A Convex Approach to Learning the Ridge based on CV

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[1] advances results in model selection by relaxing the task of optimally tuning the regularization parameter in a number of algorithms with respect to the classical cross-validation performance criterion as a convex optimization problem. The proposed strategy differs from the scope of e.g. generalized cross-validation (GCV) as it concerns the efficient optimization, not the individual evaluation of the model selection criterion.

1 Introduction

The importance of setting the ridge parameter is emphasized for decades, for a full introduction into the topic we refer to [11]. Here we confine ourself to a summary of some key citations: the ridge plays a crucial role in Tikhonov regularization [18], ridge regression [6], smoothing splines [19], regularization networks [1], SVMs [2] and LS-SVMs [17] amongst others. Different criteria were proposed to measure the appropriateness of a ridge parameter for given data, including cross-validation (CV) [16], Moody's C_p [9] and MDL [14]. A whole track of research is involved with finding good approximations to those criteria, see e.g. generalized CV (GCV) [5] or the span estimate [3], while interest is arising in closed form descriptions of the solutionpath [4] and homotopy methods, see e.g. [10] and references.

This paper reports on recent advances combining both problems in a joint formalism. Specifically, the authors proposed earlier in various publications [13, 12, 11] to formulate the model selection problem as a constraint optimization problem, and eventually relax it into a convex problem which can be solved properly using standard tools. As a consequence, we are able to recover the optimal model with respect to a classical training criterion, and simultaneously the corresponding optimal ridge with respect to a model selection criterion as CV. This paper proposes a tighter relaxation, and gives sufficient results to allow for proper analysis of the learning task. We give an application to the simple task of setting the ridge in linear ridge regression, and report on some numerical results.

The motivations for studying this problem are various. We allow ourself some optimism in order to provoke some vivid discussions on the topic.

- (Practice) Users of machine learning tools are in general not interested in being concerned with tuning the algorithm towards the application at hand. We consider this result as an important step to fully automated algorithms.
- (Convexity) It is often much easier to study worst case behavior for convex sets, instead of trying to characterize all possible local minima. This will enable a proper interpretation of the learning task of learning the ridge.
- (Complexity): Many complexity measures do not increase by considering instead of a set of solutions its convex hull, e.g. in the case of Rademacher complexity [15] and others [8].
- (Extensions): Learning the ridge serves as a bootstrap for automatizing and analyzing more complex model and structure selection problems (e.g. backward selection), which often suffer from local minima and the lack of a formal framework.
- (Approach to the global optimum): The algorithmic approach of first finding the solution of the convex relaxation and then projecting this one one the original (non-convex) solutionpath is proposed as a more efficient alternative for general purpose global optimization routines.

[1]This result emerged from discussions with various people in the field, we like to acknowledge specifically M. Pontil, U. von Luxburg and O. Chapelle.

This paper is organized as follows. In the first section, the general problem is stated and the relaxation is introduced. In the second section, we apply those ideas of learning the ridge in ridge regression with respect to a CV criterion and section 4 reports the results of a numerical case study.

2 Ridge Solution Set

The problem and corresponding convex approach is firstly stated in an abstract way.

Definition 1 (Ridge Solution Set) *Let $v \in \mathbb{R}^N$ and $\Omega = \Omega^T \in \mathbb{R}^{N \times N}$ be a positive semi-definite symmetric matrix. Tikhonov regularization schemes of linear operators typically boil down to the solution $\hat{u} \in \mathbb{R}^N$ of the following set of linear equations for a fixed $0 < \gamma < +\infty$:*

$$(\Omega + \gamma I_N)\, u = v. \tag{1}$$

The ridge solution set can then be defined as the set of all solutions $\hat{u}$ corresponding with a value $0 < \gamma < +\infty$, which we denote as the solution set $\mathcal{S}_\gamma$. Formally

$$\mathcal{S}(\gamma, u|\Omega, v) = \left\{ u_\gamma \in \mathbb{R}^N \;\middle|\; \exists\, 0 < \gamma < +\infty \quad \textit{s.t.} \quad (\Omega + \gamma I_N)\, u_\gamma = v \right\}. \tag{2}$$

and analogously but with minimal regularization constant γ_0, we define

$$\mathcal{S}_0(\gamma, u|\gamma_0, \Omega, v) = \left\{ u_\gamma \in \mathbb{R}^N \;\middle|\; \exists\, \gamma_0 < \gamma < +\infty \quad \textit{s.t.} \quad (\Omega + \gamma I_N)\, u_\gamma = v \right\}. \tag{3}$$

Let $U\Sigma U^T = \Omega$ denote the SVD of the matrix Ω with $UU^T = U^T U = I_N$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$ containing all ordered positive eigenvalues such that $\sigma_1 \geq \cdots \geq \sigma_N$.

Proposition 1 (Smoothness of the Ridge Solution Set) *The solutionset $\mathcal{S}(\gamma, u|\Omega, v)$ (when $\sigma_N > 0$) or $\mathcal{S}_0(\gamma, u|\gamma_0 > 0, \Omega, v)$ is Lipschitz smooth.*

Proof: Let γ_0 denote the minimal allowed regularization parameter (if it exists) or zero otherwise in the case $\sigma_N > 0$. The following inequality holds

$$\begin{aligned}
\left\| (\Omega + \gamma_1 I_N)^{-1} v - (\Omega + \gamma_2 I_N)^{-1} v \right\|_2 &= \left\| U \left((\Sigma + \gamma_1 I_N)^{-1} - (\Sigma + \gamma_2 I_N)^{-1} \right) U^T v \right\|_2 \\
&\leq \left\| (\Sigma + \gamma_1 I_N)^{-1} - (\Sigma + \gamma_2 I_N)^{-1} \right\|_2 \|v\|_2 \\
&\leq \max_i \left| \frac{1}{\sigma_i + \gamma_1} - \frac{1}{\sigma_i + \gamma_2} \right| \|v\|_2 \\
&\leq \frac{\|v\|_2}{(\sigma_N + \gamma_0)^2} \;\; |\gamma_1 - \gamma_2|,
\end{aligned} \tag{4}$$

by application of the Cauchy-Schwartz inequality, the definition of the 2-norm of a matrix and application of the Lipshitz condition for the function $g(x) = 1/(\sigma_N + \gamma_0 + x)$.

□

Proposition 2 (Convex relaxation to $\mathcal{S}(\gamma, u|\Omega, v)$) *Let σ_i' be equal to $\sigma_i + \gamma_0$ if the minimum value of γ is bounded by γ_0, or $\sigma_i' = \sigma_i$ for all $i = 1, \ldots, N$. The proper polygon in $\mathbb{R}^N$ described as follows contains the set $\mathcal{S}(\gamma, u|\Omega, v)$*

$$\mathcal{S}'(\Lambda, u|\Omega, v) = \begin{cases} U_i^T u = \lambda_i U_i^T v & \forall i = 1, \ldots, N \\ 0 < \lambda_i < \frac{1}{\sigma_i'} & \forall i = 1, \ldots, N \\ \left(\frac{\sigma_k'}{\sigma_i'} \right) \lambda_k \leq \lambda_i < \lambda_k & \forall \sigma_i' > \sigma_k' \\ \lambda_k = \lambda_i & \forall \sigma_k' = \sigma_i' \end{cases} \tag{5}$$

such that the maximal distance from an element in $\mathcal{S}'(\Gamma, u|\Omega, v)$ to its closest counterpart of the non-convex $\mathcal{S}(\gamma, u|\Omega, v)$ can be bounded in terms of the maximum range of the inverse eigenvalue spectrum (augmented by γ_0 in the case $\sigma_N = 0$).

Proof: The necessity of the linear constraints making up $\mathcal{S}'(\Lambda, u|\Omega, v)$ can be easily verified. Let γ' be defined as $\gamma - \gamma_0 > 0$ The necessity of the inequality $\left(\frac{\sigma'_k}{\sigma'_i}\right)\lambda_k < \lambda_i$ if $\lambda_i^* = \frac{1}{\sigma'_i+\gamma'}$ and $\lambda_k^* = \frac{1}{\sigma'_k+\gamma'}$ for all $\sigma'_i > \sigma'_k$ is proved as follows

$$\sigma'_i = \sigma'_k + |\sigma'_i - \sigma'_k| \Leftrightarrow \lambda_i^* = \frac{\lambda_k}{1+\lambda_k^*|\sigma'_i - \sigma'_k|} \geq \frac{\lambda_k^*}{1+\frac{1}{\sigma'_k}|\sigma'_i - \sigma'_k|} = \lambda_k^*\left(\frac{\sigma'_k}{\sigma'_i}\right). \tag{6}$$

The maximal difference between a solution u_Γ for given Γ, and its corresponding closest $u_{\hat{\gamma}}$ can be written as

$$\min_\gamma \left\|U\Lambda U^T v - (\Omega + \hat{\gamma} I_N)^{-1} v\right\|_2 \quad \leq \quad \min_\gamma \|v\|_2 \max_{i=1}^N \left(\left|\lambda_i - \frac{1}{\sigma'_i + \gamma}\right|\right), \tag{7}$$

which follows along the same lines as in (4). Then using the property that for any two values of λ_i and λ_k, the minimum $\min_\gamma \max(|\lambda_i - 1/(\sigma'_i - \gamma')|, |\lambda_k - 1/(\sigma'_k - \gamma')|)$ is bounded by the worst case that the solution γ passes through λ_i or through λ_k, the following inequality is obtained:

$$\begin{aligned}
\min_\gamma \max_i \left(\left|\lambda_i - \frac{1}{(\sigma'_i + \gamma)}\right|\right) \quad &\leq \quad \max_{i\neq k} \left(\left|\lambda_i - \frac{1}{(\sigma'_i + \gamma_k)}\right|\right) \\
&< \quad \max_{i\neq k} \left(\lambda_i \left|1 - \left(\frac{\sigma'_i}{\sigma'_k}\right)\right|\right) \\
&< \quad \max_{i>k} \left(\left|\frac{1}{\sigma'_i} - \frac{1}{\sigma'_k}\right|\right) \triangleq \kappa'_n,
\end{aligned} \tag{8}$$

with γ_k such that $\lambda_k = \frac{1}{\sigma'_k+\gamma_k}$, or $\gamma_k = \frac{1}{\lambda_k} - \sigma'_k$ which ought to be greater than zero by construction. Combining equations (7) and (8), one obtains the inequality

$$\forall \Gamma \in \mathbb{R}^N \;\; \exists \hat{\gamma} \;\; \text{s.t.} \;\; \|u_\Gamma - u_{\hat{\gamma}}\|_2 \leq \|v\|_2 \kappa'_n. \tag{9}$$

□

This results provide sufficient tools to conduct a thorough analysis of the relaxation, and its behavior when N grows, which will be the topic of a forthcoming journal paper.

3 Tuning the Trade-off in Ridge Regression

3.1 Learning the Ridge using a Validation Criterion

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^D \times \mathbb{R}$ be a dataset. The ridge regression estimator $f(x) = w^T x$ with $w \in \mathbb{R}^D$ minimizes the following regularized loss function

$$\hat{w} = \arg\min_w \mathcal{J}_\gamma(w) = \sum_{i=1}^n \ell\left(w^T x_i - y_i\right) + \frac{\gamma}{2} w^T w \tag{10}$$

consisting of a fitting term with loss $\ell : \mathbb{R} \to \mathbb{R}^+$ and a term penalizing the complexity measured by the 2-norm of w.

Proposition 3 (Normal equations for Ridge Regression) *In the case $\ell(\cdot) = (\cdot)^2$, necessary and sufficient conditions for w to be the unique global minimizer of (10) are given as the linear system*

$$\textit{KKT}(w|\gamma, \mathcal{D}) : \left(X^T X + \gamma I_D\right)\, w = X^T Y, \tag{11}$$

where $X \in \mathbb{R}^{n\times D}$ and $Y \in \mathbb{R}^n$ are vectors containing the data and I_D is the identity matrix of size $D \times D$.

Note that we use the notation of KKT in order to hint to the extension using other learning machines (as SVMs) which boil down to solving a convex optimization problem including inequalities. Let $\mathcal{D}^v = \{(x_j^v, y_j^v)\}_{j=1}^{n_v} \subset \mathbb{R}^D \times \mathbb{R}$ be a validation dataset. The optimization problem of finding the optimal regularization parameter with respect to a validation performance criterion can then be written as follows

$$(\hat{w}, \hat{\gamma}) = \arg\min_{w,\gamma>0} \sum_{j=1}^{n_v} \ell\left(w^T x_j^v - y_j^v\right) \quad \text{s.t.} \quad \text{KKT}(w|\gamma, \mathcal{D}) = \mathcal{S}(\gamma, w|\mathcal{D}). \tag{12}$$

and by replacing $\mathrm{KKT}(w;\gamma,\mathcal{D})$ by the convex hull defined in Proposition 2, one obtains the following convex optimization problem

$$\left(\hat{w},\hat{\Lambda}\right)=\arg\min_{w,\Lambda}\sum_{j=1}^{n_v}\ell\left(w^Tx_j^v-y_j^v\right)\quad\text{s.t.}\quad\mathcal{S}'(\Lambda,w|\mathcal{D}). \tag{13}$$

which can be solved as a quadratical programming problem when we use $\ell(z)=z^2$ as classical, or an linear programming problem when using $\ell(z)=|z|$ which may be preferred from a robustness or computational point of view.

Corollary 1 (Modified Ridge Regression yielding a Convex Solution Path) *The convex relaxation constitutes the solution path for the modified ridge regression problem*

$$\hat{w}=\arg\min_{w}\mathcal{J}_\Gamma(w)=\sum_{i=1}^{n}\ell\left(w^Tx_i-y_i\right)+\frac{1}{2}w^T\left(U\Gamma U^T\right)w \tag{14}$$

where $\Gamma=\mathrm{diag}(\gamma_1,\ldots,\gamma_D)$ *and* γ_d *satisfies the constraint* $\gamma_d=\frac{1}{\lambda_d}-\sigma_d$ *for all* $d=1,\ldots,D$*, and the following inequalities hold by translating (5):*

$$\begin{cases}\gamma_d>0 & \forall d=1,\ldots,D\\ \left(\frac{\sigma_g}{\sigma_d}\right)(\sigma_d+\gamma_d)\geq(\sigma_g+\gamma_g)>(\sigma_d+\gamma_d) & \forall\sigma_g>\sigma_d\\ \gamma_d=\gamma_g & \forall\sigma_d=\sigma_g\end{cases} \tag{15}$$

This formulation hints at the formulation of Principal Component Regression [7] and gives an automatic procedure to determine the rank in this setting when relaxing $\gamma_d>0$ to $\gamma_d\geq 0$ for all $d=1,\ldots,D$.

3.2 Extension to a Cross-Validation Setting

The extension of the validation measure to the more popular L-fold cross-validation is now studied. Let $\mathcal{D}_{(l)}$ and $\mathcal{D}_{(l)}^v$ denote the set of training and validation data respectively corresponding with the lth fold for $l=1,\ldots,L$, and such that $\bigcup_l\mathcal{D}_{(l)}^v=\mathcal{D}$, $\mathcal{D}_{(l)}^v\bigcap\mathcal{D}_{(l)}=\bigcap_l\mathcal{D}_{(l)}^v=\phi$ and $n_{(l)}=|\mathcal{D}_{(l)}|$. Let $n^L=\sum_l n_{(l)}$ define the total number of constraints. The problem of tuning γ with respect to an L-fold CV criterion can then be formalized as follows:

$$\left(\hat{w}_{(l)},\hat{\gamma}\right)=\arg\min_{w_{(l)},\gamma>0}\sum_{l=1}^{L}\frac{1}{(n-n_{(l)})}\sum_{(x_i,y_i)\in\mathcal{D}_{(l)}^v}\ell\left(w_{(l)}^Tx_i-y_i\right)\quad\text{s.t.}\quad\mathrm{KKT}\left(w_{(l)}|\gamma,\mathcal{D}_{(l)}\right),\forall l=1,\ldots,L. \tag{16}$$

Note that the unknown γ is coupled over the folds. Instead of relaxing the KKT constraints independently, we propose to couple the linear necessity constraints similarly.

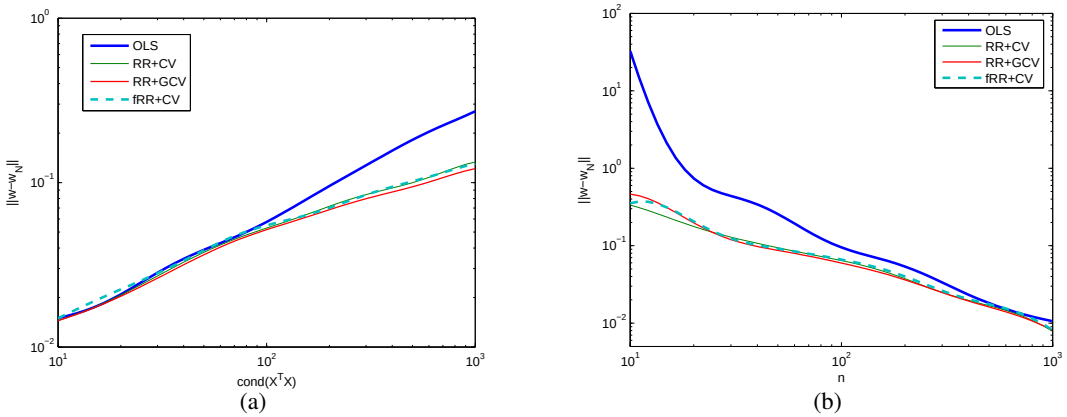
Proposition 4 (Coupling over different folds) *Let* $k=1,\ldots,n^L$ *b e a unique enumeration of the different elements of the set* $\{(l,i)\}_{l=1,i=1}^{L;n_{(l)}}$*. Let* $\Sigma^L=\{\sigma_k\}_{k=1}^{n^L}$ *equal to* $\left\{\{\sigma_{(l)i}\}_{i=1}^{n_{(l)}}\right\}_{l=1}^{L}$ *contain the pooled set of eigenvalues of the matrices* $\left\{\Omega_l=U^l\Sigma_lU^{lT}\right\}_{l=1}^{L}$ *such that* $\sigma_1\leq\cdots\leq\sigma_{n^L}$*. Then the following coupled relaxation to the set of constraints* $\left\{KKT(w_{(l)}|\gamma,\mathcal{D}_{(l)})=\mathcal{S}(\gamma,w_{(l)}|\mathcal{D})\right\}_{l=1}^{L}$ *is proposed:*

$$\mathcal{S}'\left(\Lambda^L,w_{(l)}\middle|\,\mathcal{D}_{(1)},\ldots,\mathcal{D}_{(L)}\right)=\begin{cases}U_i^{(l)T}w_{(l)}=\lambda_kU_i^{(l)T}X^{(l)T}Y^{(l)} & \forall k\leftrightarrow(l),i\\ 0<\lambda_k<\frac{1}{\sigma_k'} & \forall k=1,\ldots,n^L\\ \left(\frac{\sigma_k'}{\sigma_l'}\right)\lambda_k<\lambda_l<\lambda_k & \forall\sigma_l'>\sigma_k'\\ \lambda_k=\lambda_l & \forall\sigma_k'=\sigma_l'\end{cases} \tag{17}$$

where $Y_{(l)}\in\mathbb{R}^{n_{(l)}}$*,* $X_{(l)}\in\mathbb{R}^{n_{(l)}\times D}$ *and* $\Omega_{(l)}=X_{(l)}^TX_{(l)}\in\mathbb{R}^{D\times D}$ *correspond with* $\mathcal{D}_{(l)}$ *for all* $d=1,\ldots,D$*.*

Thus the convex relaxation to tuning the ridge with respect to an L-fold CV criterion yields to solving

$$\min_{w_{(l)},\Lambda}\sum_{l=1}^{L}\left(\frac{1}{n-n_{(l)}}\right)\sum_{(x_i,y_i)\in\mathcal{D}_{(l)}^v}\ell\left(w_{(l)}^Tx_i^v-y_i^v\right)\quad\text{s.t.}\quad\mathcal{S}'\left(\Lambda^L,w_{(l)}\middle|\,\mathcal{D}_{(1)},\ldots,\mathcal{D}_{(L)}\right). \tag{18}$$



(a) (b)

Figure 1: *Results of a comparison between OLS and RR with $D = 10$, tuned by CV (steepest descent), by GCV (using steepest descent) and the proposed method fusing training and tuning the ridge together in one convex optimization algorithm. Panel* **(a)** *shows the evolution when ranging the condition number with $n = 50$ fixed. Panel* **(b)** *displays the evolution of the performance when the number of examples ranges and $\Gamma(X^T X) = 1e^3$ is fixed. In both cases the proposed convex relaxation is performing similar as steepest descent based counterparts, while it significantly outperforms OLS in the case of low n or a high enough condition number $\Gamma(X^T X)$.*

which can either be solved by a QP or an LP depending on the choice of $\ell : \mathbb{R} \to \mathbb{R}^+$. As we do not have the optimal γ to our disposal explicitly, we suggest to use the mean regressor of the different folds $\overline{w} \frac{1}{L} \sum_{l=1}^{L} w_{(l)}$ as the final model [12].

4 Experiments

We conducted a Monte Carlo study assessing the performance of the convex relaxation of the CV model selection problem with respect to other methods as gradient descent and GCV. Every iteration constructs a "true linear regressor" for a given value of $(n, D = 10)$ defined as $f_m(x) = w_1 x^1 + \cdots + w_D x^D$ for random values of $w = (w_1, \ldots, w_D)^T \in \mathbb{R}^D$ with $w \sim \mathcal{N}(0, C)^D$ and $C \in \mathbb{R}^{m \times m}$ a rank $\Gamma(C)$ covariance matrix with $\|C\|_2 = 1$. A dataset of size n such that $\mathcal{D}_m = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^D \times \mathbb{R}$ is constructed such that $y_i = f_m(x_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$ is sampled IID for all $i = 1, \ldots, n$. We compare three methods for tuning the ridge in linear regularized estimate (10) with respect to the baseline OLS method:

(CV+sd) 10-fold CV criterion, combined with a steepest descend tuning algorithm;

(GCV+sd) Generalized CV, combined with steepest descend tuning algorithm;

(CV+f) 10-fold CV criterion, fused together with training as in (18).

Figure 1 displays average performance for a Monte Carlo sample of 20000 iterations, indicating that it performs comparable to the classical CV method using steepest descent, and evidencing that performance with respect to the baseline method (OLS) is significant when the condition number of the covariance number grows.

5 Conclusions

This paper concerns an automatic (convex) approach towards tuning the ridge with respect to a CV criterion. In addition to theoretical results, we report on the application towards linear ridge regression. Extensions towards splines and other nonlinear methods, to kernel parameter tuning and backward selection are explained in a forthcoming journal paper.

Acknowledgements Research supported by BOF PDM/05/161 (Postdoc mandaat), FWO grant V 4.090.05N (reiskrediet), IPSI Fraunhofer FgS, Darmstadt, Germany.

(Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, research communities

(ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

[1] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, Aug. 1988.

[2] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optim margin classifier. In *In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–52. ACM, 1992.

[3] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

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

[5] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[6] A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–82, 1970.

[7] I.T. Jollife. *Principal Component Analysis*. Springer-Verlag, 1986.

[8] S. Mendelson. A few notes on statistical learning theory. *in* Advanced Lectures on Machine Learning,, 2 600:1–40, 2003. Springer LNCS.

[9] J.E. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Neural Information Processing Systems*, volume 4, pages 847–854, San Mateo CA, 1992. Morgan Kaufmann.

[10] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *Journal of Computational & Graphical Statistics*, june 01 2000.

[11] K. Pelckmans. *Primal-dual Kernel Machines*. PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, May 2005. 280 p., 05-95.

[12] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization trade-off: Fusion of training and validation levels in kernel methods. *Accepted for publication in Machine Learning*, 2005.

[13] 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.

[14] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[15] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[16] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistics Society Series*, B(36):111–147, 1974.

[17] J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[18] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington DC, 1977.

[19] G. Wahba. *Spline models for observational data*. SIAM, 1990.