Neural Processing Letters (2005) 22:171–182
DOI 10.1007/s11063-005-5264-1
© Springer 2005

Primal-Dual Monotone Kernel Regression

K. PELCKMANS[1,⋆], M. ESPINOZA[1], J. DE BRABANTER[2],
J. A. K. SUYKENS[1], and B. DE MOOR[1]
[1]*K.U. Leuven, ESAT-SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium. e-mail: kristiaan.pelckmans@esat.kuleven.ac.be*
[2]*Hogeschool KaHo Sint-Lieven (Associatie KULeuven) Departement Industrieel Ingenieur*

Abstract. This paper considers the estimation of monotone nonlinear regression functions based on Support Vector Machines (SVMs), Least Squares SVMs (LS-SVMs) and other kernel machines. It illustrates how to employ the primal-dual optimization framework characterizing LS-SVMs in order to derive a globally optimal one-stage estimator for monotone regression. As a 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.

Key words. constraints, convex optimization, monotone regression, primal-dual kernel regression, support vector machines

1. Introduction

The use of non-parametric nonlinear function estimation and kernel methods is largely stimulated by recent advances in Support Vector Machines (SVMs) and related methods [1–4]. 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. [1] for the case of additive models, [5] and references therein for convergence results in the case of constrained splines).

This letter further elaborates on this but rather takes an optimization point of view. Once an appropriate global optimality principle is formalized, it is shown how one can employ two main pillars of SVMs (a) a primal-dual optimization approach and (b) the use of a feature space mapping induced by a positive definite kernel, in order to obtain a globally optimal non-parametric representation and prediction model. Both principles act also as cornerstones to the formulation of Least Squares SVM [6, 7] (LS-SVM) and its application towards the modeling of componentwise LS-SVMs [8] and Hammerstein models (Geothals et al.

⋆Corresponding author.

(2004), submitted) for nonlinear system identification. Furthermore, advances of the primal-dual framework were exploited for the purpose of regularization parameter tuning (Pelckmans et al. (2003), submitted) This letter focuses on the design of methods for estimating smooth monotone increasing or decreasing functions in the sense of a Chebychev measure [9] (also called an L_∞ or maximum norm).

The usefulness of the approach is illustrated by studying the proposed Chebychev kernel machine for estimating a smooth and monotone increasing distribution function of a given sample. A discontinuous estimate of a cumulative distribution function (cdf) is provided by the empirical cumulative distribution function (ecdf), see e.g. [10–12]. While many nice properties are associated with this classical estimator [13], one is often interested in the best smooth estimate. Applications can be found in the inversion method for generating non-uniform random variates which is based on the inverse of the cdf transforming a set of uniform generated random numbers [14], and density estimation by taking the derivative of the smoothed ecdf. Furthermore, the L_∞ measure is a natural choice for a loss function in this application [10] as it is directly related to the Kolmogorov–Smirnoff discrepancy measure between cdf's [15]. Most non-parametric approaches (see e.g. [4,11]) are based on two-stage procedures ("*smooth then monotone*") or monotone (and in general constrained) least squares semi-parametric estimators where the specific (parametric) form of the model is exploited (see e.g. [5,16–18]). With the proposed method this is done in one stage by employing a non-parametric strategy.

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 [19] regularization scheme. Section 3 tunes the estimator further towards the application of smoothing 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 for which one assumes that it can be ordered as $x_i \leqslant x_j$ if $i < j$ for all $i, j = 1, \ldots, N$ 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 $E[e_i] = 0$, $E\left[e_i^2\right] = \sigma_e^2 < \infty$. 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 [1,2] may follow along the same lines. Furthermore, the derivations are restricted to monotonically increasing functions, while 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

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

where $\varphi: \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$) dimensional feature map. Also a bias term can be considered [2,6]. Monotonicity constraints can be expressed via the following inequality constraints:

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

for a set $\mathcal{X}' = \{x'_i\}_{i=1}^{N'}$. One can impose the inequality constraints on training datapoints (i.e. $\mathcal{X}'$ equal to $\{x_i\}_{i=1}^{N}$), on an (equidistant) grid of points or at other points where one wants to evaluate the estimate. Sufficient conditions to have globally monotone estimates can be derived based on the derivatives of the estimated function [16]. 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 from the training data). Therefore, the extrapolation of the estimate to out-of-sample data-points should be treated carefully.

Consider the following regularized least squares cost function [6] constrained by the inequalities (3):

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

Construct the Lagrangian

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

with $\alpha \in \mathbb{R}^N$ and $0 \leqslant \beta \in \mathbb{R}^{N-1}$. The optimal solution is found as the saddle-point of the Lagrangian by first minimizing over the primal variables w_i and e_i and then maximizing over the dual multipliers α_i and β_i. The Lagrange dual [20]

becomes $g(\alpha,\beta)=\min_{w,e}\mathcal{L}(w,e_i;\alpha_i,\beta_i)$ with $\beta_i\geqslant 0$ for all $i=1,\ldots,N$. Taking the conditions for optimality w.r.t. w and e results in

$$\begin{aligned}&\partial\mathcal{L}/\partial e_i=0\rightarrow\gamma e_i=\alpha_i,\\&\partial\mathcal{L}/\partial w=0\rightarrow w=\sum_{i=1}^{N}\alpha_i\varphi(x_i)+\sum_{i=1}^{N-1}\beta_i\left(\varphi(x_{i+1})-\varphi(x_i)\right).\end{aligned}\tag{6}$$

When (6) holds, one can eliminate w and e in (5):

$$\begin{aligned}g(\alpha,\beta)=&\frac{1}{2}\left(\sum_{i,j=1}^{N}\alpha_i\alpha_j\varphi(x_i)^T\varphi(x_j)+2\sum_{i=1}^{N}\sum_{l=1}^{N-1}\alpha_i\beta_l\varphi(x_i)^T\varphi(x_{l+1})\right.\\&\left.+\sum_{k,l=1}^{N-1}\beta_k\beta_l(\varphi(x_k)^T\varphi(x_l)-2\varphi(x_{k+1})^T\varphi(x_l)+\varphi(x_{k+1})^T\varphi(x_{l+1}))\right)+\frac{1}{\gamma}\sum_{i=1}^{N}\alpha_i^2\\&-\sum_{i=1}^{N}\alpha_i\left(\sum_{k=1}^{N}\varphi(x_i)^T\varphi(x_k)\alpha_k+\sum_{l=1}^{N-1}\beta_l(\varphi(x_i)^T\varphi(x_{l+1})-\varphi(x_i)^T\varphi(x_l))-\frac{1}{\gamma}\alpha_i-y_i\right)\\&-\sum_{j=1}^{N-1}\beta_j\left(\sum_{k=1}^{N}\varphi(x_{j+1})^T\varphi(x_k)\alpha_k+\sum_{l=1}^{N-1}\beta_l(\varphi(x_{j+1})^T\varphi(x_{l+1})-\varphi(x_{j+1})^T\varphi(x_l))\right)\\&-\sum_{j=1}^{N-1}\beta_j\left(\sum_{k=1}^{N}\varphi(x_j)^T\varphi(x_k)\alpha_k+\sum_{l=1}^{N-1}\beta_l(\varphi(x_j)^T\varphi(x_{l+1})-\varphi(x_j)^T\varphi(x_l))\right)\\=&\frac{1}{2}\left(\alpha^T\Omega\alpha+2\alpha^T(\Omega^+-\Omega^0)\beta+\beta^T(\Omega_+^+-\Omega_+^0-\Omega_+^{0\,T}+\Omega_0^0)\beta\right)+\frac{1}{2\gamma}\alpha^T\alpha\\&-\alpha^T(\Omega+\frac{1}{\gamma}I_N)\alpha-\alpha^T(\Omega^+-\Omega^0)\beta-\beta^T(\Omega^{+T}-\Omega^{0T})\alpha\\&-\beta^T(\Omega_+^+-\Omega_+^0-\Omega_+^{0\,T}+\Omega_0^0)\beta-\alpha^TY\\=&-\frac{1}{2}\alpha^T\left(\Omega+\frac{1}{\gamma}I_N\right)\alpha-\frac{1}{2}\alpha^T(\Omega^+-\Omega^0)\beta-\frac{1}{2}\beta^T(\Omega^{+T}-\Omega^{0T})\alpha\\&-\frac{1}{2}\beta^T(\Omega_+^+-\Omega_+^0-\Omega_+^{0\,T}+\Omega_0^0)\beta+Y^T\alpha,\end{aligned}\tag{7}$$

where $\Omega\in\mathbb{R}^{N\times N}$, $\Omega^+,\Omega^0\in\mathbb{R}^{N\times(N-1)}$ and Ω_+^+, Ω_0^0, $\Omega_0^+\in\mathbb{R}^{(N-1)\times(N-1)}$ is defined as follows:

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

and the Mercer kernel $K:\mathbb{R}^d\times\mathbb{R}^d\rightarrow\mathbb{R}$ 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. [2,6]). Typical examples are the use of a polynomial kernel $K(x_i,x_j)=(\tau+x_i^Tx_j)^d$ of degree d with hyper-parameter $\tau>0$ 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 solution can be summarized in matrix notation as the solution to the following convex problem:

$$\max_{\alpha,\beta} g(\alpha,\beta) = -\frac{1}{2}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}^T H \begin{bmatrix}\alpha\\ \beta\end{bmatrix} + Y^T\alpha, \tag{8}$$

where H is defined as follows

$$H = \left[\begin{array}{c|c} \Omega + 1/\gamma I_N & (\Omega^+ - \Omega^0) \\ \hline (\Omega^+ - \Omega^0)^T & (\Omega_+^+ - \Omega_+^0 - \Omega_+^{0\,T} + \Omega_0^0) \end{array}\right]. \tag{9}$$

The unique global optimum of the dual function g w.r.t. the Lagrange multipliers α and β incorporating the inequalities $\beta \geqslant 0$ can be found by solving a Quadratic Programming problem (QP) [20].

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

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

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

The incorporation of inequalities in the optimization problem Equation (8) can result in sparseness in the unknowns β [1,2,20] while still achieving the unique global optimum. In the case of the L_2 norm, no sparseness will be present in the so-called support values $(\alpha_i + \beta_{i-1} - \beta_i)$. One can interpret the active (non-sparse) β terms as corrections to the standard LS-SVM which enforce the result to be monotonically increasing. It can happen that after applying an appropriate model selection criterion the resulting estimate with a standard LS-SVM would be monotonically increasing without having to apply the additional constraints. However, a major disadvantage of that approach over the proposed monotone estimate is that feasibility of the monotone optimum is not guaranteed and that the amount of smoothness cannot be varied independently (see e.g. Figure 1.(d)).

2.2. MONOTONE CHEBYCHEV KERNEL REGRESSION

One starts with the same primal model as (2). Consider the Chebychev measure (see [9] and citing papers) for function approximation defined as

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

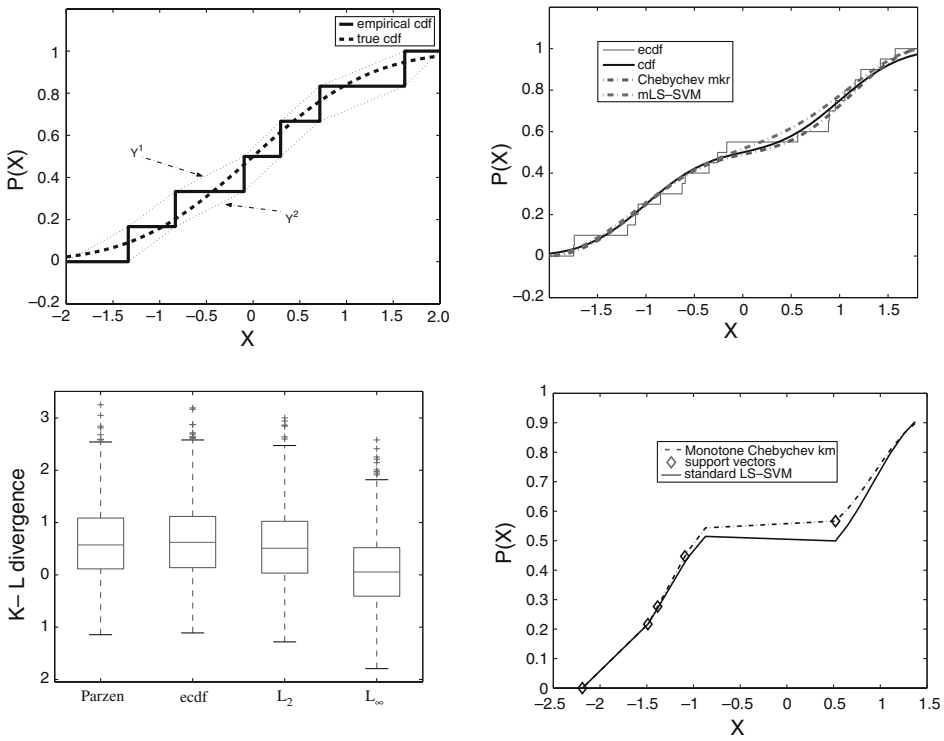


Figure 1. (a) As the ecdf is discontinuous at the sample points, the estimated cdf should lie between the upper- (Y^1) and lower-curve (Y^2) where possible while being smooth. (b) Application of the smooth estimate of the ecdf on the artificial example of Section 4.1. (c) Boxplots of the results of a Monte Carlo simulation for estimating the cdf based on, respectively, the Parzen window, ecdf, the monotone LS-SVM smoother and the monotone Chebychev kernel regressor. (d) Comparison of the smooth monotone Chebychev kernel machine and its sparse representation (using only five support vectors) and a standard LS-SVM which is not guaranteed to be monotone in general.

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

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

As usual in convex optimization (see optimization with L_1 or ϵ-insensitive loss function [20]), the L_∞ norm is translated by minimizing a variable lower- and upperbound $-t \leqslant e_i \leqslant t$ for all $i = 1, \ldots, N$ [20]. For reasons which become clear 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 $0 \leqslant \alpha^+, \ 0 \leqslant \alpha^- \in \mathbb{R}^N$ and $0 \leqslant \beta \in \mathbb{R}^{N-1}$ gives

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

with inequality constraints $\alpha^+,\alpha^-,\beta\geqslant 0$. Elimination of the high dimensional vector w and the scalar t and application of the kernel trick results in the following quadratic programming problem:

$$\max_{\alpha^+,\alpha^-,\beta} g(\alpha^+,\alpha^-,\beta)=-\frac{1}{2}\begin{bmatrix}\alpha^+\\ \hline \alpha^-\\ \hline \beta\end{bmatrix}^T H'\begin{bmatrix}\alpha^+\\ \hline \alpha^-\\ \hline \beta\end{bmatrix}+{Y^1}^T\alpha^+-{Y^2}^T\alpha^-$$
$$\text{s.t.}\quad 1_N^T(\alpha^++\alpha^-)=\gamma,\quad \alpha^+,\alpha^-,\beta\geqslant 0, \tag{14}$$

where the positive semi-definite matrix H' is defined as

$$H'=\left[\begin{array}{c|c|c} \Omega & -\Omega & (\Omega^+-\Omega^0)\\ \hline -\Omega & \Omega & -(\Omega^+-\Omega^0)\\ \hline (\Omega^+-\Omega^0)^T & -(\Omega^+-\Omega^0)^T & (\Omega_+^+-\Omega_+^0-{\Omega_+^0}^T+\Omega_0^0)\end{array}\right]$$

and the different matrices Ω and its variations are defined as in (8). The final model $\hat{f}(x)=\hat{w}^T\varphi(x)$ can be evaluated on a new datapoint x^* as in (10) where $\alpha=\alpha^+-\alpha^-$. Typically, the QP problem will lead to sparseness in the solution α^+,α^- and β. By re-ordering the representation as in (10), one will obtain a reduced set of non-sparse values which one can refer to as support values and corresponding support vectors [1,2,20] comparable with those found in Support Vector Regression (SVR) [1,2].

Remark that the derivation of a non-monotone Chebychev kernel machine may follow along the same lines by omitting the monotonicity constraints in (12). This would result in simpler convex QP problems where the β terms in (14) do not occur. This result is somewhat similar to the SVR formulation without slack variables where the ϵ tuning parameter (as in the Vapnik ϵ-insensitive loss function) can in fact be treated as an additional unknown to the (training) optimization problem [1,2].

3. Applying the Monotone Chebychev Kernel Machine for Smoothing the Ecdf

An application of the previous section is considered to the problem of estimating a smooth approximation to the distribution function of given finite datasample. 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 [10] when adopting the additive model structure [8]. Consider a random variable X with a smooth cdf. For a given realization of the sample $X_1, \ldots, X_N$, say $x_1, \ldots, x_N$, the ecdf is defined as [13]

$$\hat{F}(x) = \frac{1}{N} \sum_{k=1}^{N} I_{(\infty, x]}(x_k) \quad \text{for } -\infty < x < \infty, \tag{15}$$

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. [13]).

In order to obtain a smooth estimate based on the ecdf $\hat{F}$, a function approximation task is considered with input and dependent variables $\{x_i, \hat{y}_i^1, \hat{y}_i^2\}_{i=1}^N$. 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 $(x_1, \ldots, x_N)^T$. In order to handle the intercept term b, one notes that the average of any valid cdf F (given as $\int F(x)\mathrm{d}x$) equals 0.5, which is independent of the exact parameterization of the estimate. This motivates the choice to substract the constant 0.5 from the variables Y^1 and Y^2 as a preprocessing stage (for other appropriate transformations, see e.g. [11]). To make the setup complete, we motivate the use of an L_∞ norm as it forms the basis of the classical Kolmogorov–Smirnoff goodness-of-fit hypothesis test measuring the discrepancy between different cdf's [15].

One may motivate from different from different points of view the choice for the use of the primal-dual kernel machine framework to approach the described smoothing problem: (a) It is both statistically and numerically advantageous to start an estimation process from an unambiguous optimality principle. In this way, optimization issues and modeling assumptions become strictly separated; (b) The primal-dual framework allows for the incorporation of extra hard (linear) (in)equalities while still providing globally optimal solutions; (c) The (sparse) representations of the optimal kernel machine follows from the optimization problem and is globally optimal at the same time. In the primal-dual approach, one can easily incorporate the assumptions as enumerated in the previous paragraph in the estimation process of Equation (14). 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 to the support of F. By deriving a dual expression for this constrained optimization problem, the final optimization problem becomes as in (14) where the following definitions hold: $X^1 = (x_-, x_1, \ldots, x_{N-1})^T$, $X^2 = (x_1, x_2, \ldots, x_N, x_+)^T$, $Y^1 = (0, \hat{y}_1, \ldots, \hat{y}_{N-1})^T$ and $Y^2 = (\hat{y}_1, \ldots, \hat{y}_N, 1)^T$. Furthermore, to impose the equality constraints $w^T\varphi(x_-) = -0.5$ and $w^T\varphi(x_+) = 0.5$ exactly, one can easily see that the equality constraint of (14) should be adapted into $-1_N^T(\tilde{\alpha}^+ + \tilde{\alpha}^-) = \gamma$

where $\tilde{\alpha}^+$ and $\tilde{\alpha}^-$ are defined similar as α^+, α^- but do not contain the multipliers associated with x_+ and x_-.

4. Examples

4.1. EXAMPLE 1: TWO GAUSSIANS

While the main message of this letter concerns the ease of incorporating additional inequality constraints in the derivation of primal-dual kernel machines, some numerical experiments were conducted to motivate the ecdf smoothing application. 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 ecdf. A Monte Carlo experiment with 1000 iterations was conducted relating four cdf estimators (resp. Parzen window estimator, see e.g. [12], ecdf, L_2 smooth monotone kernel machine and smooth monotone Chebychev kernel machine) to the true underlying cdf using the Kullback–Leibler distances. The monotone kernel machines were based on the empirical cdf values down-shifted with the fixed intercept $b = -0.5$ as explained in Section 3. While the L_2 based monotone LS-SVM does not perform significantly better than the classical Parzen window estimator and the empirical cdf, the monotone Chebychev kernel regressor displays increased performance as presented by the boxplots of Figure 1(c). Figure 1(d) displays a realization of this dataset using only 10 datasamples where the standard LS-SVM estimator with tuned regularization and kernel parameters fails to catch the monotonicity. In the case of the monotone Chebychev kernel machine, the active support vector at the right hand side is correcting the non-monotone model ($\beta_i > 0$), enforcing the solution to be strictly increasing.

4.2. EXAMPLE 2: THREE UNIFORM DISTRIBUTIONS

To give a qualitative idea of the difference between the different cdf estimators (resp. ecdf, integrated Parzen window estimator and L_∞ kernel machine), Figure 2 displays the estimates of a complex discontinuous distribution function based on the union of three disjunct uniform parts (dashed-dotted line) with some background noise. While the ecdf is non-smooth in nature and the Parzen window estimate fails to catch the four *knees*, the L_∞ monotone kernel machine leads to a smooth estimate which model the discontinuities.

4.3. EXAMPLE 3: THE SUICIDE DATA

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. [12]) by taking the numerical derivative of the smooth estimate. In this case the support of the data was known

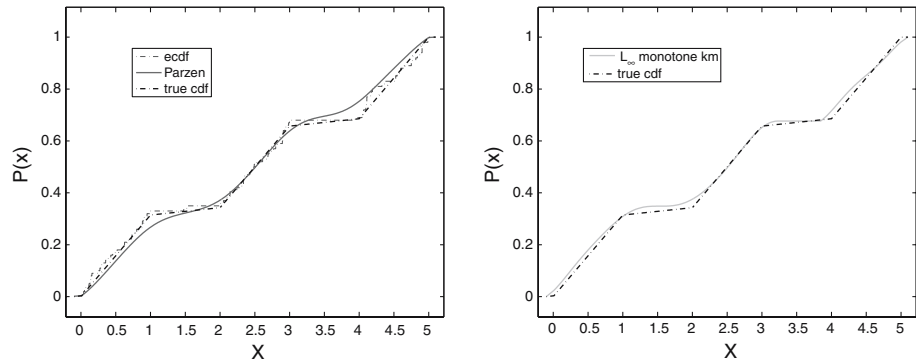


Figure 2. Example of a distribution function estimation task described in Section 4.2. (a) The ecdf is nonsmooth, while the Parzen estimator fails to catch the four knees. (b) Both the monotone L_2 and L_∞ kernel machine succeed in capturing the *knees*.

to have an exact lower bound at 0 which can be nicely incorporated in this framework as shown in Figure 3(b). A main advantage of this technique over the use of the Parzen kernel estimator becomes apparant in this study. As well known in literature, this strictly positive dataset manifest a tri-modal structure [12]. As shown in Figure 3b and c one cannot find a single bandwidth of the Parzen window estimator which result in a plausible density satisfying both constraints, while the monotone Chebychev kernel machine manages to do so in Figure 2.

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 KUL. 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

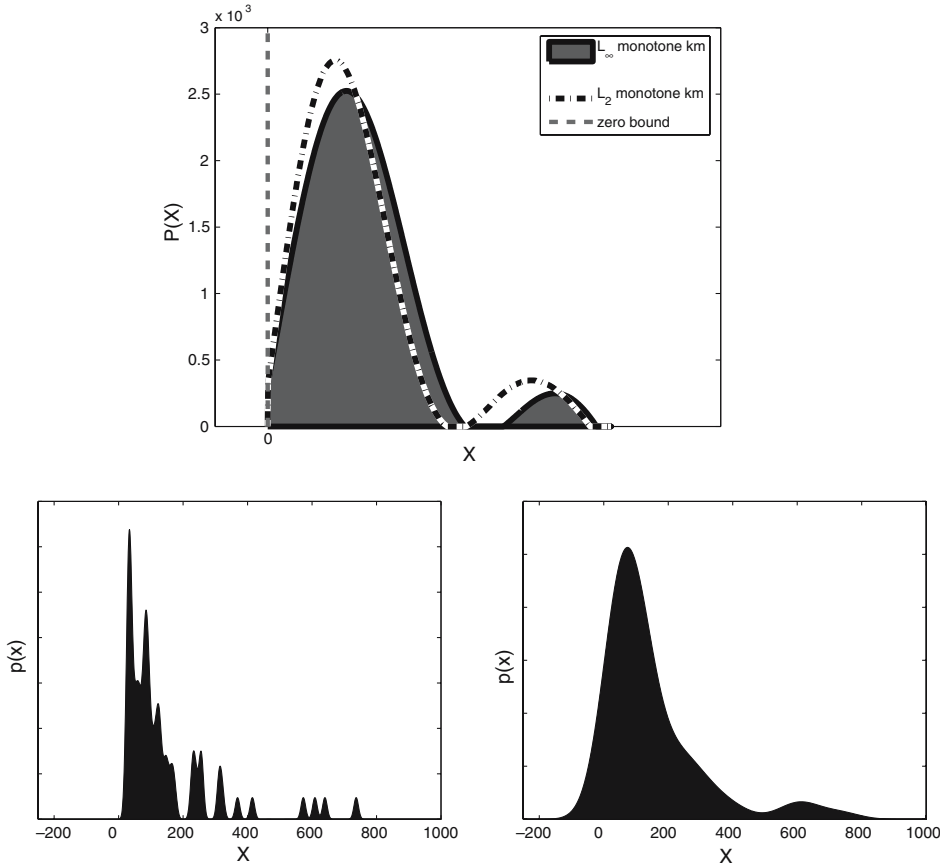


Figure 3. (a) Density estimation of the suicide data using the derivative of the monotone Chebychev kernel regressor and the monotone LS-SVM technique. Both estimates reflect the trimodal structure as well as the positive support. A well-known drawback of the Parzen window estimator in this case is seen in that no single bandwidth parameter of the Parzen window results in both a strictly positive density (one has to under-smooth, (b)) and a smooth trimodal structure (one has to over-smooth,(c)).

(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. Vapnik, V. N.: *Statistical Learning Theory*, New York, Wiley: 1998.
2. Schölkopf, B. and Smola, A.: *Learning with Kernels*, Cambridge, MA, MIT Press: 2002.
3. Poggio, T. and Girosi, F.: Networks for approximation and learning, *Proceedings of the IEEE*, **78**, (september 1990), pp. 1481–1497.

4. Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, 2001.
5. Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P.: A general projection framework for constrained smoothing, *Statistical Science* **16**(3) (2001), 232–248.
6. Suykens, J. A. K., Van Gestel, T. De Brabanter, J. De Moor, B. and Vandewalle, J.: *Least Squares Support Vector Machines*, World Scientific, 2002.
7. Suykens, J. A. K., Horvath, G. Basu, S. Micchelli, C. Vandewalle, J. (eds.), *Advances in Learning Theory: Methods, Models and Applications*, NATO Science Series III: Computer & Systems Sciences, p. 190, IOS Press Amsterdam, 2003.
8. Pelckmans, K. Goethals, I. De Brabanter, J. Suykens, J. A. K. De Moor, B.: Componentwise least squares support vector machines, *Internal Report 04-75, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*, 2004.
9. Chebyshev, P. L.: Sur les questions de minima qui se rattachent la representation approximative des fonctions, *Oeuvres de P. L. Tchebychef*, **1**, 273–378, Chelsea, New York, 1961 (1859).
10. Scott, D. W.: *Multivariate Density Estimation, Theory, Practice and Visualization*, Wiley series inb probability and mathematical statistics, 1992.
11. Gaylord, C. K. Ramirez, D. E.: Monotone regression splines for smoothed bootstrapping, *Computational Statistics Quarterly* **6**(2) (1991), 85–97.
12. Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, 1986, p. 26.
13. Billingsley, P.: *Probability and Measure*, Wiley, New York p. 26, 1986.
14. Devroye, L.: *Non-Uniform Random Variate Generation*, Springer-Verlag, 1986.
15. Conover, W. J.: *Practical Nonparametric Statistics*, New York, Wiley: 1971.
16. De Boor, C. and Schwartz, B.: Piecewise monotone interpolation, *J. Approximation Theory* **21** (1977), 411–416.
17. Ramsay, J. O.: Monotone regression splines in action. *Statistical Science*, **3** (1988) 425–461.
18. Vapnik, V. and Mukherjee, S.: Support vector method for multivariate density estimation In: S. A. Solla T. K. Leen and K.-R. Müller (eds.), In *Advances in Neural Information Processing Systems*, Vol. 12, pp. 659–665, 1999.
19. Tikhonov, A. N. and Arsenin, V. Y.: *Solution of Ill-Posed Problems*, Winston, Washington DC, 1977.
20. Boyd, S. and Vandenberghe, L.: *Convex Optimization*, Cambridge University Press, 2004.