Primal-Dual Monotone Kernel Regression

K. Pelckmans, M. Espinoza, J. De Brabanter†, J.A.K. Suykens, B. De Moor
K.U. Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Tel: 32/16/32 86 58 - Fax: 32/16/32 19 70

June, 2004

Abstract. This paper considers the estimation of monotone nonlinear regression based on Support Vector Machines (SVMs), Least Squares SVMs (LS-SVMs) and kernel machines. It illustrates how to employ the primal-dual optimization framework characterizing (LS-)SVMs in order to derive a globally optimal one-stage algorithm. As an practical application, this letter considers the smooth estimation of the cumulative distribution functions (cdf) which leads to a kernel regressor that incorporates a Kolmogorov-Smirnoff discrepancy measure, a Tikhonov based regularization scheme and a monotonicity constraint.

Keywords: Monotone modeling, primal-dual kernel regression, convex optimization, constraints, Support Vector Machines

1. Introduction

The use of nonparameteric nonlinear function estimation and kernel methods were largely stimulated by recent advances in Support Vector Machines and related methods [24, 18, 6, 11]. The theory of statistical learning has been a key issue for these methods as it provides bounds on the generalization performance which are based on hypothesis space complexity measures and empirical risk minimization. In this sense, it is plausible to make all assumptions of the modeling task at hand as explicit as possible during the estimation stage: by restricting the hypothesis space as much as possible, the generalization performance is likely to improve (see e.g. [24] for the case of additive models, [12] and references therein for convergence results in the case of constrained splines).

This paper further explores this line of thoughts by considering a dedicated task where prior knowledge can be imposed. An estimation of the cumulative distribution functions (cdf) is commonly made using empirical cumulative distribution functions (ecdf) see e.g. [17, 9, 19]. While many nice properties are associated to this classical estimator [2], one is often interested in the best smooth estimate as motivated by the

† Corresponding author: kristiaan.pelckmans@esat.kuleuven.ac.be



© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

assumption that the underlying distribution has a smooth form. Applications can be found in the inversion method for generating nonuniform random variates which is based on the inverse of the cdf transforming a set of uniform generated random numbers [8], and density estimation by deriving the smoothed ecdf.

This paper illustrates how one can exploit the other pillars of SVMs for this dedicated approximation task: a primal-dual optimization approach and use of a feature space mapping. This is also the case explicitly in the formulation of Least Squares Support Vector Machines [21] in which case the optimal estimate is described using simpler formulations as a consequence of taking the additional assumption of a Gaussian i.i.d. (noise) model. Apart from this advantage, the latter is also more flexible for absorbing additional (linear) equality and inequality constraints as was shown in [10, 14, 13].

For certain types of approximation problems the Chebychev measure [4] is more appropriate (as originally stated in the theory of function approximation using polynomial expansions and later on known as the L_∞ or maximum norm). In the setting of estimating the cdf it is a natural choice for a loss function [17] as it is directly related to the Kolmogorov-Smirnoff discrepancy measure between cdf's [5]. Most nonparameteric approaches (see e.g. [11, 9]) are based on two-stage procedures (*"smooth then monotone"*) or monotone (and in general constrained) least squares semi-parametric estimators (see e.g. [7, 16, 12] and references therein).

This paper is organized as follows: Section 2 derives the optimal solution of a monotone function based on a least squares and Chebychev norm and a Tikhonov [23] regularization scheme. Section 3 tunes the estimator further towards the smoother of the ecdf, while Section 4 gives some experimental results.

2. Primal-Dual Derivations

Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the training data with inputs which can be ordered $x_i \leq x_j$ if $i < j$ and outputs y_i. Consider the regression model

$$y_i = f(x_i) + e_i, \tag{1}$$

where $x_1, \ldots, x_N$ are deterministic points, $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function and $e_1, \ldots, e_N$ are uncorrelated random errors with $E[e_i] = 0$, $E[e_i^2] = \sigma_e^2 < \infty$. Let $Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$.

This section considers the constrained estimation problem of monotone kernel regression based on convex optimization techniques. First, the extension of the LS-SVM regressor towards monotone estimation

using primal-dual convex optimization techniques is discussed. The second part considers an L_∞ norm as it is an appropriate measure for the application at hand. Extensions to other convex loss functions [24, 18, 6] may follow along the lines. Furthermore and without loss of generality, the derivations are restricted to monotonically increasing functions, the case of monotonically decreasing functions can be done in a similar way.

2.1. Monotone LS-SVM regression

The primal LS-SVM regression model is given as

$$f(x) = w^T \varphi(x), \tag{2}$$

in the primal space where $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$) dimensional feature map. Also a bias can be considered [6, 18, 21]

Monotonicity constraints can be expressed via the following inequality constraints:

$$w^T \varphi(x'_i) \leq w^T \varphi(x'_{i+1}), \quad \forall i = 1, ..., N' - 1, \tag{3}$$

for a set $\mathcal{X}' = \{x'_i\}_{i=1}^{N'}$. One can impose the inequality constraints on training points ($\mathcal{X}'$ equal to $\{x_i\}_{i=1}^{N}$), on an (equidistant) grid of points, or on the points where one wants to evaluate the estimate. Sufficient conditions to have globally monotone estimates can be constructed based on the derivatives of the estimated function [7]. However, as this will depend in our setting on the chosen kernel, this path is not further pursued here. The derivation here proceeds with the first choice (monotone estimate on the training data) Therefor, the extrapolation of the estimate to out-of-sample data-points should be treated carefully.

Consider the regularized least squares cost function [21] constrained by the inequalities (3):

$$\begin{aligned} \min_{w,e_i} \mathcal{J}(w,e) &= \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \\ \text{s.t.} \quad & \begin{cases} w^T \varphi(x_i) + e_i = y_i, & \forall i = 1, ..., N \\ w^T \varphi(x_{i+1}) \geq w^T \varphi(x_i) & \forall i = 1, ..., N-1. \end{cases} \end{aligned} \tag{4}$$

Construct the Lagrangian

$$\mathcal{L}(w, e_i; \alpha_i, \beta_i) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i (w^T \varphi(x_i) + e_i - y_i) - \sum_{i=1}^{N-1} \beta_i (w^T \varphi(x_{i+1}) - w^T \varphi(x_i)), \quad (5)$$

with $\alpha \in \mathbb{R}^N$ and $0 \leq \beta \in \mathbb{R}^{N-1}$. The Lagrange dual [3] becomes

$$g(\alpha, \beta) \; = \; \inf_{w,e} \mathcal{L}(w, e_i; \alpha_i, \beta_i), \quad (6)$$

with $\beta_i \geq 0$ for all $i = 1, \ldots, N$. Taking the conditions for optimality w.r.t. w and e results in

$$\begin{cases} \partial \mathcal{L}/\partial e_i = 0 \; \rightarrow \gamma e_i = \alpha_i \\ \partial \mathcal{L}/\partial w = 0 \; \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) + \sum_{i=1}^{N-1} \beta_i \left(\varphi(x_{i+1}) - \varphi(x_i)\right). \end{cases} \quad (7)$$

When (7) holds, one can eliminate w and e by substitution of (7) in (5) and (6):

$$\begin{aligned} g(\alpha, \beta) \; = \; & \frac{1}{2} \left(\sum_{i,j=1}^{N} \alpha_i \alpha_j \varphi(x_i)^T \varphi(x_j) + 2 \sum_{i=1}^{N} \sum_{l=1}^{N-1} \alpha_i \beta_l \varphi(x_i)^T \varphi(x_{l+1}) \right. \\ & \left. + \sum_{k,l=1}^{N-1} \beta_k \beta_l (\varphi(x_k)^T \varphi(x_l) - 2\varphi(x_{k+1})^T \varphi(x_l) + \varphi(x_{k+1})^T \varphi(x_{l+1})) \right) + \frac{1}{\gamma} \sum_{i=1}^{N} \alpha_i^2 \\ & - \sum_{i=1}^{N} \alpha_i \left(\sum_{k=1}^{N} \varphi(x_i)^T \varphi(x_k) \alpha_k + \sum_{l=1}^{N-1} \beta_l (\varphi(x_i)^T \varphi(x_{l+1}) - \varphi(x_i)^T \varphi(x_l)) - \frac{1}{\gamma} \alpha_i - y_i \right) \\ & - \sum_{j=1}^{N-1} \beta_j \left(\sum_{k=1}^{N} \varphi(x_{j+1})^T \varphi(x_k) \alpha_k + \sum_{l=1}^{N-1} \beta_l (\varphi(x_{j+1})^T \varphi(x_{l+1}) - \varphi(x_{j+1})^T \varphi(x_l)) \right) \\ & - \sum_{j=1}^{N-1} \beta_j \left(\sum_{k=1}^{N} \varphi(x_j)^T \varphi(x_k) \alpha_k + \sum_{l=1}^{N-1} \beta_l (\varphi(x_j)^T \varphi(x_{l+1}) - \varphi(x_j)^T \varphi(x_l)) \right) \\ = \; & \frac{1}{2} \left(\alpha^T \Omega \alpha + 2\alpha^T (\Omega^+ - \Omega^0) \beta + \beta^T (\Omega_+^+ - \Omega_+^0 - \Omega_+^{0\,T} + \Omega_0^0) \beta \right) + \frac{1}{2\gamma} \alpha^T \alpha \\ & - \alpha^T (\Omega + \frac{1}{\gamma} I_N) \alpha - \alpha^T (\Omega^+ - \Omega^0) \beta - \beta^T (\Omega^{+T} - \Omega^{0T}) \alpha \\ & - \beta^T (\Omega_+^+ - \Omega_+^0 - \Omega_+^{0\,T} + \Omega_0^0) \beta - \alpha^T Y \\ = \; & -\frac{1}{2} \alpha^T (\Omega + \frac{1}{\gamma} I_N) \alpha - \frac{1}{2} \alpha^T (\Omega^+ - \Omega^0) \beta - \frac{1}{2} \beta^T (\Omega^{+T} - \Omega^{0T}) \alpha \\ & - \frac{1}{2} \beta^T (\Omega_+^+ - \Omega_+^0 - \Omega_+^{0\,T} + \Omega_0^0) \beta + Y^T \alpha, \end{aligned} \quad (8)$$

where $\Omega \in \mathbb{R}^{N\times N}, \Omega^+, \Omega^0 \in \mathbb{R}^{N\times N-1}$ and $\Omega_+^+,\ \Omega_0^0,\ \Omega_0^+ \in \mathbb{R}^{(N-1)\times(N-1)}$ is defined as follows:

$$\begin{cases} \Omega_{ij} &= K(x_i, x_j), & \forall i,j = 1,\ldots,N \\ \Omega_{il}^+ &= K(x_i, x_{l+1}), & \forall i = 1,\ldots,N, \forall l = 1,\ldots,N-1 \\ \Omega_{il}^0 &= K(x_i, x_l), & \forall i,j = 1,\ldots,N, \forall l = 1,\ldots,N-1 \\ \Omega_{0,kl}^0 &= K(x_k, x_l), & \forall k,l = 1,\ldots,N-1 \\ \Omega_{+,kl}^+ &= K(x_{k+1}, x_{l+1}), & \forall k,l = 1,\ldots,N-1 \\ \Omega_{0,kl}^+ &= K(x_k, x_{l+1}), & \forall k,l = 1,\ldots,N-1, \end{cases}$$

where the kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as the inner product $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ for all $i,j = 1,\ldots,N$. For the choice of an appropriate kernel K see e.g. [18, 21]. Typical examples are the use of a polynomial kernel $K(x_i, x_j) = (\tau + x_i^T x_j)^d$ of degree d with hyperparameter $\tau > 0$ or the Radial Basis Function (RBF) kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$ where σ denotes the bandwidth of the kernel. The dual function can be summarized in matrix notation as follows: $g(\alpha, \beta)$

$$-\frac{1}{2}\begin{bmatrix}\alpha \\ \beta\end{bmatrix}^T \left[\begin{array}{c|c} \Omega + 1/\gamma I_N & (\Omega^+ - \Omega^0) \\ \hline (\Omega^+ - \Omega^0)^T & (\Omega_+^+ - \Omega_+^0 - \Omega_+^{0\,T} + \Omega_0^0) \end{array}\right] \begin{bmatrix}\alpha \\ \beta\end{bmatrix} + Y^T\alpha. \quad (9)$$

The global maximum of the dual function g w.r.t. the Lagrange multipliers α and β incorporating the inequalities $\beta \geq 0$ can be found by solving a Quadratic Programming problem (QP) [3].

The final model $\hat{f}(x) = \hat{w}^T\varphi(x)$ can be evaluated in a new datapoint x^* as follows:

$$\begin{aligned} \hat{f}(x^*) &= \sum_{i=1}^{N} \alpha_i K(x_i, x^*) + \sum_{l=1}^{N-1} \beta_l \left(K(x_{l+1}, x^*) - K(x_l, x^*)\right) \\ &= \sum_{i=1}^{N} (\alpha_i + \beta_{i-1} - \beta_i) K(x_i, x^*), \end{aligned} \quad (10)$$

where $\beta_0 = \beta_N = 0$ by definition.

2.2. Monotone Chebychev kernel regression

One starts with the same primal model as (2). Consider the Chebychev measure (see [4] and citing papers) for function approximation defined as

$$\|e\|_\infty = \max_i |f(x_i) - y_i|, \quad (11)$$

over the given data-samples. The following constrained optimization problem can be formulated

$$\min_{w,e} \mathcal{J}(e,w) = \frac{1}{2} w^T w + \gamma |e|_\infty$$
$$\text{s.t.} \quad \begin{cases} w^T\varphi(x_i) + e_i = y_i & \forall i = 1,...,N \\ w^T\varphi(x_{i+1}) \geq w^T\varphi(x_i) & \forall i = 1,...,N-1. \end{cases} \tag{12}$$

As usual in convex optimization [3] and L_1 or the ϵ-insensitive loss function as in SVMs [24, 18, 6], the L_∞ norm is translated by minimizing a variable upper-bound $-t < e_i \leq t$ for all $i = 1, \ldots, N$ [3]. For reasons which become apparent in the next section, a notational distinction is made between $Y^1 = (y_1^1, \ldots, y_N^1)$ and $Y^2 = (y_1^2, \ldots, y_N^2)$ which are both taken equal to Y for the moment. Constructing the Lagrangian with Lagrange multipliers $0 \leq \alpha^+, \ 0 \leq \alpha^- \in \mathbb{R}^N$ and $\beta \in \mathbb{R}^{N-1}$ gives

$$\mathcal{L}(w,t;\alpha^+,\alpha^-,\beta) = \frac{1}{2} w^T w + \gamma t - \begin{cases} \sum_{i=1}^N \alpha_i^+ \left(t - (w^T\varphi(x_i) - y_i^1) \right) \\ + \sum_{i=1}^N \alpha_i^- \left(t + (w^T\varphi(x_i) - y_i^2) \right) \\ + \sum_{i=1}^{N-1} \beta_i \left(w^T\varphi(x_{i+1}) - w^T\varphi(x_i) \right), \end{cases} \tag{13}$$

with inequality constraints $\alpha^+, \alpha^-, \beta \geq 0$. One considers the Lagrange dual of (13). Let $\varphi = (\varphi(x_1), \ldots, \varphi(x_N))^T \in \mathbb{R}^{N \times n_f}$ and the reduced versions $\varphi^+ = (\varphi(x_2), \ldots, \varphi(x_N))^T$ and $\varphi^0 = (\varphi(x_1), \ldots, \varphi(x_{N-1}))^T \in \mathbb{R}^{N-1 \times n_f}$.

$$\begin{aligned} g(w,t;\alpha^+,\alpha^-,\beta) \ = \ & \frac{1}{2} w^T w + \gamma t - \alpha^{+T} \left(t - (w^T\varphi - Y^1) \right) \\ & - \alpha^{-T} \left(t + (w^T\varphi - Y^2) \right) - \beta^T \left(w^T\varphi^+ - w^T\varphi^0 \right) \\ \text{s.t.} \quad & \begin{cases} \alpha^-, \alpha^+, \beta \geq 0 \\ \gamma = 1_N^T\alpha^+ + 1_N^T\alpha^- \\ w = \varphi^T(\alpha^- - \alpha^+) + (\varphi^+ - \varphi^0)^T \beta. \end{cases} \end{aligned} \tag{14}$$

Elimination of the high dimensional vector w and the scalar t and application of the kernel trick results in the following quadratic programming problem

$$\max_{\alpha^+,\alpha^-,\beta} g(\alpha^+,\alpha^-,\beta) = -\frac{1}{2} \left[\begin{array}{c} \alpha^+ \\ \hline \alpha^- \\ \hline \beta \end{array} \right]^T H \left[\begin{array}{c} \alpha^+ \\ \hline \alpha^- \\ \hline \beta \end{array} \right] + Y^{1T}\alpha^+ - Y^{2T}\alpha^-$$
$$\text{s.t.} \quad 1_N^T(\alpha^+ + \alpha^-) = \gamma, \quad \alpha^+, \alpha^-, \beta \geq 0, \tag{15}$$

where H is defined as

$$H = \left[\begin{array}{c|c|c} \Omega & \text{-}\Omega & (\Omega^+ - \Omega^0) \\ \hline \text{-}\Omega & \Omega & \text{-}(\Omega^+ - \Omega^0) \\ \hline (\Omega^+ - \Omega^0)^T & \text{-}(\Omega^+ - \Omega^0)^T & (\Omega^+_+ - \Omega^0_+ - \Omega^{0\,T}_+ + \Omega^0_0) \end{array}\right].$$

The final model $\hat{f}(x) = \hat{w}^T \varphi(x)$ can be evaluated on a new datapoint x^* as in (10) where $\alpha = \alpha^+ - \alpha^-$.

3. Smoothing the Empirical Cumulative Distribution Function

This section extends the monotone Chebychev kernel machine towards the task of smoothing the ecdf. For notational convenience and to keep the derivation conceptually simple, only the univariate case is considered here, although the multivariate case may follow along the same lines [17, 13]. Consider a random variable X with a smooth Cumulative Distribution Function (cdf). For a given realization of the sample $X_1, \ldots, X_N$, say $x_1, \ldots, x_N$, the ecdf is defined as [2]

$$\hat{F}(x) = \frac{1}{N} \sum_{k=1}^{N} I_{(\infty,x]}(x_k), \quad \text{for} \quad -\infty < x < \infty \tag{16}$$

where the indicator function $I_{(\infty,x]}(x)$ equals 1 if $x \in (\infty, x]$ and 0 otherwise. This estimator has the following properties: (i) it is uniquely defined; (ii) its range is $[0, 1]$; (iii) it is non-decreasing and continuous on the right; (iv) it is piecewise constant with jumps at the observed points, i.e. it enjoys all properties of its theoretical counterpart, the cdf. Furthermore, $\|F(x_i) - \hat{F}(x_i)\|_\infty \to 0$ with probability one as stated in the Glivenko-Cantelli Theorem (see e.g. [2]). In order to build a smooth estimate based on the ecdf $\hat{F}$, a function approximation task is considered with input and dependent variables $\{x_i, \hat{y}^1_i, \hat{y}^2_i\}_{i=1}^N$. Now it becomes apparent why one makes a distinction between $Y^1 = (0, y_1, \ldots, y_{N-1})^T$ and $Y^2 = (y_1, \ldots, y_{N-1}, 1)^T$, which acts as lower- and upper bounds at the observed values of the ecdf in the points $(x_1, \ldots, x_N)^T$.

One transforms the estimates by shifting Y^1 and Y^2 downwards by 0.5 (for other appropriate transformations, see e.g. [9]) as can be argued by the consideration that the average $\int F(x)dx$ of a cdf is always 0.5. As such, the bias term as in [18, 21] needs not to be considered further. A classical discrepancy measure between different cdf's is the Kolmogorov-Smirnoff goodness-of-fit hypothesis test [5] which is based on a Chebychev L_∞ norm.

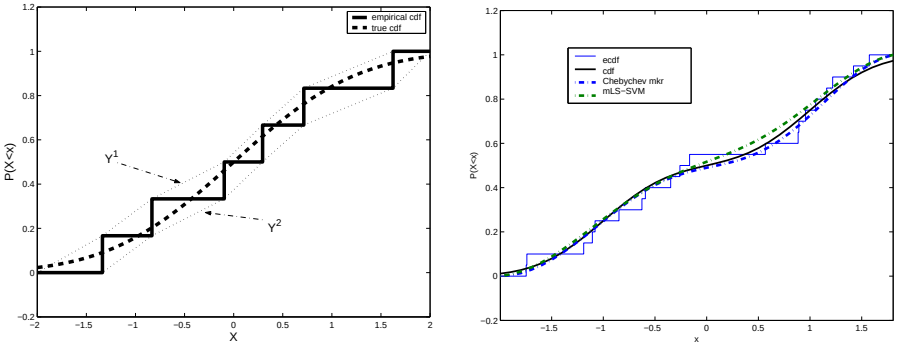

Figure 1. **(a)** *The smoothed ecdf is insensitive to the jumps (in between* Y^1 *and* Y^2*) in the ecdf and aims at minimizing the distance from this zone while remaining as smooth as possible;* **(b)** *application of the smooth estimation of the ecdf.*

Incorporating these assumptions in the estimation process leads to a modification of (15). The constraints $w^T\varphi(x_-) = -0.5$ and $w^T\varphi(x_+) = 0.5$ are added where x_- and x_+ are respectively lower- and upper bounds for the support of F. By deriving a dual expression for this constrained optimization problem the final optimization problem becomes as in (15) with $X^1 = (x_-, x_1, \ldots, x_{N-1})^T$, $X^2 = (x_1, x_2, \ldots, x_N, x_+)^T$, $Y^1 = (0, \hat{y}_1, \ldots, \hat{y}_{N-1})^T$ and $Y^2 = (\hat{y}_1, \ldots, \hat{y}_N, 1)^T$. Furthermore, to impose the equality constraints $w^T\varphi(x_-) = 0$ and $w^T\varphi(x_+) = 1$ exactly, one can easily see that the equality constraint of (15) should be adapted as $-1_N^T(\tilde{\alpha}^+ + \tilde{\alpha}^-) = \gamma$ where $\tilde{\alpha}^+$ and $\tilde{\alpha}^-$ are defined similar as α^+, α^- but do not contain the multipliers associated with x_+ and x_-.

4. Illustrative Example

To illustrate the use of the presented smoothing technique in the context of cdf estimation, the following experiments were designed. At first, consider a dataset which consists of a realization of both $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$ with a total of 30 samples. The hyper-parameters of the smoothing techniques are determined by minimizing a 10-fold cross-validation criterion. Figure 1.b shows the ecdf and the smoothed cdf. A Monte Carlo experiment was conducted relating three empirical cdf estimators to the true underlying cdf using the Kullback-Leibler distances. While the L_2 based monotone LS-SVM does not differ substantially from the classical empirical cdf, the monotone Chebychev kernel regressor increases in performance as shown in the boxplot of Figure 2.a.

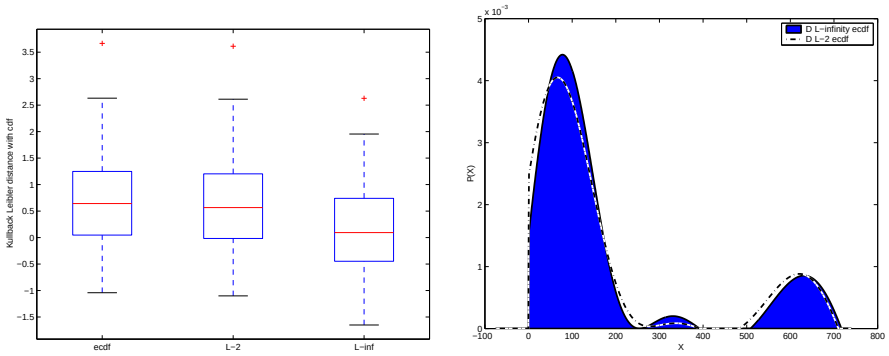

Figure 2. **(a)** *Boxplots of the results of a Monte Carlo simulation for estimating the cdf based on respectively the ecdf, the monotone LS-SVM smoother and the monotone Chebychev kernel regressor.* **(b)** *Density estimation of the suicide data using the derivative of the monotone Chebychev kernel regressor and the monotone LS-SVM technique.*

The technique based on the L_2 norm and the L_∞ norm was applied to generate a density estimate of the suicide data (see e.g. [19]) by deriving the smoothed estimate. In this case the support of the data was known to have an exact lower bound at 0 which can be nicely incorporated in this framework as shown in Figure 2.b. Note that the well-known trimodal structure of the distribution [19] becomes more explicit when using the Chebychev norm.

5. Conclusions

This paper described the derivation of monotone kernel regressors based on primal-dual optimization theory for the case of a least squares loss function (monotone LS-SVM regression) as well as an L_∞ norm (monotone Chebychev kernel regression). This is illustrated in the context of smoothly estimating the cdf.

Acknowledgments. This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0499.04 (robust

statistics), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); Direct 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. R.E. Barlow, D.J. Bartholomew, J.M. Bremner and H.D. Brunk. *Statistical inference under order restrictions* Wiley, London, 1977.
2. P. Billingsley. *Probability and Measure*. Wiley & Sons, 1986.
3. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
4. P. L. Chebyshev. "Sur les questions de minima qui se rattachent la reprsentation approximative des fonctions" *Oeuvres de P. L. Tchebychef*, 1, 273-378, Chelsea, New York, 1961 (1859).
5. W.J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1971.
6. Cristianini N., Shawe-Taylor J. *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.
7. C. De Boor and B. Schwartz. Piecewise monotone interpolation *Journal of Approximation Theory*, 21, 411-416, 1977.
8. L. Devroye. *Non-Uniform Random Variate Generation* Springer-Verlag, 1986
9. C.K. Gaylord D.E. Ramirez. "Monotone regression splines for smoothed bootstrapping", *Computational Statistics Quarterly* 6(2), 85-97, 1991.
10. I. Goethals, K. Pelckmans, J.A.K. Suykens, B. De Moor, "Identification of MIMO Hammerstein Models using Least Squares Support Vector Machines", *Internal Report 04-45, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*, 2004, Submitted.
11. T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
12. E. Mammen, J. S. Marron, B. A. Turlach, and M. P. Wand. "A General Projection Framework for Constrained Smoothing", *Statistical Science* 16(3), 232-248, 2001.
13. K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, B. De Moor, "Componentwise Least Squares Support Vector Machines", *Internal Report 04-75, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*, 2004.
14. K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization: fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SCD-SISTA, K.U.Leuven (Leuven, Belgium)*, 2003, submitted.
15. T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, volume 78, pages 1481–1497, september 1990.
16. J.O. Ramsay. Monotone Regression Splines in Action. *Statistical Science*, 3, 425-461, 1988

17. D.W. Scott. *Multivariate Density Estimation, theory, practice and visualization*. Wiley series inb probability and mathematical statistics, 1992.
18. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
19. B.W. Silverman. *Density Estimation for Statistics and Data Analysis* Monographs on Statistics and Applied Probability, 26, Chapman & Hall, 1986.
20. J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
21. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.
22. J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Eds.) *Advances in Learning Theory: Methods, Models and Applications*. NATO Science Series III: Computer & Systems Sciences, 190, IOS Press Amsterdam, 2003.
23. A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington DC, 1977.
24. V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.