predictive properties is easy to asses. The method uses information from both the dependent and independent variables but in a way that is easy to understand. It should be noted that any orthogonalisation of X can be used as long as it is independent of Y. If we are fitting a linear model where we have particular interest in quadratic terms of the x_i:s we could choose an

orthogonalisation where we the quadratic terms are one of the orthogonal base vectors. This might be fruitful when we have prior knowledge of the problem. From a computational point of view it is $\max(n, p)$, $X \in \mathbb{R}^{n \times p}$ that determines the complexity of the problem. In, for instance, spectroscopical problems the limiting factor is normally the number of samples in the calibration set. if the sample size is moderate (< 500) the method is fast to calculate. If we have prior knowledge of the number of the rank of similar problems we could use the step-down method described in the paper. The step-down method does not disturb the structure of either X or Y which might be the case in cross validation, see Adolfsson (1999). On a reasonable fast computer the simulations for the step-down method is done rather fast.

References

T. Adolfsson. Partial least squares and its implementations from a statistical point of view. Technical report, Chalmers University of Technology and Göteborg University, 1999.

T. Adolfsson. *A testing procedure for dermening the chemical rank of spectrometrical absorption matrices - a heuristic approach.* PhD thesis, Chalmers University of Technology and Göteborg University, 2002a.

T. Adolfsson. *Variance estimation and multiple inference testing in saturated orthogonal designs.* PhD thesis, Chalmers University of Technology and Göteborg University, 2002b.

L. Aziz and R. Cléroux. The pls multivariate regression model: testing the significans of successive pls components. *J. Chemometrics*, 2001.

M. C. Denham. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemometrics*, 14:351–361, 2000.

N. M. Faber. Critical evaluation of a significance test for partial least squares regression. *Anal. Chim. Acta*, pages 235–240, 2001.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 1993.

I. S. Helland. On the structure of partial least squares regression. *Comm. Statist. B-simulation Comput.*, 1988.

I. S. Helland. Partial least squares regression and statistical models. *Scand. J. Statist.*, 1990.

I. S. Helland and T. Almøy. Comparison of prediction methods when only a few components are relevant. *J. Amer. Statist. Assoc.*, 1994.

R. A. Horn and J. A Johnson. *Matrix analysis.* Cambridge university press, 1985.

R. A. Horn and J. A Johnson. *Topics in matrix analysis.* Cambridge university press, 1991.

A. M. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, 1972.

D. von Rosen. Partial least squares and a linear model. In *Proceedings at the International Conference on Linear Statistical Inference LINSTAT 93*, 1994.