A NEW APPROACH TO THE VARIABLE SELECTION USING THE TLS APPROACH

Sébastien MARIA, *Jean-Jacques* FUCHS

IRISA/Université de Rennes I
Campus de Beaulieu - 35042 Rennes Cedex - France
smaria@irisa.fr, fuchs@irisa.fr

ABSTRACT

The problem of variable selection is one of the most pervasive model selection problems in statistical applications. It arises when one wants to explain the observations or data adequately by a subset of possible regression variables. The objective is to identify factors of importance and to include only variables that contribute significantly to the reduction of the prediction error.

Numerous selection procedures have been proposed in the classical multiple linear regression model. We extend one of the most popular methods developped in this context, the backward selection procedure, to a more general class of models.

In the basic linear regression model, errors are present on the observations only, if errors are present on the regressors as well, one gets the errors-in-variables model which for Gaussian noise becomes the total-least-squares model, this is the context considered here.

1. INTRODUCTION

The aim of variable selection is to model the relationship between an observation vector and a subset of potential regressors. Not every regressor is relevant, some of them might be redundant or unrelated. Therefore, it becomes necessary to choose a good subset of regressors especially when their number is important. This process is called variable selection or subset selection.

In the litterature, a lot of models have already been proposed, the most widely used being the linear regression. The variable selection problem is most familiar in the linear regression context where attention is restricted to normal linear models.

More formaly, let y be an observation vector and $X_1, ..., X_p$ a set of potential regressors participating the observation, each component of y corresponds to an experimentation. The design matrix $X \in \mathbb{R}^{n \times p}$ contains all the possible regressors. Selecting variable is then equivalent to finding the good subset of columns which actually takes part to the regression. The model can be written as:

$$y = X\beta^* + e, \tag{1}$$

where y is a n-dimensional vector, β^* the p-dimensional vector of regression coefficients and $e \sim N(0, \sigma^2 I)$, the measurement error vector. We assume that there are more experimentations than possible regressors, $n > p$, and that X is full column-rank.

Ideally to get the best subset according to a given criterion, one has to examine all the $2^p - 1$ possible models. The most common criterion are the Residual Sum of Squares, RSS or R^2 [1] , and Mallows' $C(p)$ statistic [2]. But the computational complexity of such an exhaustive search makes this approach impratical for even reasonable numbers of regressors. To circumvent this difficulty, many algorithms have been proposed in the litterature. A good compromise between computation time and efficiency is provided by the Stepwise Procedures. These include forward selection procedure, backward elimination procedure and stepwise regression. All these procedures add or remove variables one-at-a-time until some stopping rule is satisfied. In the sequel, we will consider the Backward Elimination Procedures (BEP) in which the idea is to begin with the complete model and to remove the least relevant regressor at each iteration until a preselected significance level tells us that no further removal is justified. It is well known that such procedures can be far from optimal and no global significance level or overall power can be guaranteed [3, 4, 5].

An extension of the BEP is considered here. We replace the basic linear regression model (1) in which only the measurements in y are assumed to be corrupted by errors, by the so-called errors-in-variables model [6] in which the regression matrix X itself is not known precisely and is contaminated by errors. The model becomes

$$y = X\beta^* + e \quad \text{with} \quad Z = X + E, \tag{2}$$

where the vector y and the matrix Z are known or observed, e and E modeling the errors. When these errors are assumed to be Gaussian, the least squares (LS) model associated with (1) becomes then the total least squares (TLS) model [7, 8] . The idea which is developed here is to apply the BEP to the TLS model. This paper is organized as follows: the known results in the LS problem and the backward algorithm are described in section 2, the extension to the TLS model and the corresponding algorithm are proposed in section 3. Then, in section 4, the efficiency of both algorithms are compared on two types of data: simulated data and Hald's data and we conclude in section 5.

2. THE LS CASE

In the basic multiple regression model (1) with Gaussian noise on the observations $e \sim N(0, \sigma^2 I)$, the Maximum Likelihood (ML) estimate of the vector β^* of weights is obtained by solving the Least Squares (LS) problem

$$\min_{\beta} \|y - X\beta\|_2^2 \tag{3}$$

$$\Leftrightarrow \quad \min_{r\ \beta} \|r\|_2^2 \quad \text{subject to} \quad X\beta = y + r,$$

where $\|r\|_2^2 = \sum r_i^2$ denotes the square of the ℓ_2-norm. The optimum is attained at $\hat{\beta} = X^+ y$ where $X^+ = (X^T X)^{-1} X^T$ is the Moore-Penrose inverse of X which is assumed to be full column-rank. Some properties of this estimate are:

- $E(\hat{\beta}) = \beta^*$, no bias
- $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$,
- $\hat{\sigma}^2 = \frac{RSS}{n-p}$ where $RSS = \|y - X\hat{\beta}\|_2^2$,
- $\hat{\beta} \sim N(\beta^*, \sigma^2(X^TX)^{-1})$,
- $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

After the full weight vector $\hat{\beta}$ has been obtained, the BEP decides if a component of $\hat{\beta}_j$ or equivalently a column X_j of X can be removed from the model. The procedure admits two distinct statistical justifications that yield exactly the same decisions. The first method tests if $\hat{\beta}_j$ should be declared equal to zero or not and eliminates the associated regressor if the test is positive. This is called the Student Test. The idea of the second method is to remove a regressor whose contribution to the prediction is probably non-relevant, i.e. if the increase of the residual sum of squares following its deletion is small: this is done in the Fisher test.

2.1. Backward Elimination procedure with the Student test

In the Student test approach, the emphasis is put on the weights, the components in $\hat{\beta}$. Some statistical results must be given to explain the test.

The Student law of parameter k is the law of $U/\sqrt{V/k}$ with $U \sim N(0,1)$ and $V \sim \chi^2_k$, U and V being independent. Observing that $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$, the properties of the full model parameter $\hat{\beta}$ properties given above allows to etablish the following proposition:

Proposition 1 *In the LS context developped above,* $\forall j = 1, .., p$*:*

$$T_j = \frac{\hat{\beta}_j - \beta^*_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$$

where $\hat{V}(\hat{\beta}_j) = \hat{\sigma}^2((X^TX)^{-1})_{j,j}$ *follows a Student law of parameter* $(n-p)$. □

The idea is to test if $\beta^*_j = 0$ by comparing $T_j = \hat{\beta}_j/\sqrt{\hat{V}(\hat{\beta}_j)}$ to a sample from a Student law of parameter $(n-p)$. Hence the following algorithm starts with the full model and $p_0 = p$:

1. Compute the current complete model estimate $\hat{\beta}$
2. Test of the lowest T_j associated with its p_k components
   - If $T_j \leq t_{n-p_k}(1-\alpha)$, removal of the j-th regressor X_j from the current regressor matrix, set $p_{k+1} = p_k - 1$ and return to 2.
   - Else stop and accept the current reduced model,

where $t_{n-p}(1-\alpha)$ is the quantile fonction of the Student law of parameter $(n-p)$ and $(1-\alpha)$ is the confidence parameter. In this test, the stress is on the value of $\hat{\beta}$ components. If the value is in a confidence interval, it is declared to come from a zero component of β^*.

2.2. Backward Elimination procedure with the Fisher test

In the Fisher test the emphasis is put on the regressors influence. The idea is to find the regressor whose elimination leads to the smallest increase in the residual sum of squares, and to eliminate it if this increase is not significantly larger than the noise variance. The Fisher-Snedecor law with parameters (k,l) is the law of $(U/k)/(V/l)$ where $U \sim \chi^2_k$ and $V \sim \chi^2_l$ are independent. Using the statistical properties of the LS estimate, we already indicated that $\frac{RSS}{\sigma^2} = (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Now let us introduce RSS_j, the residual sum of squares obtained after removal of X_j the j-th column in X

$$RSS_j = \min_{\beta} \|y - X\beta\|^2 \text{ subject to } \beta_j = 0,$$

if $\beta^*_j = 0$, $\frac{RSS_j - RSS}{\sigma^2} \sim \chi^2_1$ and the random variable

$$F_j = \frac{RSS_j - RSS}{RSS/(n-p)}$$

follows a Fisher-Snedecor law of parameters $(1, n-p)$. The BEP based on the Fisher test is build around F_j. It is an iterative algorithm that proceeds as in the Student procedure described above with step 2 therein replaced by

2. Compute $F_j = \frac{RSS_j - RSS}{RSS/(n-p)}$ where RSS_j is the lowest Residual Sum of Squares obtained by removing one of regressors from the current model
   - If $F_j \leq f_{1,n-p}(1-\alpha)$, removal of the corresponding regressor and variable associated and return to 1.
   - Else stop and accept the current model,

where $f_{1,n-p}(1-\alpha)$ is the quantile fonction of the Fisher-Snedecor law of parameters $(1, n-p)$ and $(1-\alpha)$ are the confidence parameter. Due to lack of space, we do not establish that at a given step in both iteration procedures, one actually has $T_j^2 = F_j$ and thus strict equivalence of both procedure for identical α's. This can be established using for instance, results concerning the pseudo-inverse of partitionned matrices, to link $\hat{\beta}$ and the optimal parameter vector obtained when X_j is deleted, and similarly RSS and RSS_j.

3. THE TLS MODEL

3.1. The new model

In model (2), we introduce the possibility for the regressors to be corrupted by additive noise as well as the observations. For obvious practical reasons[1], we generalize this model to the case where some regressors are perfectly known (so far the standard assumption) those in the T matrix below, and the others, those in the X matrix, are subject to noise. The model (2) thus becomes:

$$\begin{aligned} y &= X\beta^* + T\beta^*_0 + e \\ Z &= X + E, \end{aligned} \tag{4}$$

which can be rewritten as, isolating row i:

$$\begin{aligned} y_i &= X_i^T\beta^* + T_i^T\beta^*_0 + e_i \\ Z_i &= X_i + E_i. \end{aligned}$$

We assume that the columns of the noisy matrix E are independent. The errors e_i and E_i are then independent Gaussian noises with variance σ^2 and covariance matrix $\sigma^2\Sigma^2$ respectively (one needs to assume that Σ^2 is known since it is not identifiable). The ML estimates of the parameters are part of the solution of:

$$\min_{\beta_0,\beta,X_i} \sum_i (y_i - X_i^T\beta - T_i^T\beta_0)^2 + (Z_i - X_i)^T\Sigma^{-2}(Z_i - X_i),$$

[1] In multiple linear regression one systematically introduces a column of ones in the X matrix to allow for an intercept parameter. This is but one example of perfectly known regressor.

or in matrix notations

$$\min_{\beta_0,\beta,X} \|y - X\beta - T\beta_0\|_2^2 + \|(Z - X)\Sigma^{-1}\|_F^2,$$

where $\|X\|_F^2 = \sum_{i,j} x_{i,j}^2 = \text{trace}\,(XX^T)$ denotes the square of the Frobenius norm of X. The problem is separable and the minimum with respect to β_0 is attained at: $\hat{\beta}_0 = T^+(y - X\beta)$. Before we replace $\hat{\beta}_0$ by its optimum, let us define $y^{\|} = TT^+y$ the projection of the column vector y on the range space of T and $y^{\perp} = (I - TT^+)y$ the projection on the orthogonal of the range of T. We define similarly $X^{\|}$, $X^{\perp}$, $Z^{\|}$ and $Z^{\perp}$. Then, after subtitution we get:

$$\min_{\beta,X^{\perp},X^{\|}} \|Y^{\perp} - X^{\perp}\beta\|_2^2 + \|(Z^{\perp} - X^{\perp})\Sigma^{-1}\|_F^2 + \|(Z^{\|} - X^{\|})\Sigma^{-1}\|_F^2. \quad (5)$$

Here again, the minimum with respect to $X^{\|}$ is attained at: $X^{\|} = Z^{\|} = TT^+Z$, and the last term in (5) vanishes. We now introduce Q a full column rank orthogonal matrix that is such that $I - TT^+ = QQ^T$. Thus (5) becomes:

$$\min_{\beta,Q^TX} \|Q^T(y - X\beta)\|_2^2 + \|Q^T(Z - X)\Sigma^{-1}\|_F^2.$$

Defining $\bar{e} = Q^Te = Q^T(y - X\beta)$, $\bar{W} = Q^T(Z - X)\Sigma^{-1}$ and $\bar{\beta} = \Sigma\beta$, this problem is in turn equivalent to:

$$\min_{\bar{\beta},\bar{e},\bar{W}} \|\bar{e}\|_2^2 + \|\bar{W}\|_F^2 \quad \text{s.t.} \quad Q^Ty - \bar{e} = (Q^TZ\Sigma^{-1} - \bar{W})\bar{\beta}.$$

The purpose of these transformations is to reduce the number of unknowns to their minimum, making them independent. In this last form, it is now possible to recognize the standard formulation that links the TLS problem with the Singular Value Decomposition [7, 8]. One seeks the minimal Frobenius norm perturbation matrix say $\bar{\Delta} = [\bar{W} \ : \ \bar{e}]$ that makes

$$Q^Ty - \bar{e} \in \text{Range}(Q^TZ\Sigma^{-1} - \bar{W}),$$

The optimum [7, 8] is a rank one matrix built using the smallest singular triplet of the matrix $\hat{\bar{C}} = [Q^TZ\Sigma^{-1} \ : \ Q^Ty]$. The optimal $\bar{\beta}$, the ML estimate of $\Sigma\beta^*$, is deduced from the *smallest* right singular vector v_{p+1} by the following relation

$$[\,\bar{\beta}^T \ : \ -1] = -v_{p+1}^T/v_{p+1}(p+1), \quad (6)$$

i.e., one normalizes the last component of v_{p+1} to -1.

3.2. The TLS Backward Elimination procedure

The idea is to use a backward approach with the student Test similar to the one used in the LS case when testing the components. In the sequel, to simplify the presentation, we will suppose that $\Sigma = I$ and that there are no regressor stricly known in (4) except for the constant regressor, we denote T, corresponding to a column of ones which allows to capture the intercept parameter that is considered to be systematically present in any multiple linear regression model [3]. To take care of this peculiarity, one projects (see (5)) the observed matrices on the orthogonal of the range space of T by using $P = I - \frac{1}{n}TT^T = QQ^T$. And the ML parameter estimate $\hat{\beta}$ is deduced (see (6)) from the *smallest* or *minimal* right singular vector of the matrix $\hat{C} = [Q^TZ \ : \ Q^Ty]$. To assess its statistical properties one can either, say ML techniques and compute the Fisher Information matrix [10] or use results from matrix perturbation theory. We adopt this second approach with the following perturbation model $\hat{C} = C + \sigma\Delta$ where C is the exact but unknown underlying matrix $C = [Q^TX \ : \ Q^TX\beta]$ and $\sigma\Delta = [Q^TE \ : \ Q^Te]$. The common standard deviation σ has been factorized to highlight the fact that it is this quantity that has to be small for the perturbation results to hold. Let us introduce the following notations: $C = USV^T$ and $\hat{C} = \hat{U}\hat{S}\hat{V}^T$ denote the SVD of C and $\hat{C}$ respectively with $U, \hat{U}, V$ and $\hat{V}$ square orthogonal matrices and $S, \hat{S}$ matrices of the dimensions of C:

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & s_{p+1} \\ 0 & 0 \end{bmatrix}, \quad \hat{S} = \begin{bmatrix} \hat{S}_1 & 0 \\ 0 & \hat{s}_{p+1} \\ 0 & 0 \end{bmatrix},$$

where S_1 and $\hat{S}_1$ order-p diagonal matrices. $S_1 = \text{diag}(s_1, s_2, ..., s_p)$ and $s_1 \geq s_2 \geq \ ... \ \geq s_p > s_{p+1} = 0$. Note that C, the exact matrix has one zero singular value s_{p+1} and that we are precisely interested in the way its right singular vector is perturbed. For small σ one can expect that the difference between $\hat{v}_{p+1}$ and v_{p+1} is of the order of σ. This is indeed true and one can establish [11, 12, 13, 14] the following results.

Proposition 2 *If $s_p \gg \sigma\sqrt{n-p}$, then:*

- $E(\hat{v}_{p+1}) = v_{p+1} + O(\sigma^2)$ *no first order bias,*
- $E(\tilde{v}\tilde{v}^T) = \sigma^2(V_1S_1^{-2}V_1^T) + O(\sigma^3)$ *with* $\tilde{v} = \hat{v}_{p+1} - v_{p+1}$,
- $E(\hat{s}_{p+1}^2) = (n-p)\sigma^2 + O(\sigma^3)$,
- $\hat{v}_{p+1}$ *and* $\hat{s}_{p+1}$ *are independent,*

□

where V_1 is such as $C = USV^T = U_1S_1V_1^T$. Similarly to the results of the LS case where $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/(n-p)$ is an unbiased estimate of σ, in the TLS case, an unbiased estimate of the common variance σ^2 is given by $\hat{\sigma}^2 = \hat{s}_{p+1}^2/(n-p)$. It follows that

$$T_j = \frac{\hat{v}_{p+1}(j) - v_{p+1}(j)}{\hat{\sigma}(V_1S_1^{-2}V_1^T)_{j,j}^{\frac{1}{2}}}$$

follows a Student law of paramater $(n-p)$. It is then logical to propose the same algorithm than in LS case with the Student test. Yet as we don't have access to the C matrix, we remplace V_1 and S_1 by $\hat{V}_1$ and $\hat{S}_1$ in T_j. Note also that since a normalization is performed, it is equivalent to test $\hat{\beta}(j)$ or $\hat{v}_{p+1}(j)$.

4. EXPERIMENTAL RESULTS

Results are presented for both the basic LS backward elimination procedure and our approach, applied to two sorts of data: simulated data and the Hald data set. In both cases, the $T = \mathbf{1}$ regressor is assumed to participate to the regression and consequently it will not be tested.

In the simulated case, the data set corresponds to $n = 15$ observations and $p = 7$ possible regressors including the 3 regressors which describe effectively the regression. The components of the X matrix are independent samples from a $N(0, I)$. Gaussian noise was added to both X and y according to the TLS model assumptions, making these simulations favorable to our approach. The tests were done with a Student test with signifiance level at 10 %. The TLS-backward elimination procedure (TLS-BEP) retrieves systematically the true selection when the noise

level satisfies the assumptions given in Proposition 2. The LS-backward elimination procedure generally retains too many variables. A typical result is the following where the true selection is $(X_2,\ X_4,\ X_5)$.

Simulated Data	X_2, X_4, X_5	
	Selected Data	Ordered removed Data
LS-BEP	X_2, X_4, X_5, X_7	X_1, X_6, X_3
TLS-BEP	X_2, X_4, X_5	X_1, X_6, X_3, X_7

Note that both approaches reject the regressors in the same order but the LS-BEP stops earlier and keeps a wrong regressor in the selection.

We now present some results using real data: the Hald data set [3]. It is a data set with $n = 13$ observations and $p = 4$ regressors. The tests are done with a Student test for significance at the 10 % level. The exact model is of course unknown. But since the measures made to get the components of the regressors are of the same kind as those made to get the observations, the TLS model seems quite natural and justified. Remember that in the TLS model (4) the noise variance on the measurements y is denoted σ^2 and the noise on the rows X_i^T of the regressor matrix X is $\sigma^2\Sigma^2$. In the sequel, we take $\Sigma^2 = \frac{I_p}{d^2}$ so that d^2 represents the ratio between the observation noise variance and the common regressor noise variance (as $d \to \infty$, one tends towards the LS model). Three different cases have been encountered. We can note that the results obtained by the LS-BEP are of course unaffected by these modifications and remain the same.

Case 1: $1 \leq d \leq 3$

	Selected Data	Ordered removed Data
LS-BEP	X_1, X_2	X_3, X_4
TLS-BEP	X_1, X_2, X_3, X_4	

In this case, the different noises are of similar magnitudes. The BEP declares two variables as unrelevant whereas the TLS-BEP selects all the regressors.

Case 2: $4 \leq d \leq 18$

	Selected Data	Ordered removed Data
LS-BEP	X_1, X_2	X_3, X_4
TLS-BEP	X_1, X_2	X_4, X_3

The case is very interessant because both algorithms remove the same variables but not in the same order. If we suppose that the perturbations on the regressors are small but still significant enough, the X_4 regressor is considered as the least relevant regressor in the TLS approach whereas it is X_3 for the classical BEP.

Case 3: $d \geq 19$

	Selected Data	Ordered removed Data
LS-BEP	X_1, X_2	X_3, X_4
TLS-BEP	X_1, X_2	X_3, X_4

In this third case both algorithms find the same result. Indeed for d large, the TLS model is close to the LS model. It is then logical to have the same results.
More investigations on different dataset need to be performed to assess the potentialities of the proposed approach.

5. CONCLUSION

The Backward Elimination Procedure is a useful and powerful algorithm that allows to select variables in multiple linear regression schemes. It has been developped and is used for Least Squares models where it is assumed that (Gaussian) errors are only present in the observation vector y that one wants to explain.

In many practical situations, the components of some regressors X_j are essentially of the same nature as the components of the observation vector y. There is thus no reason to consider these regressors to be known exactly as is done in the standard Least Squares model. More generally it seems natural, depending upon the type of the regressors to consider that some are subject to noise while others (e.g. those having integer values) are known exactly and to develop selection procedures that allow to take this possibility into account.

The multiple linear regression model where Gaussian noise is assumed to be present not only on the observations but also on part of the regressors is known as the Total Least Squares (TLS) model. We have developped the Backward Elimination procedure for this type of models. For this we have analysed the statistical properties of the corresponding Maximum Likelihood parameter vector estimate using results from matrix perturbation analysis and developed a Student test that allows to decide if a component of the estimated vector should be declared equal to zero.

It is also possible to develop a Fisher test and, besides performing intensive results evaluations. We plan to analyse the advantages or disadvantages of both approaches.

6. REFERENCES

[1] L. Wilkinson and G.E. Dallal, Tests of Significance in Forward Selection Regression with an F-to-enter Stopping Rule, *Technometrics*, 23, 377-380, 1981.

[2] C.L. Mallows, Some comments on C_p. *Technometrics*, 15, 661-675, 1973.

[3] N.Draper and H.Smith, Applied regression analysis. *Wiley.*, 1981.

[4] A. Miller, Subset Selection in Regression. *Chapman and Hall, London*, 1990.

[5] M.L. Thompson, Selection of variables in multiple regression, Part I and II. *Internat. Statist. Rev.*, 46, 1-19, 129-146, 1978.

[6] W.A. Fuller, Measurement Errors Models *Wiley and Sons, New York*, 1987.

[7] G.H. Golub and C.F. Van Loan, An analysis of the Total Least Squares problem. *SIAM J. Numer. Anal.*, vol. 17, 6, 883-893, 1980.

[8] S. Van Huffel and J. Vanderwalle. The Total Least Squares problem. *SIAM* , 1991 .

[9] V. Voievodine. Principes numeriques d'algebre lineaire. *Edition Mir, Moscou*, 1980.

[10] L.J. Gleser, Estimation in a multivariate "errors in variables" regression model: large sample results. *The Annals of Statistics*, vol. 9, 1, 24-44, 1981.

[11] J.H. Wilkinson, The algebraic Eigenvalue Problem. *Oxford Univ. Press, Oxford*, 1965.

[12] J.J. Fuchs, Rectangular Pisarenko Method Applied to Source Localization. *IEEE Trans. on S.P.*, vol. 44, 10, october 1996.

[13] G.H. Golub and C.F. Van Loan. Matrix Computations. *John Hopkins University Press*, 1983.

[14] G.W. Stewart and J.G. Sun, Matrix Pertubation Theory. *Academic Press*, 1990.