Selection amongst Alternative Variables in Regression Analysis

K. Pelckmans, J.A.K. Suykens, B. De Moor
K.U.Leuven - ESAT - SCD/SISTA
Kasteelpark Arenberg 10, B-3001, Leuven (Heverlee), Belgium
kristiaan.pelckmans@esat.kuleuven.ac.be
http://www.esat.kuleuven.ac.be/sista/lssvmlab

Abstract

This paper presents an approach to the task of regression analysis by imposing a dedicated sparseness pattern is imposed on the solution. It imposes the estimate to be expressed in terms of only one single variable amongst the proposed set of alternative variables. The task of Selection amongst Alternative variables in Regression Analysis (SARA) is elaborated for linear as well as kernel based methods using the primal-dual argument according to optimization theory. Selection of inputs amongst alternative variables is achieved through the use of a dedicated regularization scheme by shrinking the active set of variables gradually towards a singleton.

Keywords: Feature Selection , Sparseness, Shrinkage, Kernel Methods.

1 Introduction

Selection of variables for regression and classification problems is a problem with a long research history. Classical statistics often takes a strategy based on significance testing as heavily employed in the method of ANOVA, see e.g. [12, 15]. A disadvantage there is the sensitivity of the methodology on the Gaussian assumption and the parametric model [15]. The backward and forward selection strategy then constitute general computer-intensive approaches, see e.g. [9]. Recent advances in Bayesian statistics propose methods as Automatic Relevance Determination often resulting in heavy non-convex optimization problems, see e.g. [10]. Recent advances in the field of learning theory and machine learning were reported in a special issue [7].

An alternative perspective to the problem was found in the method of regularization. While Tikhonov regularization schemes based on the 2-norm of the parameters are commonly used in order to improve estimates (statistically as well as numerically), interest in L_1-based regularization schemes has emerged recently as seen in the formulation and study of LASSO (Least Absolute Shrinkage and Selection Operator) estimators [16], SURE (Stein Unbiased Risk Estimator) [4, 1] and basis pursuit [6, 3] algorithms. Least angle regression [5] then analyses the complete solution path of the LASSO estimator for all regularization constants. Sparseness in the resulting estimate of the parameter vector is often regarded as a form of structure detection and can be used as (an initialization in) a variable selection procedure. A non-convex extension to the use of the zero-norm was described in [17].

The contribution of this paper is twofold. Section 2 discusses the mechanism of shrinkage to a single variable in the context of linear parametric models using a least squares approach. Section 3 then continues the discussion to a setting of non-parametric additive kernel machines by relying on the technique of convex optimization with a primal and dual model representation. Section 4 illustrates the practical relevance of the approach by relating it to LASSO estimators and assessing the performance of the method on a benchmark dataset.

2 Selection between Alternatives in Least Squares Regression

Given a set of observed input/output data-samples $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}$. Assume the model underlying the dataset $\mathscr{D}$ takes a linear parametric form

$$f(x) = \omega^T x, \tag{1}$$

where $\omega \in \mathbb{R}^D$ denotes the parameter vector of the model. Let the data-samples satisfy the relation $y_i = \omega^T x_i + \varepsilon_i$ for all $i = 1, \ldots, N$ with $\{\varepsilon_i\}_{i=1}^N \subset \mathbb{R}$ a set of i.i.d. error terms. For notational convenience, this paper does not include an intercept term in the derivations and a proper normalization of the data is assumed.

The task of estimating the optimal predictor based on a single variable amongst given alternatives is considered. Formally, one solves the problem

$$\min_w \mathscr{J}(w) = \sum (w^T x_i - y_i)^2 \quad \text{s.t.} \quad w_i\, w_j = 0,\ \forall i, j = 1, \ldots, D,\ i \neq j, \tag{2}$$

using the following related formulation:

Lemma 1 *(Selection amongst Alternative variables in Regression Analysis, SARA) The solution to the problem*

$$(\hat{w}, \hat{t}) = \arg\min_{w,t} J(w) = \frac{1}{2} \left\| w^T x_i - y_i \right\|_2^2 \quad s.t. \quad \begin{cases} t^T\ 1_{D\times D}\, t \leq w^T w \\ -t_i \leq w_i \leq t_i \end{cases} \quad \forall i = 1, \ldots, D \tag{3}$$

where $1_{D\times D} \in \mathbb{R}^{D\times D}$ contains all ones, searches for on optimal parameter estimate in 2-norm such that the constraints in (2) are satisfied.

Proof: Let $X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N\times D}$ and $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ be vectors. The Lagrangian of the constrained optimization problem (3) becomes $\mathscr{L}(w, t; \lambda, \alpha^+, \alpha^-)$ $= \frac{1}{2} \|Xw - Y\|_2^2 + \sum_{i=1}^D \alpha_i^-(-t_i - w_i) + \sum_{i=1}^D \alpha_i^+(-t_i + w_i) + \frac{\lambda}{2} \left(t^T\ 1_{D\times D}\, t - w^T w\right)$ where $\alpha^+, \alpha^- \in \mathbb{R}^{+,D}$ and $\lambda \in \mathbb{R}^+$ are positive Lagrange multipliers. Let $1_D \in \mathbb{R}^D$ denote the vector containing ones. The first order (necessary) conditions for optimality are given by the Karush-Kuhn-Tucker conditions (KKT), see e.g. [2]:

$$\begin{cases} \left(X^T X - \lambda I_D\right) w - X^T Y = \alpha^- - \alpha^+ & (a) \\ \alpha_i^- + \alpha_i^+ = \lambda 1_D^T\, t \qquad \forall i = 1, \ldots, D & (b) \\ -t_i \leq w_i \leq t_i \qquad \forall i = 1, \ldots, D & (c) \\ \alpha_i^+, \alpha_i^- \geq 0 \qquad \forall i = 1, \ldots, D & (d) \\ \alpha_i^-(t_i + w_i) = 0 \qquad \forall i = 1, \ldots, D & (e) \\ \alpha_i^+(t_i - w_i) = 0 \qquad \forall i = 1, \ldots, D & (f) \\ t^T\ 1_{D\times D}\, t \leq w^T w & (g) \\ \lambda \geq 0,\ \ \lambda(t^T\ 1_{D\times D}\, t - w^T w) = 0 & (h) \end{cases} \tag{4}$$

where the equalities (4.efh) are referred to as the complementary slackness constraints. By combining conditions (4.ef) and (4.bcd), it follows that $t_i = |w_i|$ for all $i = 1,\ldots,D$. From condition (4.g) it then follows that

$$t^T\ 1_{D\times D}\, t \leq w^T w = t^T t \ \ \Rightarrow\ \ t^T\,(1_{D\times D}\ - I_D)\, t \leq 0. \tag{5}$$

As the vector t and the matrix $(1_{D\times D} - I_D)$ contains all positive numbers, only $t^T\,(1_{D\times D}\ - I_D)\, t = 0$ is to be considered. As such, conditions (2) are satisfied in (any) optimum to (3). This concludes the proof. □ The relationship with the least squares estimator when the relevant variable were known beforehand is given in the following lemma.

Corollary 1 *(Relation to Univariate Least Squares) Let γ^* satisfy the KKT conditions (4) such that $(X^T X - \lambda^* I_D) \succeq 0$ and that the constraint $t^T\ 1_{D\times D}\, t \leq w^T w$ is satisfied, then the solution to (3) corresponds with the univariate least squares estimator based on only the variable with corresponding nonzero parameter.*

proof: Let the solution to (3) have finally one variable denoted as $X_{(1)} \in \mathbb{R}^N$. Let $X_{(0)} \in \mathbb{R}^{N\times(D-1)}$ be a matrix denoting all other candidate variables. Condition (4.a) can then be rewritten as

$$\begin{bmatrix} (X_{(1)}^T X_{(1)} - \lambda) & X_{(1)}^T X_{(0)} \\ X_{(0)}^T X_{(1)} & (X_{(0)}^T X_{(0)} - \lambda I_{D'\times D'}) \end{bmatrix} \begin{bmatrix} w_{(1)} \\ w_{(0)} \end{bmatrix} = \begin{bmatrix} X_{(1)}^T Y \\ X_{(0)}^T Y \end{bmatrix} + \begin{bmatrix} \alpha_{(1)}^- - \alpha_{(1)}^+ \\ \alpha_{(0)}^- - \alpha_{(0)}^+ \end{bmatrix} \tag{6}$$

where the parameters $w_{(1)} \in \mathbb{R}$ and $w_{(0)} \in \mathbb{R}^{D-1}$ relate to $X_{(1)}$ and $X_{(0)}$ respectively. In case the parameters $w_{(0)}$ are zero and $w_{(1)}$ is nonzero, the following property holds

$$\left(X_{(1)}^T X_{(1)}\right) w_{(1)} = X_{(1)}^T Y, \tag{7}$$

as $\alpha_{(1)}^+ - \alpha_{(1)}^- = \lambda w_{(1)}$ from application of (4.bef) and the property that $|w_{(1)}| = 1^T t$ in the solution to (4). Then note that (7) corresponds with the normal equations of the least squares problem $\min_w \|X_{(0)} w - Y\|_2^2$. In case all estimates equal zero, the Lemma follows trivially as $\alpha^- + \alpha^- = 0_D$. □ This result is strongly related to the derivations of oracle inequalities as studied thoroughly in [4, 1].

2.1 A convex approach

So far, we did not discuss the choice of the Lagrange multiplier λ satisfying (4.g). However, it turns out that the global optimum can be computed efficiently in many cases. A related optimization problem to (3) with constant $\gamma > 0$ becomes

$$(\hat{w},\hat{t}) = \arg\min_{w,t} J_\gamma(w) = \frac{1}{2}\|Xw - Y\|_2^2 + \frac{\gamma}{2}\left(t^T\ 1_{D\times D}\, t - w^T w\right)$$
$$\text{s.t.} \quad -t_i \leq w_i \leq t_i \quad \forall i = 1,\ldots,D, \tag{8}$$

which is convex as long as $(X^T X - \gamma I_D)$ is symmetric positive semi-definite [2]. Note that the KKT conditions corresponding to (3) and (8) equal when γ is chosen equal to $\hat{\lambda}$ solving (4). Furthermore, if $\gamma \geq \hat{\lambda}$ where λ solves (4) and $(X^T X - \lambda I_D)$ is positive semi-definite, it follows also from the KKT conditions that the solution to the problems associated with constants λ and γ are equal.

This results in a practical algorithmic approach to estimate the solution to the original problem (3) if it is unique. Hereto, let σ^- denote the smallest eigenvalue of the sample covariance matrix $X^T X$. Then it is easily seen that $\gamma = \sigma^-$ is the largest value for which the problem (8) is convex. If not so, the problem (8) is not convex and one can use local optimization strategies to search the global solution. Note that if the covariance matrix is ill-conditioned (σ^- small), almost no freedom is allowed to apply the above result.

3 Primal-Dual Kernel Machines

3.1 Additive Models and Structure Detection

Additive models are defined as follows [8]:

Definition 1 *(Additive Models) Let the i-th observation consists of multiple components of dimension D_q. Let this be denoted as $x_i = \left(x_i^{(1)}, \ldots, x_i^{(Q)}\right)$ with $x_i^{(q)} \in \mathbb{R}^{D_q}$ (in the simplest case $n_q = 1$ such that $x_i^{(q)} = x_i^q$). The class of additive models using those components is defined as*

$$\mathscr{F}^Q = \left\{ f : \mathbb{R}^D \to \mathbb{R} \;\middle|\; f(x) = \sum_{q=1}^{Q} f_q\left(x^{(q)}\right) + b, \right.$$
$$\left. f_q : \mathbb{R}^{D_q} \to \mathbb{R}, b \in \mathbb{R}, \forall x = \left(x^{(1)}, \ldots, x^{(Q)}\right) \in \mathbb{R}^D \right\}. \quad (9)$$

Let furthermore X^q denote the random variable (vector) corresponding to the q-th component for all $q = 1, \ldots, Q$.

Extensions towards primal-dual kernel machines and the related fast implementation is described in [13]. In [14], the use of the following criterion is proposed:

Definition 2 *(Maximal Variation) The maximal variation of a function $f_q : \mathbb{R}^{D_q} \to \mathbb{R}$ over a distribution $P_{X^{(q)}}$ is defined as follows*

$$\mathscr{M}_q = \sup_{x^{(q)} \in P_{X^{(q)}}} \left| f_q\left(x^{(q)}\right) \right|, \quad (10)$$

for all $x^{(q)} \in \mathbb{R}^{D_q}$ sampled from the distribution P_{X^q} corresponding to the q-th component. The empirical maximal variation can be defined as

$$\hat{\mathscr{M}}_q = \max_{x_i \in \mathscr{D}_N} \left| f_q\left(x_i^{(q)}\right) \right|, \quad (11)$$

with $x_i^{(q)}$ denoting the q-th component of a the sample x_i.

A formal proof that this empirical measure indeed contains information on the theoretical maximal variation on the range of the training set remains to be given. This work focuses on how to formulate kernel machines using this measure. Structural limits can be obtained by using results described in [11].

Given those definitions, one can express the task of structure detection as the task of detecting which components in the additive model can be driven to zero without affecting the global risk of the model too much. The measure of (empirical) maximal variation then quantifies the expected maximal contribution of each component on the dataset.

Proposition 1 *(Selection between Alternatives) Let the matrix $N_Q \in \mathbb{R}^{Q \times Q}$ be defined as*

$$N_Q = \begin{bmatrix} 0 & 1 & \ldots & 1 \\ 1 & 0 & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \ldots & 1 & 0 \end{bmatrix}. \quad (12)$$

Let $t = \left(\hat{\mathscr{M}}_1, \ldots, \hat{\mathscr{M}}_q\right) \in \mathbb{R}^Q$ denote the vector collecting the empirical maximal variations for all components. Then $t^T N_Q t = 0$. is a necessary and sufficient condition for the model to be using only one component.

The inverse N_Q^{-1} to the matrix such that $N_Q^T N_Q^{-1} = N_Q^{-1^T} N_Q = I_Q$ can be found analyticaly. Remark that similar matrices can be designed to enforce the use of a series of variables with only a part restricted to a single alternative. This path is further elaborated in the examples.

3.2 Selection between alternatives in Primal-Dual Kernel Machines

Adopting again the weighting procedure where the quadratical constraint is integrated in the cost-function with a predefined weighting term $\lambda > 0$, the following primal problem formulation is obtained where a least squares criterion is used:

$$\begin{aligned} \mathscr{J}_{\gamma,\lambda}(w_q, e_i, b, t) &= \frac{1}{2}\sum_{q=1}^{Q} w_q^T w_q + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 + \frac{\lambda}{2} t^T N_Q t \\ \text{s.t.} \quad & \begin{cases} \sum_{q=1}^{Q} w_q^T \varphi_q\left(x_i^{(q)}\right) + b + e_i = y_i & \forall i = 1,\ldots,N \\ -t_q \leq w_q^T \varphi_q\left(x_i^{(q)}\right) \leq t_q & \forall i = 1,..,N, \forall q = 1,..,Q \end{cases} \end{aligned} \tag{13}$$

where $\varphi_q(\cdot) : \mathbb{R}^{D_q} \to \mathbb{R}^{D_\varphi^q}$ denotes the feature space mapping per component. Its dual problem enables the use of the kernel trick as follows:

Lemma 2 *The dual problem to (13) is given by*

$$\begin{aligned} \max_{\alpha,\rho^+,\rho^-} \mathscr{J}^D_{\gamma,\lambda}(\alpha, \rho^-_{qN}, \rho^+_{qN}) &= \frac{1}{2}\sum_{q=1}^{Q} (\alpha + \rho^-_{qN} - \rho^+_{qN})^T \Omega_q (\alpha + \rho^-_{qN} - \rho^+_{qN}) + \frac{1}{2\gamma}\sum_{i=1}^{N} \alpha_i^2 \\ + \frac{1}{2\lambda}\sum_{q,r=1}^{Q} \left(\rho^+_{qN} + \rho^-_{qN}\right)^T 1_N N^{-1}_{Q,qr} 1_N^T \left(\rho^+_{rN} + \rho^-_{rN}\right) - \alpha^T Y \quad s.t. \quad & \begin{cases} \rho^-_{qi}, \rho^+_{qi} \geq 0 & \forall i, q \\ \sum_{i=1}^{N} \alpha_i = 0 \end{cases} \end{aligned} \tag{14}$$

where one defines $\Omega_{q,ij} = K_q\left(x_i^{(q)}, x_j^{(q)}\right)$*, the vector* $\rho^+_{qN} = \left(\rho^+_{1q}, \ldots, \rho^+_{nq}\right)^T \in \mathbb{R}^N$ *and* ρ^+_{qN} *analogously. The estimated function can be used for predicting the outcome as follows*

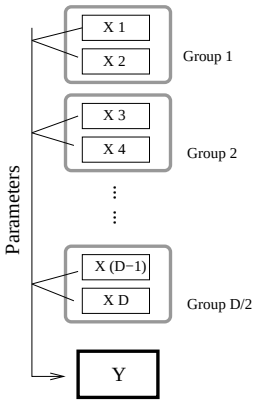
$$\hat{f}(x_*) = \sum_{q=1}^{Q}\sum_{i=1}^{N} (\hat{\alpha}_i + \hat{\rho}^-_{qi} - \hat{\rho}^+_{qi}) K_q\left(x_i^{(q)}, x_*^{(q)}\right) + \hat{b}, \tag{15}$$

where $\hat{b}, \hat{\alpha}, \hat{\rho}^+$ *and* $\hat{\rho}^-$ *are the solution to (14).*

Proof: The Lagrangian of the problem becomes

$$\begin{aligned} \mathscr{L}_{\gamma,\lambda}(w_q, b, e_i, t; \alpha, \rho^-, \rho^-) = \mathscr{J}_{\gamma,\lambda}(w_q, e_i, b, t) - \sum_{i=1}^{N} \alpha_i \left(\sum_{q=1}^{Q} w_q^T \varphi_q\left(x_i^{(q)}\right) + b + e_i - y_i\right) \\ - \sum_{q=1}^{Q}\sum_{i=1}^{N} \rho^+_{qi}\left(t_q + w_q^T \varphi_q\left(x_i^{(q)}\right)\right) - \sum_{q=1}^{Q}\sum_{i=1}^{N} \rho^-_{qi}\left(t_q - w_q^T \varphi_q\left(x_i^{(q)}\right)\right). \end{aligned} \tag{16}$$

The minimum of the Lagrangian with respect to the primal variables w_q, e_i, b and t_q is obtained when the following conditions for optimality hold $\frac{\partial \mathscr{L}_{\gamma,\lambda}}{\partial w_q} = 0 \to w_q = \sum_{i=1}^{N} \left(\alpha_i + \rho^-_{qi} - \rho^+_{qi}\right) \varphi_q\left(x_i^{(q)}\right)$ for all $q = 1,\ldots,Q$; $\frac{\partial \mathscr{L}_{\gamma,\lambda}}{\partial e_i} = 0 \to \gamma e_i = \alpha_i$ for all $i =$



(a)

	RR	LASSO	SARA
	D=10		
mean[$\|\omega-\hat{w}\|_2$]	0.0052	0.0649	0.0022
std[$\|\omega-\hat{w}\|_2$]	0.0026	0.0908	0.0012
Mean[Struct]%	0.5081	0.6253	0.9636
	D=20		
mean[$\|\omega-\hat{w}\|_2$]	0.0066	0.0766	0.0018
std[$\|\omega-\hat{w}\|_2$]	0.0026	0.1071	0.0008
Mean[Struct]%	0.5038	0.6141	0.9308
	D=40		
mean[$\|\omega-\hat{w}\|_2$]	0.0128	0.0839	0.0025
std[$\|\omega-\hat{w}\|_2$]	0.0036	0.1327	0.0010
Mean[Struct]%	0.5011	0.6072	0.9061

(b)

Figure 1: **(a)** *Schematical representation of the imposed sparseness pattern: each group with a consecutive pair of variables may contain only one non-zero parameter.* **(b)** *Numerical results on an artificial dataset as described in Subsection 4.1. Given a dataset of $N = 80$ and $D = 10, 20$ and 40 respectively, the performance of the estimate of Ridge Regression (RR), LASSO and the proposed method SARA method is expressed as $\|\omega - \hat{w}\|_2$. The regularization constants in RR and LASSO are tuned using 10-fold cross validation. The ability to recover the structure is expressed as the percentage of the inputs where the sparseness corresponds with the true structure. From the Monte Carlo iteration, a serious increase in both performance as well as ability to recover the underlying structure can be noted. In this toy example, the method is robust when increasing the dimensionality D.*

$$1,\ldots,N;\ \frac{\partial \mathscr{L}_{\gamma,\lambda}}{\partial b}=0 \rightarrow \sum_{i=1}^{N}\alpha_i=0\ \frac{\partial \mathscr{L}_{\gamma,\lambda}}{\partial t_q}=0 \rightarrow \sum_{p\neq q} t_p=\sum_{i=1}^{N}(\rho_{qi}^{+}+\rho_{qi}^{-}) \text{ for all } q=1,\ldots,Q.$$

By application of the kernel trick $\varphi_q\left(x_i^{(q)}\right)^T \varphi_q\left(x_j^{(q)}\right) \triangleq K_q\left(x_i^{(q)}, x_j^{(q)}\right)$, elimination of the primal variables w_q, e_i, b, t_q and using the inverse matrix $N_Q^{-1} \in \mathbb{R}^{Q\times Q}$ the dual problem (14) follows. □

4 Illustrative Examples

4.1 Artificial dataset: linear model

An artificial example is given, relating the proposed approach with existing methods and illustrating its use. Given is a dataset containing D input variables and one output variable. Suppose one is interested in a model using exactly one variable of each disjunct pair of consecutive input variables. This user-defined sparseness pattern can be imposed on the estimated model using the technique of SARA as illustrated in Figure 1.

The dataset is constructed as follows. A set of D inputs are sampled independently from a uniform distribution with $N = 80$ and $D = 10, 20$ and 40 respectively. The outputs are generated using a parameter vector where each consecutive disjunct pair (e.g. $(w_1, w_2), (w_3, w_4), \ldots$) contains exactly one zero value and one non-zero random value between -2 and 2. The outputs contain additional noise terms sampled from a standard distribution. The results of a Monte Carlo experiment using 1000 iterations is given. From

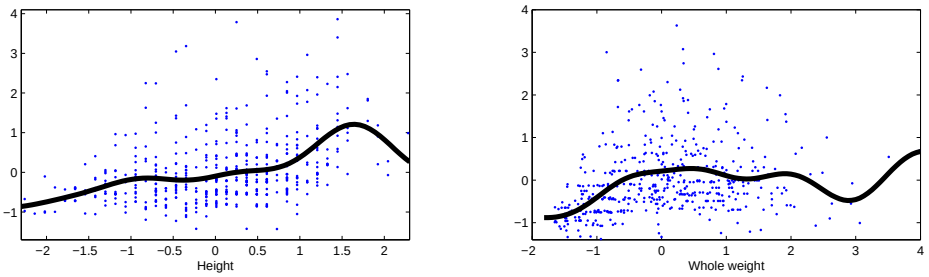


Figure 2: *Application on the abalone dataset. The inputs are divided in two groups: the first four inputs are related to the physical appearance, while the last four characterize the weight in different forms. The technique of SARA was applied to estimate a componentwise kernel model in function of exact two variables each taken from the respective group. The method resulted in the shown additive model using only the height and the whole weight of the individual abalone.*

the results, it can be seen that the method performs much better if the specific sparseness pattern is imposed, while more traditional methods as ridge regression and LASSO are not able to recover the structure well as the prior knowledge cannot be imposed straightforwardly. Note that an iterative scheme ("use a least squares estimator on all allowed subsets") would require the solution of $2^{(}N/2)$ sets of linear equations.

4.2 Exlorative Data Analysis using Componentwise Kernel Models

Secondly, we illustrate the use of the approach on the Abalone dataset (UCI) using an additive model. Assume the end user specifies the following sparseness pattern: only two variables may be used where the first may be chosen amongst $X_1, X_2 X_3, X_4$ (lengths), and the second amongst $X_5, X_6 X_7, X_8$ (weights). The proposed primal-dual kernel machine elaborated in the previous section is used. A subset of 500 datapoints is used for training to make the algorithm practical. The other datapoints are used for tuning the regularization constant γ and the bandwidth of the RBF kernel. The performance is slightly worse compared with a standard componentwise kernel machine [13, 14] or a standard SVM, but the structural constraint is not satisfied for the latter two.

5 Conclusions

This paper discussed a method resulting in estimates which have a prespecified sparseness pattern. This meachanism enabled the user to impose a certain user-defined structure ("selection amongst alternatives") on the final estimate as illustrated on a number of experiments. The method is contrasted to the approach of LASSO and is extended towards a primal-dual kernel machine based on the measure of maximal variation. An important result was that the estimate corresponds with a univariate least squares estimator if only one set of alternatives were considered.

Acknowledgments. This research work was carried out at the ESAT laboratory of the KULeuven. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006), Program Sustainable Development PODO-II (CP/40); Di-

rect contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

[1] A. Antoniadis and J. Fan. Regularized wavelet approximations (with discussion). *Journal of the American Statistical Association*, 96:939–967, 2001.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[4] D. Donoho and I. Johnstone. Ideal spatial adaption by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

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

[6] J. H. Friedmann and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[7] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[8] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, Heidelberg, 2001.

[10] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:698–714, 1992.

[11] O.L. Mangasarian, J.W. Shavlik, and E.W. Wild. Knowledge-based kernel approximation. *Journal of Machine Learning Research*, 5:1127–1141, September 2004.

[12] J. Neter, W. Wasserman, and M.H. Kutner. *Applied Linear Models, Regression, Analysis of Variance and Experimental Designs*. Irwin, third edition, 1974.

[13] K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Componentwise least squares support vector machines. Chapter in *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer, pages 77–98, 2004.

[14] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing*, 64:137–159, 2005.

[15] J.A. Rice. *Mathematical statistics and data analysis*. Duxbury Press, Pacific Grove, California, 1988.

[16] R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[17] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Methods*, 3:1439–1461, March 2003.