MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1653 April 1999
C.B.C.L Paper No. 170

Support Vector Method for Multivariate Density Estimation

Sayan Mukherjee and Vladimir Vapnik

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1653.ps

Abstract

A new method for multivariate density estimation is developed based on the Support Vector Machine (SVM) solution of inverse ill-posed problems. This algorithm is consistent and results in a sparse solution. This method compared favorably to both Parzen's method and the Gaussian Mixture Model (GMM) method in experiments. The SVM and Parzen's method are shown to be closely related, result in solutions of similar quality, however the SVM solution is sparse. The SVM and GMM approachs are both sparse but the SVM approach is consistent and in general results in more accurate estimates.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from the Office of Naval Research under contract No. N00014-93-13085. Additional support is provided by Eastman Kodak Company, Daimler-Benz AG, Siemens Corporate Research Inc., AT&T, Digital Equipment and Central Research Institute of Electric Power Industry.

1 Introduction

The problem of multivariate density estimation is important for many applications, in particular, for speech recognition [1], [8]. In this paper we introduce two ideas, the Support Vector Method for density estimation, and a new type of kernel whose form depends upon the distance between a point and its nearest neighbor.

There exist two basic approaches to density estimation, parametric approaches and nonparametric approaches.

In the parametric approach one assumes that the unknown density belongs to a parametric set satisfying certain conditions and estimates these parameters based upon the maximum likelihood (ML) principle. Researchers believe that real-life densities can be approximated by a mixture of a small number of elements (Gaussians for Gaussian Mixture Models (GMM)) from the parametric set. Therefore, the following parametric density model was introduced

$$P(t, a, \Sigma) = \sum_{i=1}^{m} \alpha_i P(t, a_i, \Sigma_i), \qquad \alpha \geq 0, \ \sum_{i=1}^{m} \alpha_i = 1, \tag{1}$$

where $P(t, a_i, \Sigma_i)$ are Gaussians with different position vectors a_i and different diagonal covariance matrices Σ_i; α_i is the proportion of the i-th Gaussian in the mixture. It is known [11] that for general forms of Gaussian mixtures the ML estimate does not exist and the method is not consistent. To use the ML method researchers specify two values: a lower bound on diagonal elements of the covariance matrix and an upper bound on the number mixture elements. Under these constraints one can estimate the mixture parameters using the EM algorithm. This solution, however, is based on predefined parameters.

The nonparametric approaches do not have the limitation of assuming that the density belongs to a mixture from a parametric set. The most popular approach, Parzen's method [6], is a consistent estimator [11] of any continuous density. Parzen's method [6], suggests the following estimate given data $x_1, ..., x_\ell$:

$$P(t, \gamma_\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} K_{\gamma_\ell}(t - x_i), \tag{2}$$

where $K_{\gamma_\ell}(t - x_i)$ is smooth function such that $\int K_{\gamma_\ell}(t - x_i)dt = 1$. Under some conditions on parameters γ_ℓ and function $K_{\gamma_\ell}(t - x_i)$, Parzen's method converges with a fast asymptotic rate. A Gaussian with covariance matrix I and one free parameter γ_ℓ^2 (the width) defines such a function

$$K_{\gamma_\ell}(t - x_i) = \frac{1}{(2\pi)^{n/2}\gamma_\ell^n} \exp\left\{-(t - x_i)^T \left(2\gamma_\ell^2 I\right)^{-1} (t - x_i)\right\}. \tag{3}$$

The structure of the Parzen estimator, however, is too complex: the number of terms in (3) is equal to the number of observations (which can be hundreds of thousands).

We introduce an SVM approach to obtain a sparse estimate of a density. The approach has no free parameters. In our experiments it performs as well as Parzen's method but with a sparse solution and outperforms the GMM method. These experiments are carried out on 1, 2, 6, 12 and 40 dimensional distributions. We also prove the consitency of the SVM method.

2 Estimation of Density is an Ill-posed Problem

A density $p(t)$ is defined as the solution of the equation

$$\int_{-\infty}^{x} p(t)\,dt = F(x), \tag{4}$$

where $F(x)$ is the probability distribution function. Estimating a density from data involves solving equation (4) on a given set of densities $p(t,\alpha)$, $\alpha \in \Lambda$, when the distribution function $F(x)$ is unknown but a random i.i.d. sample is given, $x_1, ..., x_\ell$. The empirical distribution function

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} \theta(x - x_i),$$

where $\theta(u)$ is the step-function is a good approximation of the actual distribution $F(x)$. In the one dimensional case, the distribution of the supremum error between $F(x)$ and $F_\ell(x)$ is given by the Kolmogorov-Smirnov distribution

$$P\{\sqrt{\ell}\sup_x |F(x) - F_\ell(x)| > \varepsilon\} = 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\{-2\varepsilon^2 k^2\}. \tag{5}$$

In appendix (A) we prove the distribution of the supremum error is independent of $F(x)$ for multivariate distributions.
Hence, the problem of density estimation can be considered as a problem of solving equation (4) where instead of $F(x)$ one uses the empirical distribution function $F_\ell(x)$ which converges to the true one with the (fast) rate $o(1/\sqrt{\ell})$.
The problem of solving linear operator equation $Ap = F$ with approximation $F_\ell(x)$ is ill-posed. In the 1960's methods were proposed for solving ill-posed problems using approximations F_ℓ converging to F as ℓ increases. The idea of these methods was to introduce a regularizing functional $\Omega(p)$ (semi-continuous, positive functional for which $\Omega(p) \leq c, \;\; c > 0$ is a compactum) and define the solution p_ℓ which is a trade-off between values $\Omega(p)$ and $||Ap - F_\ell||$.
The following two methods which are asymptotically equivalent [14] were proposed by Tikhonov [10] and Phillips [7] respectively

$$\min_p \left[||Ap - F_\ell||^2 + \gamma_\ell \Omega(p)\right], \quad \gamma_\ell > 0, \;\; \gamma_\ell \longrightarrow 0, \tag{6}$$

$$\min_p \Omega(p) \;\; s.t. \;\; ||Ap - F_\ell|| < \varepsilon_\ell, \;\; \varepsilon_\ell > 0, \quad \varepsilon_\ell \longrightarrow 0. \tag{7}$$

Both methods can be generalized for the stochastic case. In particular, it was shown for the Tikhonov method that if function $F_\ell(x)$ converges in probability to $F(x)$ and $\gamma_\ell \longrightarrow 0$ then for sufficiently large ℓ and arbitrary ν and μ the following inequality holds [13], [11]

$$P(\rho_{E_1}(p, p_\ell) > \nu) \leq P(\rho_{E_2}(F, F_\ell) > \sqrt{\gamma_\ell \mu}) \tag{8}$$

where $\ell > \ell_0(\nu, \mu)$ and $\rho_{E_1}(p, p_\ell), \;\; \rho_{E_2}(F, F_\ell)$ are metrics in the spaces p and F.
Since the empirical distribution $F_\ell(x)$ converges to $F(x)$ with the rate $o(\frac{1}{\sqrt{\ell}})$, it follows from (8) that if $\gamma_\ell > o(\frac{1}{\sqrt{\ell}})$ the solutions of equation (4) are consistent. See appendix (B) for the consistency proofs for both the Tikhonov and Phillips methods for density estimation.

3 Choice of Regularization Parameters

To choose the appropriate parameter (γ_ℓ in (6) or ε_ℓ in (7)) Morozov [4] proposed the so called *residual method* which suggests using a regularization parameter (γ_ℓ or ε_ℓ) that defines the solution p_ℓ for which the equality holds

$$||Ap_\ell - F_\ell|| = ||F(x) - F_\ell(x)|| = \sigma_\ell, \tag{9}$$

where σ_ℓ is the known accuracy of approximation of $F(x)$ by $F_\ell(x)$. For density estimation one can obtain a good estimate of σ_ℓ. In the one dimensional case from distribution (5) it follows that $\sigma_\ell = c_1/\sqrt{\ell}$, where c_1 corresponds to the appropriate quantile of this distribution. In the d-dimensional case a procedure for evaluating these quantiles is known [5]. One can either evaluate the appropriate quantile c_d using this procedure or by computer simulations.

In solving the density estimation problem we select a regularization parameter (γ_ℓ or ε_ℓ) for which the solution p_ℓ satisfies constraint (9) for an estimated value σ_ℓ.

The density estimation problem can be solved using either regularization method (6) or (7). It is known [13], [11] that using method (6) with a L_2 norm in image space F and regularization functional (Vp, Vp) where V is a convolution operator, one obtains Parzen's method with kernels defined by operator V. Also it was shown [12] that regularization parameters defined by rule (9) provide an almost optimal asymptotic rate of convergence for method (6). Experiments demonstrate that this rule provides a good density estimate for a fixed sample size [3].

4 SVM for Density Estimation

We now estimate densities using method (7) for which the SVM technique can be applied. We solve equation (4) in a set of functions belonging to a Reproducing Kernel Hilbert Space (RKHS) using the functional

$$\Omega(p) = ||p||_H^2. \tag{10}$$

To define the RKHS one has to define a positive definite kernel $K(x, y)$ and an inner product $(f, g)_H$ in Hilbert space H such that

$$(p(x), K(x, y))_H = p(y) \quad \forall p \in H. \tag{11}$$

Note that any positive definite function $K(x, y)$ has an expansion

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y) \tag{12}$$

where λ_i and $\phi_i(x)$ are eigenvalues and eigenfunctions of $K(x, y)$. Consider the set of functions

$$f(x, c) = \sum_{i=1}^{\infty} c_i \phi_i(x) \tag{13}$$

and the inner product

$$(f(x, c^*), f(x, c^{**}) = \sum_{i=1}^{\infty} \frac{c_i^* c_i^{**}}{\lambda_i}. \tag{14}$$

Kernel (12), inner product (14), and set (13) define a RKHS. Indeed

$$(p(x), K(x,y))_H = \left(\sum_{i=1}^{\infty} c_i\phi_i(x), K(x,y)\right)_H =$$

$$\left(\sum_{i=1}^{\infty} c_i\phi_i(x), \sum_{i=1}^{\infty} \lambda_i\phi_i(x)\phi_i(y)\right)_H = \sum_{i=1}^{\infty} \frac{c_i\lambda_i\phi_i(y)}{\lambda_i} = p(y).$$

For functions from a RKHS the functional (10) has the form

$$\Omega(p) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}, \tag{15}$$

where λ_i is the i-th eigenvalue of the kernel $K(x,y)$. Therefore the choice of the kernel defines smoothness properties on the solution.
To use method (7) for the density estimation problem (4) in a RKHS with a solution satisfying condition (9) we minimize functional

$$\Omega(p) = (p,p)_H$$

subject to constraints

$$\left| F_\ell(x) - \int_{-\infty}^{x} p(t)dt \right|_{x=x_i} \leq \sigma_\ell, \quad 1 \leq i \leq \ell.$$

We look for a solution of equation (4) with the form

$$p(t) = \sum_{i=1}^{\ell} \beta_i K_{\gamma_\ell}(x_i, t). \tag{16}$$

Taking into account (11) and (16) we rewrite functional (10) and minimize

$$\Omega(p,p) = (p,p)_H = \sum_{i,j=1}^{\ell} \beta_i\beta_j K_{\gamma_\ell}(x_i, x_j) \tag{17}$$

subject to constraints

$$\left| F_\ell(x) - \sum_{j=1}^{\ell} \beta_j \int_{-\infty}^{x} K_{\gamma_\ell}(x_j, t)dt \right|_{x=x_i} \leq \sigma_\ell, \quad 1 \leq i \leq \ell. \tag{18}$$

This optimization problem is closely related to the SV regression problem with an ε_ℓ-insensitive zone [11]. It can be solved using the standard SVM technique. Usually most of the β_i values in the SVM estimate will be zero, the x_i values corresponding to the nonzero β_i values are called support vectors.
To obtain the solution as a mixture of densities we choose a kernel that is a density and add two more constraints $\beta_i \geq 0, \quad \sum_{i=1}^{\ell} \beta_i = 1$.
In general kernels have a parameter γ_ℓ, the width. We call the value of this parameter admissible if in (18) at least for one x_i we have a strict equality (the solution satisfies condition (9)). The admissible set $\gamma_{min} \leq \gamma_\ell \leq \gamma_{max}$ is not empty since for Parzen's method (which also has form (16)) such a value does exist. Recall that value γ_ℓ in kernel (2) determines the the smoothness

requirements of the solution: the larger the γ the more rapidly the eigenvalues λ_i of the kernel decrease and therefore functional (15) imposes stronger smoothness requirements. For any admissible γ the SVM technique provides the unique solution with some number of elements in the mixture. We choose the smoothest solution amongst those with the minimum number of mixture elements, number of support vectors. This is done by selecting the γ_ℓ which minimizes $\Omega(p)$ among the γ_ℓ in the admissable set for which the number of support vectors is minimum.

5 Variable Kernels

We now introduce kernels that change their form based upon the distance to their nearest neighbor. These kernels are introduced to imporve performance in high dimensional problems, where data are sparse.

The empirical distribution functions in both Parzen's and the SVM approach was constructed were piecewise constant, a sum of theta functions. This approximation is discontinuous. However, any distribution function that has a density is continuous so it makes sense to construct the empirical distribution function as a monotonic continuous function.

Let us consider the one-dimensional case. Given data $x_1, ..., x_\ell$ which is an ordered set, consider the following distribution function, the linear empirical distribution function:

$$F_\ell^1(x) = \begin{cases} \frac{k}{\ell} + \frac{1}{\ell}\frac{x - x_k - \tau_k/2}{\tau_k} & \text{if } x \in [x_k - \tau_k/2, x_k + \tau_k/2) \\ \frac{k}{\ell} & \text{if } x \in [x_k, x_{k+1}) \text{ and } x \notin [x_k - \tau_k/2, x_k + \tau_k/2), \end{cases} \tag{19}$$

where x_k is the k-th data point and τ_k is the distance between the k-th data point and it's nearest neighbor. (See Figure (1) for a comparison between the linear empirical distribution function and the standard empirical distribution function). In the Parzen's case, minimizing the following functional, the Tikhonov form:

$$R_{\gamma_\ell}(p, F_\ell^1) = \frac{1}{\ell}\sum_{i=1}^{\ell}(\int_{-\infty}^{x} p(t)dt - F_\ell^1(x))^2 + \gamma_\ell(p,p)_H \tag{20}$$

results in a solution of the form

$$p(t) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{2}{\tau_i}\int_{x_i-\tau_i/2}^{x_i+\tau_i/2} K_{\gamma_\ell}(x,t)dx, \tag{21}$$

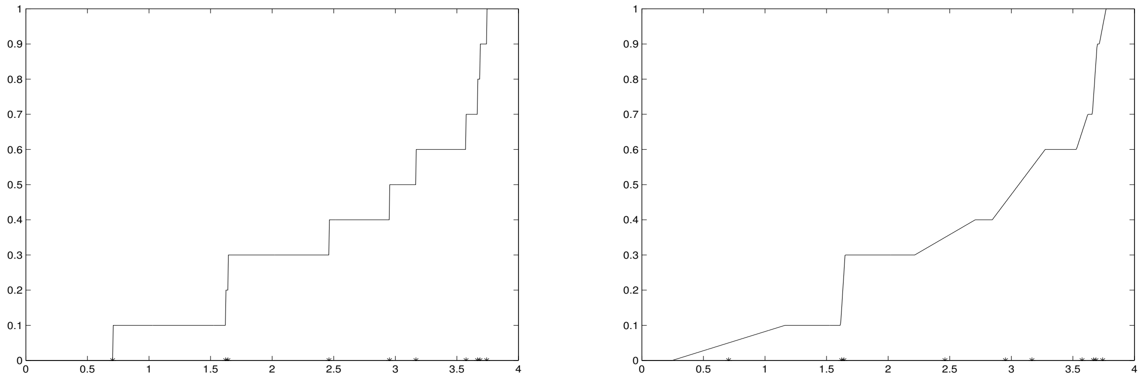
where $K_{\gamma_\ell}(x,t)$ is the kernel corresponding to the RKHS in equation (20).

For the SVM case the following form is minimized:

$$\Omega(p,p) = (p,p)_H = \sum_{i,j=1}^{\ell}\beta_i\beta_j M(x_i,x_j) \tag{22}$$

subject to constraints

$$\left| F_\ell^1(x) - \sum_{j=1}^{\ell}\beta_j\int_{-\infty}^{x} L(x_j,t)dt \right|_{x=x_i} \le \sigma_\ell, \quad 1 \le i \le \ell, \tag{23}$$



(a) (b)

Figure 1: (a) The standard empirical distribution function for 10 points (b) The linear empirical distribution function for the same 10 points.

and $\beta_i \geq 0, \quad \sum_{i=1}^{\ell} \beta_i = 1$ where

$$L(x_j, t) = \frac{1}{\tau_j} \int_{x_j - \tau_j/2}^{x_j + \tau_j/2} K_{\gamma_\ell}(x, t) dx$$

and

$$M(x_i, x_j) = \frac{1}{\tau_i \tau_j} \int_{x_j - \tau_j/2}^{x_j + \tau_j/2} \int_{x_i - \tau_i/2}^{x_i + \tau_i/2} K_{\gamma_\ell}(t, t') dt dt'.$$

The density estimate has the form

$$p(t) = \sum_{i=1}^{\ell} \beta_i L(x_i, t). \tag{24}$$

The width parameter, γ_ℓ, is set in the same way as was done for standard kernels. Appendix (C) lists analytic forms for L and M from various K_{γ_ℓ} and also plots some examples of $L(x_i, t)$.

6 Experiments

We include three sets of experiments. The first two sets serve mainly to illustrate graphically the difference between the three approaches to density estimation (SVM, Parzen, and GMM). The third set is a more statistically thorough comparison of the techniques for a variety of distributions.

The first set examines and plots density estimates for the three approaches for a particular drawing from a one dimensional distribution. The second set examines and plots density estimates for a two dimensional distribution. The third set plots statistical properties of the estimates over many trials for distributions in $\mathbb{R}^2$, $\mathbb{R}^6$, $\mathbb{R}^{12}$, and $\mathbb{R}^{40}$.

6.1 One Dimensional Example

We first compare the SVM method with Parzen's method and the GMM method for a one dimensional density. The choice of the parameter γ for the SVM approach was set by constraint

(9) in which $\varepsilon = \frac{\mu_{\ell d}}{2}$, where $\mu_{\ell d}$ is the mean of the Kolmogorov-Smirnov distribution for ℓ points and dimensionality d. For the Parzen case γ was selected using a leave-one out procedure on the Parzen estimator. The GMM method uses the EM algorithm and sets all parameters except n, the upper bound on the number of terms in the mixture.
The density to be estimated was the following

$$p(t) = .5\frac{1}{\sqrt{2\pi}}e^{-(t-2)^2/2} + .5\frac{.7}{2}e^{-|t+2|}.$$

Figures (2) and (3) show plots of the Parzen, SVM, and GMM estimates for the Gaussian kernel when 20 and 100 points drawn from the distribution are used as training data. The test error is an $L1_p$ norm on 10,000 points drawn from the distribution. Table (1) summarizes the results, nsv is the number of functions in the approximation, points stands for the number of training data. See appendix (D) for a table showing the estimates for a wide variety of kernels.

	points	nsv	σ	γ	$L1_p$
Gaussian Parzen	20	20	$\frac{.36}{\sqrt{20}}$	.55	3.361×10^{-2}
Gausian SVM	20	5	$\frac{.36}{\sqrt{20}}$	1.2	3.128×10^{-2}
GMM	20	2			6.519×10^{-2}
Gaussian Parzen	100	100	$\frac{.33}{\sqrt{100}}$	.6	2.208×10^{-2}
Gausian SVM	100	5	$\frac{.33}{\sqrt{100}}$	1.25	2.165×10^{-2}
GMM	100	4			3.806×10^{-2}
GMM	100	2			5.860×10^{-2}

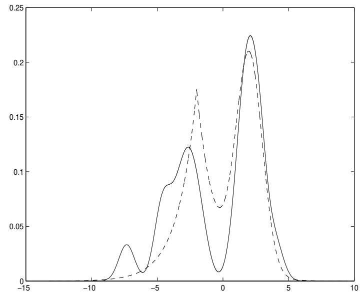
Table 1: Performance of the SVM algorithm, Parzen's windows, and the GMM algorithm for a Gaussian kernel.

6.2 Two Dimensional Example

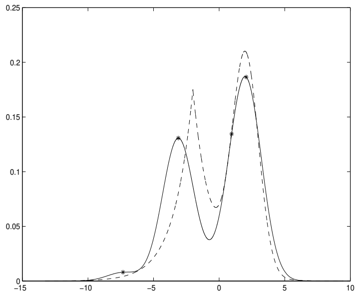
The same comparison was made for a 2-dimensional distribution. The choice of the parameter γ for the SVM approach was set by constraint (9), again set to half the mean of the Kolomogorov-Smirnov distribution. For the Parzen case we again used a leave-one out procedure to set γ. The GMM method uses the EM algorithm and sets all parameters except n, the upper bound on the number of terms in the mixture.
The density to be estimated was the following

$$p(t_x, t_y) = .5\frac{1}{2\pi}e^{-(t_x-2)^2/2}e^{-(t_y-2)^2/2} + .5\frac{.35}{4}e^{-.7|t_x+2|}e^{-.5|t_y+2|}.$$

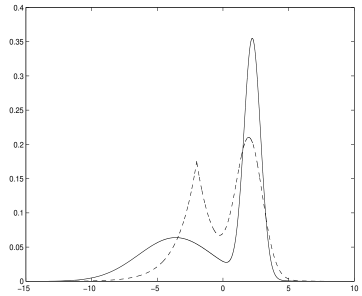
Figure (4) shows plots of the Parzen, SVM, and GMM estimates for the Gaussian kernel when 60 points drawn from the distribution are used as training data. The test error is an $L1_p$ norm on 10,000 points drawn from the distribution. Table (2) summarizes the results, nsv is the number of functions in the approximation, points stands for the number of training data. See appendix (D) for a table showing the estimates for a wide variety of kernels.



(a)



(b)



(c)

Figure 2: (a) The true distribution and estimate for the Parzen case (b) the true distribution and estimate for the SVM case (the asterix show the position of the support vectors) (c) the true distribution and estimate for the GMM case with 4 mixtures for 20 points.

	points	nsv	σ	γ	$L1_p$
Gaussian Parzen	60	60		.72	7.182×10^{-3}
Gausian SVM	60	7	$\frac{.36}{\sqrt{60}}$	1.0	5.690×10^{-3}
GMM	60	2			2.250×10^{-2}
GMM	60	4			1.249×10^{-2}

Table 2: Performance of the SVM algorithm, Parzen's windows, and the GMM algorithm for a Gaussian kernel.

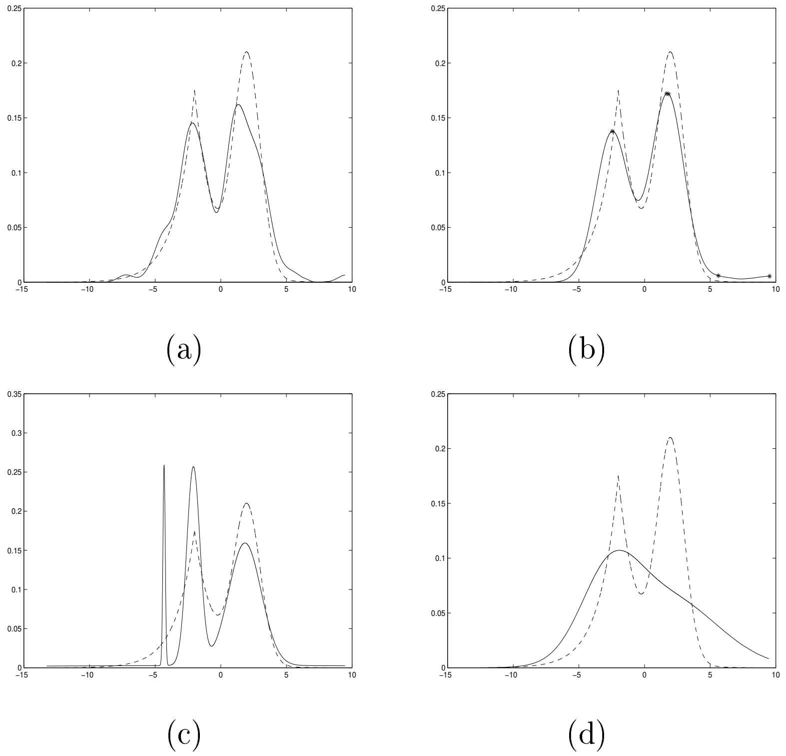
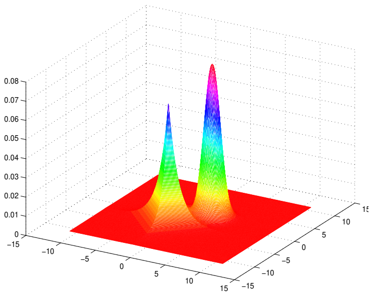
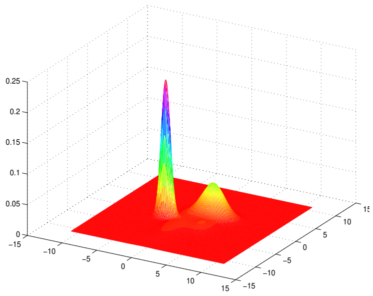


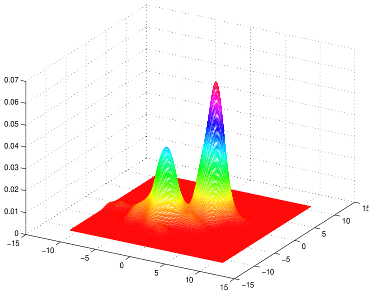
Figure 3: (a) The true distribution and estimate for the Parzen case (b) the true distribution and estimate for the SVM case (the asterix show the position of the support vectors) (c) the true distribution and estimate for the GMM case with 4 mixtures (d) the true distribution and estimate for the GMM case with 2 mixtures for 100 points.



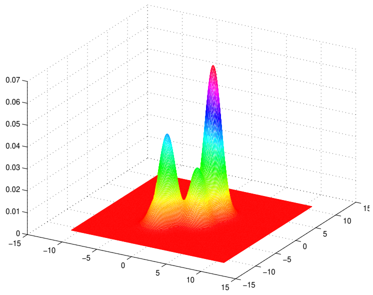
(a)



(b)



(c)



(d)

Figure 4: (a) The true distribution (b) the GMM case with 4 mixtures (c) the Parzen case (d) the SVM case for 60 points.

6.3 Estimates Over Several Trials

To fairly assess the performance of the three approaches it is not enough to look at the error of the estimator for one sampling of ℓ points. Several trials of estimates constructed from sampling the distribution need to be examined. Each trail consisted of sampling ℓ points from the distribution to be estimated and using these points as training data for the estimator. The error of the estimate is measured as the $L1_p$ norm on $10,000$ out of sample points drawn from the distribution. We ran many trials and the made boxplots of the errors. The boxplots describe the distribution of distances between estimates and actual densities. The horizontal lines of the boxplot indicate 5%, 25%, 50%, 75%, and 95% quantiles of the error distribution.

The choice of the parameter γ for the SVM approach was set by constraint (9), again set to half the mean of the Kolomogorov-Smirnov distribution. For Parzen's method γ was set using a leave-one-out procedure. This GMM method sets all parameters except n – the upper bound on the number of terms in the mixture. We used Gaussian kernels in all the methods.

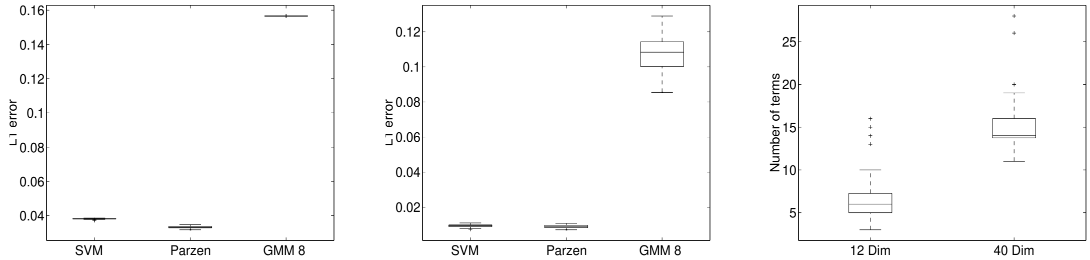
Figure (5a) shows four boxplots describing the results of estimating a density defined by a mixture of two Laplacians in a two dimensional space using 200 observations. Each box-plot shows outcomes of 100 trials: for the SVM method, Parzen's method and the GMM method with parameters $n = 2$, and $n = 4$. Figure (5c) shows the distribution of the number of terms for the SVM method.

Figure (5b) shows boxplots of outcomes of estimating a density defined by the mixture of four Gaussians in a six dimensional space using 600 observations. Each box-plot shows outcomes of 50 trials: for the SVM method, Parzen's method, and the GMM method with parameters $n = 4$, and $n = 8$. Figure (5c) shows the distribution of the number of terms for the SVM method.

Figure (6) and (7) demonstrate the trade-off between accuracy and sparsity for the two dimensional case for the SVM method. Figure (6a) displays the distribution of the $L1_p$ error and figure (6b) displays the distribution of the number of terms for the SVM method with $\gamma_\ell = \gamma_{\max}$, $\gamma_\ell = 1.1$, $\gamma_\ell = .9$, and Parzen's method. Figure (7a) displays the distribution of the $L1_p$ error and figure (7b) displays the distribution of the number of terms when in constraint 9 one uses $\varepsilon = mc_d/\sqrt{\ell}$ with $m = 1, 1.5, 2.3$.

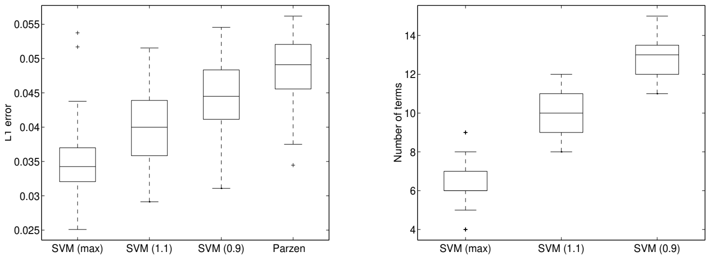
Figure (8a) shows boxplots of outcomes of estimating a density defined by the mixture of four Gaussians and four Laplacians in a twelve dimensional space using 400 observations. Each box-plot shows outcomes of 50 trials: for the SVM method, Parzen's method, and the GMM method with parameter $n = 8$. Figure (8c) shows the distribution of the number of terms for the SVM method.

Figure (8b) shows boxplots of outcomes of estimating a density defined by the mixture of four Gaussians and four Laplacians in a forty dimensional space using 480 observations. Each box-plot shows outcomes of 50 trials: for the SVM method, Parzen's method, and the GMM method with parameter $n = 8$. Figure (8c) shows the distribution of the number of terms for the SVM method.



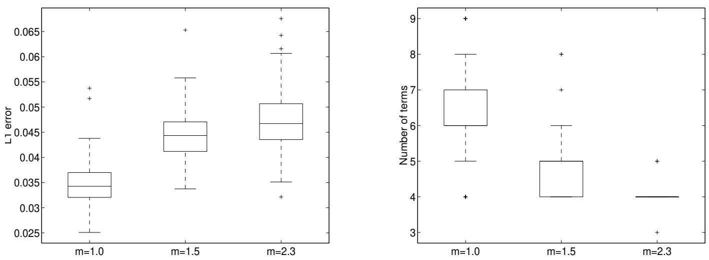
(a) (b) (c)

Figure 5: (a) Boxplots of the $L1_p$ error for the mixture of two Laplacians in two dimensions for the SVM method, Parzen's method, and the GMM method with 2 and 4 Gaussians. (b) Boxplots of the $L1_p$ error for mixture of four Gaussians in six dimensions with the SVM method, Parzen's method, and the GMM method with 4 Gaussians. (c) Boxplots of distribution of the number of terms for the SVM method for the two and six dimensional cases.



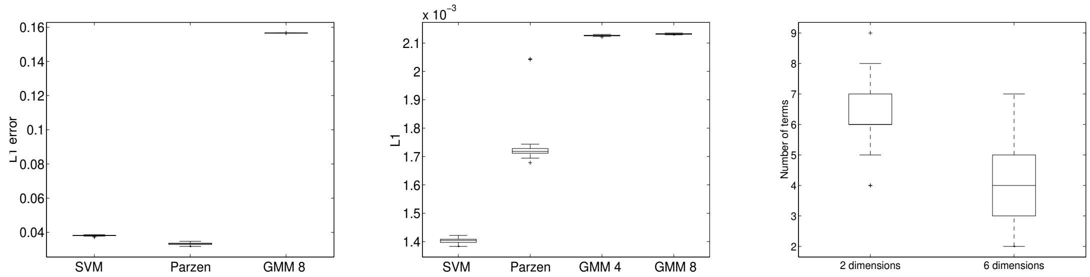
(a) (b)

Figure 6: (a) Boxplots of the $L1_p$ error for the two dimensional case for the SVM method with $(\gamma_\ell = \gamma_{\max})$, $\gamma_\ell = 1.1$, $\gamma_\ell = .9$, and Parzen's method. (b) Boxplots of the distribution of the number of terms.



(a) (b)

Figure 7: (a) Boxplots of the L_1 error for the SVM method for the two dimensional case with $\sigma_\ell = mc_d/\sqrt{\ell}$ with $m = 1, 1.5, 2.3$. (b) A boxplot of the number of terms.



(a) (b) (c)

Figure 8: (a) Boxplots of the $L1_p$ error for the mixture of four Laplacians and four Gaussians in twelve dimensions for the SVM method, Parzen's method, and the GMM method with 8 Gaussians. (b) Boxplots of the $L1_p$ error for mixture of four Gaussians and four Laplacians in forty dimensions with the SVM method, Parzen's method, and the GMM method with 8 Gaussians. (c) Boxplots of distribution of the number of terms for the SVM method for the twelve and forty dimensional cases.

7 Summary

Multivariate density estimation using the Support Vector Method is introduced. This algorithm is sparse in the number of terms in the estimate and seems to be more accurate that the GMM approach over a set of trials. Another advantage of this algorithm is that it has no free parameters: the width parameter in the algorithm is set based upon the Kolmogorov-Smirnov statistic.

A Universality of the Kolmogorov-Smirnov Statistic

We first state Kolmogorov's proof of the unversality of the Kolomogorov-Smirnov statistic in one dimension [2].

Definition A.1 *The probability function $\Phi_\ell(\lambda)$ is defined as follows:*

$$\Phi_\ell(\lambda) = \mathbb{P}\{\sup_x |F_\ell(x) - F(x)| < \lambda/\sqrt{\ell}\},$$

where $F(x)$ is the distribution function and $F_\ell(x)$ is the empirical distribution function.

Lemma A.1 *(Kolmogorov , 1933) The probability function $\Phi_\ell(\lambda)$ is independent of the distribution function $F(x)$ on condition the latter is continuous.*

Proof. Let X be a random variable with continuous distribution function $F(x)$. For the random variable $Y = F(x)$ corresponds the distribution function $F^0(x)$ where $F^0(x) = x_+$. Given empirical distribution functions $F_\ell(x)$ and $F_\ell^0(x)$ for X and Y after ℓ observations, the following hold

$$F_\ell(x) - F(x) = F_\ell^0[F(x)] - F^0[F(x)] = F_\ell^0(y) - F(y)$$
$$\sup_x |F_\ell(x) - F(x)| = \sup_x |F_\ell^0(y) - F^0(y)|.$$

So the probability function $\Phi_\ell(\lambda)$ for any continuous distribution function is identical to that of the function $F^0(x)$. □

We now state the multivariate version of this lemma.

Definition A.2 *The probability function $\Phi_\ell(\lambda)$ is defined as follows:*

$$\Phi_\ell(\lambda) = \mathbb{P}\{\sup_{\mathbf{x}} |F_\ell(\mathbf{x}) - F(\mathbf{x})| < \lambda/\sqrt{\ell}\},$$

where $F(\mathbf{x})$ is the distribution function and $F_\ell(\mathbf{x})$ is the empirical distribution function. If $\mathbf{x} \in \mathbb{R}^d$ then for a random vector $\mathbf{X} \in \mathbb{R}^d$ the distribution function is

$$F(\mathbf{x}) = F(\mathbf{x}_1)F(\mathbf{x}_2|\mathbf{x}_1)F(\mathbf{x}_3|\mathbf{x}_2|\mathbf{x}_1)...F(\mathbf{x}_d|\mathbf{x}_{d-1}|...|\mathbf{x}_2|\mathbf{x}_1)$$

and the empirical distribution function is

$$F_\ell(\mathbf{x}) = F_\ell(\mathbf{x}_1)F_\ell(\mathbf{x}_2)...F_\ell(\mathbf{x}_d)$$

where $F_\ell(\mathbf{x}_i)$ is the empirical distribution for the random variable corresponding to the i-th element in the random vector.

Lemma A.2 *The probability function $\Phi_\ell(\lambda)$ is independent of the multivariate distribution function $F(\mathbf{x})$ on condition the latter is continuous.*

Proof. Let $\mathbf{X}$ be a random vector with continuous distribution function $F(\mathbf{x})$. For the random vector $\mathbf{Y} = T(\mathbf{X})$ where the transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ is as follows

$$\begin{aligned} T_1 : \mathbf{Y}_1 &= F(\mathbf{X}_1) \\ T_2 : \mathbf{Y}_2 &= F(\mathbf{X}_2|\mathbf{X}_1) \\ \vdots &= \vdots \\ T_d : \mathbf{Y}_d &= F(\mathbf{X}_d|\mathbf{X}_{d-1}|...|\mathbf{X}_2|\mathbf{X}_1), \end{aligned}$$

over d dimensions. We now introduce the distribution function $F^0(\mathbf{x})$ such that

$$F^0(\mathbf{x}) = \mathbf{x}_{1+}\mathbf{x}_{2+}...\mathbf{x}_{d+}.$$

Given empirical distribution functions $F_\ell(\mathbf{x})$ and $F^0_\ell(\mathbf{x})$ for $\mathbf{X}$ and $\mathbf{Y}$ after ℓ observations, the following hold

$$F_\ell(\mathbf{x}) - F(\mathbf{x}) = F^0_\ell[T(\mathbf{x})] - F^0[T(\mathbf{x})] = F^0_\ell(\mathbf{y}) - F(\mathbf{y})$$
$$\sup_{\mathbf{x}} |F_\ell(\mathbf{x}) - F(\mathbf{x})| = \sup_{\mathbf{x}} |F^0_\ell(\mathbf{y}) - F^0(\mathbf{y})|.$$

So the probability function $\Phi_\ell(\lambda)$ for any continuous distribution function is identical to that of the function $F^0(\mathbf{x})$. $\square$

We now show that for ℓ large enough $\sup_{\mathbf{x}} |F_\ell(\mathbf{x}) - F(\mathbf{x})|$ scales as $\frac{1}{\sqrt{\ell}}$.
Consider a class of functions $\mathcal{C}$ and a probability space (Ω, Σ, P), let us look at

$$\tau_{\mathcal{C}}(\lambda) = \mathbb{P}\{\sup_{f\in\mathcal{C}} |\sum_{i\leq\ell} f(\mathbf{x}_i) - \ell Ef| \geq \lambda\sqrt{\ell}\}.$$

Theorem A.1 *(Talagrand , 1994)[9] Consider a class $\mathcal{C}$ that is a Vapnik-Červonenkis (VC) class of index $v \geq 1$ then the following holds.*
i) Given $\varepsilon > 0$, and any probability Q on Ω we have

$$N(\mathcal{C}, d_Q, \varepsilon) \leq (\frac{4e}{\varepsilon})^v,$$

where

$$d_Q(C_1, C_2) = Q(C_1 \triangle C_2).$$

ii)Given $\varepsilon > 0$, $\mathcal{C}$ can be covered by at most $(\frac{4e}{\varepsilon})^v$ brackets $[C_1, C_2]$ for which $\mathbb{P}\{C_2 \setminus C_1\} \leq \varepsilon$. Then for all $\lambda > 0$ we have

$$\tau_{\mathcal{C}}(\lambda) \leq (K(4e))^{v+1}\lambda^{2v-1}v^{-v}e^{-2\lambda^2}.$$

To apply this theorem to our problem we need to characterize the empirical distribution function in terms of VC classes. We do this by introducing the class of lower orthants and stating its VC index. This allows us to apply Talagrand's theorem to this case.

The class of orthants in $\mathbb{R}^d$ has a VC index of d.

Theorem A.2 *In $\mathbb{R}^d$ let $\mathcal{Y}$ be the set of all left half-lines $(-\infty, x]$ for $x \in \mathbb{R}$. Let $\mathcal{C} = \{\Pi_{i=1}^d L_i : L_i \in \mathcal{Y}, i = 1, ..., d\}$, $\mathcal{C}$ is the class of lower orthants parallel to the axes. $v(\mathcal{C}) = d$.*

Proof. Let $\mathcal{D}_i$ be classes of subsets of a set X and $\mathcal{D} = \{\bigcap_{i=1}^n D_i : D_i \in \mathcal{D}_i, i = 1, ..., n\}$, where each $\mathcal{D}_i$ is linearly ordered by inclusion. Then $v(\mathcal{D}) = n$.□

If $f(\mathbf{x}_i) = 1_{C(i)}$ where $C(i) \in \mathcal{C}$ then the empirical distribution function is simply $\frac{1}{\ell} \sum_{i \leq \ell} f(\mathbf{x}_i)$ and the distribution function is Ef. So for the class of lower orthants with VC index d we can restate Talagrand's result

Theorem A.3 *Consider the class $\mathcal{C}$ of lower orthants then the following holds. For all $\lambda > 0$ we have*

$$\tau_{\mathcal{C}}(\lambda) \leq (K(4e))^{d+1} \lambda^{2d-1} d^{-d} e^{-2\lambda^2},$$

where

$$\tau_{\mathcal{C}}(\lambda) = \mathbb{P}\{\sup_{\mathbf{x}} |F_\ell(\mathbf{x}) - F(\mathbf{x})| \geq \lambda/\sqrt{\ell}\}.$$

So it is clear that $\sup_{\mathbf{x}} |F_\ell(\mathbf{x}) - F(\mathbf{x})|$ scales as $\frac{1}{\sqrt{\ell}}$.

It is important to realize that we don't use these bounds in our algorithm but rather simulate the multivariate Kolmogorov-Smirnov distribution. The reason for this is the above proof applied a more general result to a more specific case, for this reason the bounds are probably looser than they can be.

B Consistency of Regularization Methods

We prove first the consistency of the Tikhonov method applied to the density estimation problem [11].
Consider the Tikhonov form of the functional

$$R_{\gamma_\ell}(p, F_\ell) = \rho_{E_2}^2(Ap, F_\ell) + \gamma_\ell W(p), \tag{25}$$

where $W(p)$ is a lower semicontinuous functional, $\rho_{E_2}^2(\cdot, \cdot)$ is a metric in the space E_2, $F_\ell = F_\ell(\mathbf{x})$ is a random function (here the empirical distribution function), $\gamma_\ell > 0$ is the regularization parameter, and p_ℓ minimizes the above functional.

Theorem B.1 *The solution $p_\ell(\mathbf{x})$ converges in probability to the true distribution $p(\mathbf{x})$ in a metric $\rho_{E_1}(p_\ell, p)$ if $\gamma_\ell \to 0$ and $\ell\gamma_\ell \to \infty$ as $\ell \to \infty$,where $\mathbf{x} \in \mathbb{R}^d$.*

Proof. We use the two theorems to show the convergence in probability of p_ℓ.

Theorem B.2 *(Vapnik and Stefanyuk , 1978) For any positive ϵ and μ there exists a positive number $n(\epsilon, \mu)$ such that for all $\ell > n(\epsilon, \mu)$ the inequality*

$$\mathbb{P}\{\rho_{E_1}(p_\ell, p) > \epsilon\} \leq \mathbb{P}\{\rho_{E_2}(F_\ell, F) > \sqrt{\gamma_\ell \mu}\}$$

is satisfied.

From theorem A.3 we know for $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{P}\{\rho_{E_2}(F_\ell, F) \geq \varepsilon\} \leq K(\varepsilon\sqrt{\ell})^{2d-1} d^{-d} e^{-2\varepsilon^2 \ell}$$

where K is a constant and $\rho_{E_2}(F_\ell, F) = \sup_{\mathbf{x}} |F(\mathbf{x}) - F_\ell(\mathbf{x})|$.
So for sufficiently large ℓ the inequality

$$\mathbb{P}\{\rho_{E_1}(p_\ell, p) > \epsilon\} \leq K(\gamma_\ell \mu \ell)^{d-1/2} d^{-d} e^{-2\gamma_\ell \mu \ell}$$

holds. So p_ℓ converges in probability to p in the metric $\rho_{E_1}(p, p_\ell)$ if

$$\begin{aligned} \gamma_\ell &\rightarrow 0, \ as\ \ell \rightarrow \infty \\ \gamma_\ell \ell &\rightarrow \infty, \ as\ \ell \rightarrow \infty. \ \square \end{aligned}$$

We now prove the consistency of the Phillip's method applied to the density estimation problem. We have proven that the density estimate p_ℓ minimizing the Tikhonov functional is consistent under certain conditions of γ_ℓ.
Consider the Phillip's form of the functional

$$\inf W(p), \tag{26}$$

such that

$$\rho_{E_2}(Ap, F_\ell) < \delta, \ \ \delta > 0 \tag{27}$$

where $W(p)$ is a lower semicontinuous functional, $\rho^2_{E_2}(\cdot, \cdot)$ is a metric in the space E_2, $F_\ell = F_\ell(\mathbf{x})$ is a random function (here the empirical distribution function), $\gamma_\ell > 0$ is the regularization parameter, and p_ℓ minimizes the above functional.

Theorem B.3 *If p^0_ℓ is the solution that minimizes functional (25) and $\delta = \rho_{E_2}(Ap, F_\ell)$, then p^0_ℓ is the solution of functional (26) subject to constraint (27).*

Proof.
We use Kuhn-Tucker conditions

Theorem B.4 *(Kuhn and Tucker) Let $g(x)$ and $f(x)$ be concave functionals defined on a convex set D of a normed space X and there exists $\bar{x} \in D$ for which $f(\bar{x}) > 0$. Under these assumptions the element x^0 maximizes the functional $g(x)$ for $h(x) \geq 0$ and $x \in D$ if and only if a scalar λ^0 exists such that for the Lagrange function $\psi(x, \lambda) = g(x) + \lambda h(x)$ the inequality*

$$\psi(x, \lambda^0) \leq \psi(x^0, \lambda^0) \leq \psi(x^0, \lambda)$$

holds for all $x \in D$, and $\lambda \geq 0$.

Let $g(f) = -W(f)$ and let $h(f) = -\rho_{E_2}(Af, F_\ell) + \delta$. Since A is a linear operator and $g(f)$ and $h(f)$ are concave functionals we can use the condition in the Kuhn-Tucker theorem. So if a scalar $\lambda^0 \geq 0$ exists such that (f^0, λ^0) is a saddle point for $\psi(f, \lambda) = g(f) + \lambda h(f)$ then the inequalities

$$\psi(f, \lambda^0) \leq \psi(f^0, \lambda^0) \leq \psi(f^0, \lambda) \tag{28}$$

hold.

Let us assume that f^0 minimizes functional (25) for some $\gamma_\ell > 0$, then

$$-W(f) - \lambda^0 \rho_{E_2}(Af, F_\ell) + \lambda^0 \delta \le -W(f_\ell^0) - \lambda^0 \rho_{E_2}(Af_\ell^0, F_\ell) + \lambda^0 \delta,$$

where $\delta = \rho_{E_2}(Af_\ell^0, F_\ell)$, and $\lambda^0 = \frac{1}{\gamma_\ell}$. This satisfies the left side of inequality (28). Now $\delta = \rho_{E_2}(Af_\ell^0, F_\ell)$, so

$$-W(f^0) - \lambda^0 \rho_{E_2}(Af^0, F_\ell) + \lambda^0 \delta \le -W(f_\ell^0) - \lambda \rho_{E_2}(Af_\ell^0, F_\ell) + \lambda \delta,$$

which satisfies the right side of inequality (28). Since the pair (f^0, λ^0) is a saddle point for the Lagrange function $\psi(f, \lambda) = -W(f) + \lambda(-\rho_{E_2}(Af^0, F_\ell) + \delta)$ from the Kuhn-Tucker theorem it follows that f^0 minimizes functional (26) subject to constraint (27). □

C Variable Kernels

In this section we list the analytic forms for the functions $L(x_i, x)$ and $M(x_i, x_j)$ given a variety of kernels $K_{\gamma_\ell}(x, y)$.

(1) Laplacian:

$$K_\gamma(x, y) = \frac{1}{2\gamma} e^{|x-y|/\gamma}.$$

$$L(x_i, x) = \begin{cases} \frac{1}{2\tau_i} e^{-\frac{|x_i - x|}{\gamma}} \cosh(\frac{\tau_i}{2\gamma}) & \text{if } |x - x_i| > \frac{1}{2\tau_i} \\ \frac{1}{\tau_i}(1 - e^{\frac{\tau_i}{2\gamma}} \cosh(x - x_i)) & \text{if } |x - x_i| \le \frac{1}{2}\tau_i. \end{cases}$$

$$M(x_i, x_j) = \begin{cases} \frac{\gamma}{\tau_i \tau_j}(-\frac{\tau_i}{\gamma} + e^{\frac{\tau_i}{\gamma}} - 1) & \text{if } x_i = x_j \\ \frac{4}{\tau_j \tau_i} e^{-\frac{|x_i - x_j|}{\gamma}} \cosh(\frac{\tau_i}{2\gamma}) \cosh(\frac{\tau_j}{2\gamma}) & \text{otherwise.} \end{cases}$$

(2) Epanechnikov:

$$K_\gamma(x, y) = \begin{cases} \frac{3}{4\sqrt{\gamma}}(1 - \frac{1}{\gamma}(x - y)^2) & \text{if } |x - y| \le \sqrt{\gamma} \\ 0 & \text{otherwise.} \end{cases}$$

$$L(x_i, x) = \begin{cases} \frac{3}{4\sqrt{\gamma}\tau_i}(x_i + \frac{1}{2}\tau_i - n + \frac{1}{3\gamma}((x - x_i - \frac{1}{2}\tau_i)^3 - (x - n)^3)) & \text{if } x_i + \frac{1}{2}\tau_i < x \\ \frac{3}{4\sqrt{\gamma}\tau_i}(m - n + \frac{1}{3\gamma}((x - n)^3 - (m - x)^3) & \text{if } x_i - \frac{1}{2}\tau_i \le x \le x_i + \frac{1}{2}\tau_i \\ \frac{3}{4\sqrt{\gamma}\tau_i}(m - x_i + \frac{1}{2}\tau_i + \frac{1}{3\gamma}((m - x)^3 - (x_i - \frac{1}{2}\tau_i - x)^3)) & \text{if } x < x_i - \frac{1}{2}\tau_i \\ 0 & \text{if } |x_i - x| > \sqrt{\gamma} + \frac{1}{2}\tau_i, \end{cases}$$

where $m = \min(x_i + \frac{1}{2}\tau_i, x + \sqrt{\gamma})$ and $n = \max(x_i - \frac{1}{2}\tau_i, x - \sqrt{\gamma})$. For $x_i \ge x_j$ (note since $M(x_i, x_j)$ is symmetric this case is sufficient for evaluating the entire matrix). If $x_i = x_j$

$$M(x_i, x_j) = \begin{cases} \frac{3}{4\tau_i^2}(2\tau_i - \sqrt{\gamma}) & \text{if } \sqrt{\gamma} < \frac{1}{2}\tau_i \\ \frac{3}{4\tau_i}(2\sqrt{\gamma} - \tau_i) + \frac{3}{4\sqrt{\gamma}\tau_i^2}(-\frac{\gamma}{2} - \frac{\tau_i^4}{6\gamma} + 2\frac{\tau_i^3\sqrt{\gamma}}{3}) & \text{if } \frac{1}{2}\tau_i \ge \sqrt{\gamma} \ge \tau_i \\ \frac{3}{4\sqrt{\gamma}}(1 - \frac{\tau_i^2}{6\gamma}) & \text{if } \sqrt{\gamma} > \tau_i. \end{cases}$$

If $x_i > x_j$

$$M(x_i,x_j) = \begin{cases} \frac{\gamma + x_j - x_i}{\gamma^2} & \text{if } \sqrt{\gamma} \geq x_i - x_j + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ \frac{1}{\gamma^2\tau_i\tau_j}((\frac{1}{2}d^2 + \frac{1}{2}\gamma^2 + d\gamma)(b-a) & \text{if } \sqrt{\gamma} \leq x_i - x_j - \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ -\frac{1}{2}(\gamma+d)(b^2-a^2) + \frac{1}{6}(b^3-a^3)) & \\ \tau_j(x_j+\gamma)(c-a) - \tau_j^2(c^2 - \frac{1}{2}a^2) & \text{if } x_i - x_j - \frac{1}{2}\tau_i + \frac{1}{2}\tau_j < \sqrt{\gamma} < x_i - x_j + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ +\frac{1}{\gamma^2\tau_i\tau_j}((\frac{1}{2}d^2 + \frac{1}{2}\gamma^2 + \gamma d)(b-c) & \\ -\frac{1}{2}(\gamma+d)(b^2-c^2) + \frac{1}{6}(b^3-c^3)) & \\ 0 & \text{if } |x_i - x| > \sqrt{\gamma} + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j, \end{cases}$$

where $a = x_i - \frac{1}{2}\tau_i$, $b = \min(x_i + \frac{1}{2}\tau_i, x_j + \frac{1}{2}\tau_j + \sqrt{\gamma})$, $d = x_j + \frac{1}{2}\tau_j$, and $c = x_j - \tau_j + \sqrt{\gamma}$.
(3) 1st order β-spline:

$$K_\gamma(x,y) = \sum_{i=0}^{2} -1^i \binom{2}{i} (x - y + \gamma(1-i))_+$$

$$L(x_i,x) = \begin{cases} \frac{1}{\gamma^2\tau_i}(\frac{1}{2}(m_1^2 - n_1^2) + (\gamma - x)(m_1 - n_1))\Theta(m_1 - n_1) & \\ -\frac{2}{\gamma^2\tau_i}(\frac{1}{2}(m_1^2 - n_2^2) + x(n_2 - m_1))\Theta(m_1 - n_2) & \text{if } |x - x_i| \leq \gamma + \frac{1}{2}\tau_i \\ 0 & \text{if } |x_i - x_j| > \gamma + \frac{1}{2}\tau_i, \end{cases}$$

where $m_1 = \min(x_i + \frac{1}{2}\tau_i, x + \gamma)$, $n_1 = \max(x_i - \frac{1}{2}\tau_i, x - \gamma)$, and $n_2 = \max(x_i - \frac{1}{2}\tau_i, x)$. For $x_i \geq x_j$ (note since $M(x_i,x_j)$ is symmetric this case is sufficient for evaluating the entire matrix).
If $x_i = x_j$

$$M(x_i,x_j) = \begin{cases} \frac{1}{\gamma} - \frac{\tau_i}{3\gamma^2} & \text{if } \gamma > \tau_i \\ \frac{1}{\gamma\tau_i^2}(\tau_i^2 - \gamma^2) + \frac{1}{\gamma\tau_i}(2\gamma - \tau_i) - \frac{1}{\tau_i^2}(\tau_i - \gamma) - \frac{1}{3\tau_i^2}\gamma^3 & \text{if } \frac{1}{2}\tau_i \leq \gamma \leq \tau_i \\ \frac{16\gamma}{6\tau_i^2} + \frac{2}{\tau_i^2}(\tau_i - 2\gamma) + \frac{1}{\tau_i^2}(\gamma - \tau_i) & \text{if } \gamma < \frac{1}{2}\tau_i. \end{cases}$$

If $x_i > x_j$

$$M(x_i,x_j) = \begin{cases} \frac{1}{\gamma^2}(\gamma + x_j - x_i) & \text{if } \gamma \geq x_i - x_j + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ \frac{1}{\gamma^2\tau_i\tau_j}((\frac{1}{2}d^2 + \frac{1}{2}\gamma^2 + \gamma d)(b-a) & \text{if } \gamma \leq x_i - x_j - \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ -\frac{1}{2}(\gamma+d)(b^2-a^2) + \frac{1}{6}(b^3-a^3)) & \\ \tau_j(x_j+\gamma)(b_1-a) - \tau_j^2(b_1^2 - \frac{1}{2}a^2) & \text{if } x_i - x_j - \frac{1}{2}\tau_i + \frac{1}{2}\tau_j < \gamma < x_i - x_j + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j \\ +\frac{1}{\gamma^2\tau_i\tau_j}((\frac{1}{2}d^2 + \frac{1}{2}\gamma^2 + \gamma d)(b-a_1) & \\ -\frac{1}{2}(\gamma+d)(b^2-a_1^2) + \frac{1}{6}(b^3-a_1^3)) & \\ 0 & \text{if } |x_i - x_j| > \gamma + \frac{1}{2}\tau_i + \frac{1}{2}\tau_j, \end{cases}$$

where $a = x_i - \frac{1}{2}\tau_i$, $b = \min(x_i + \frac{1}{2}\tau_i, x_j + \frac{1}{2}\tau_j + \gamma)$, $d = x_j + \frac{1}{2}\tau_j$, $a_1 = x_j + \frac{1}{2}\tau_j + \gamma$, and $b_1 = x_j - \tau_j + \gamma$.
(4) 3rd order β-spline:

$$K_\gamma(x,y) = \sum_{i=0}^{4} \frac{-1^i}{3} \binom{4}{i} (x - y + \gamma(2-i))_+$$

$$L(x_i,x)=\begin{cases}\frac{1}{2\gamma^4\tau_i}(-\frac{1}{12}((x-m_1+2\gamma)^4-(x-n+2\gamma)^4)\Theta(m_1-n) & \text{if } |x-x_i|\le\gamma+\frac{1}{2}\tau_i\\ +\frac{1}{3}((x-m_2+\gamma)^4-(x-n+\gamma)^4)\Theta(m_2-n)- & \\ \frac{1}{2}((x-m_3)^4-(x-n)^4)\Theta(m_3-n)+ & \\ \frac{1}{3}((x-m_4-\gamma)^4-(x-n-\gamma)^4)\Theta(m_4-n) & \\ -\frac{1}{12}((x-m_5-2\gamma)^4-(x-n-2\gamma)^4)\Theta(m_5-n)) & \\ 0 & \text{if } |x-x_i|>\gamma+\tau_i/2,\end{cases}$$

where $m_1=\min(x_i+\frac{1}{2}\tau_i,x+2\gamma)$, $m_2=\min(x_i+\frac{1}{2}\tau_i,x+\gamma)$, $m_3=\min(x_i+\frac{1}{2}\tau_i,x)$, $m_4=\min(x_i+\frac{1}{2}\tau_i,x-\gamma)$, $m_5=\min(x_i+\frac{1}{2}\tau_i,x-2\gamma)$, and $n=x_i-\frac{1}{2}\tau_i$. For $x_i\ge x_j$ (note since $M(x_i,x_j)$ is symmetric this case is sufficient for evaluating the entire matrix).

$$M(x_i,x_j)=\begin{cases}\frac{1}{2\gamma^4\tau_i^2}(\frac{1}{60}(\tau_i-2\gamma)^5\Theta(\tau_i-2\gamma)-\frac{1}{15}(\tau_i-\gamma)^5\Theta(\tau_i-\gamma) & \text{if } x_i=x_j\\ +\frac{1}{10}\tau_i^5+\frac{2}{15}\gamma^5-\frac{1}{15}(\tau_i+\gamma)^5-\frac{1}{15}(c_1-x_i-\frac{1}{2}\tau_i+\gamma)^5 & \\ -\frac{1}{30}(2\gamma)^5+\frac{1}{60}(\tau_i+2\gamma)^5+\frac{1}{60}(c_2-x_i-\frac{1}{2}\tau_i+2\gamma)^5) & \\ \frac{1}{2\gamma^4\tau_i^2}(p_1+p_2+p_3+p_4+p_5) & \text{if } |x_j-x_i|\le 2\gamma+\frac{1}{2}\tau_i+\frac{1}{2}\tau_j\\ 0 & \text{if } |x_j-x_i|>2\gamma+\frac{1}{2}\tau_i+\frac{1}{2}\tau_j,\end{cases}$$

where $c_1=\max(x_i-\frac{1}{2}\tau_i,x_i+\frac{1}{2}\tau_i-\gamma)$, $c_2=\max(x_i-\frac{1}{2}\tau_i,x_i+\frac{1}{2}\tau_i-2\gamma)$, and

$$\begin{aligned}p_1&=\frac{1}{60}(-(a-c+2\gamma)^5+(b-c+2\gamma)^5+(a-d+2\gamma)^5-(b-d+2\gamma)^5)\\ p_2&=\frac{1}{15}((a-c+\gamma)^5-(b-c+\gamma)^5-(a-d+\gamma)^5+(b-d+\gamma)^5)\\ p_3&=\frac{1}{10}(-(a-c)^5+(b-c)^5+(a-d)^5-(b-d)^5)\end{aligned}$$

if $b>c+\gamma$

$$p_4=\frac{1}{15}((a-c-\gamma)^5-(b-c-\gamma)^5-(a-d-\gamma)^5+(b-d-\gamma)^5)$$

else

$$p_4=\frac{1}{15}(-(e_1-d-\gamma)^5+(g_1-d-\gamma)^5\Theta(e_1-g_1)+(a-c-\gamma)^5-(e_1-c-\gamma)^5-(a-d-\gamma)^5+(e_1-d-\gamma)^5)$$

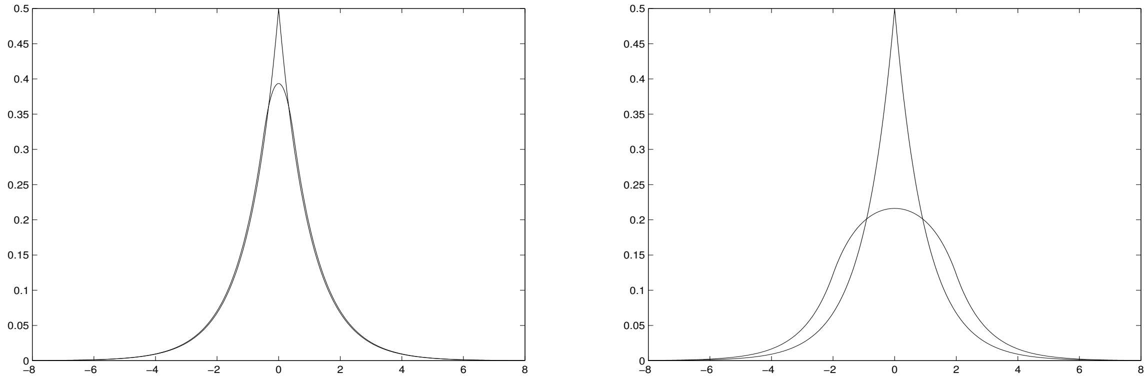
if $b>c+2\gamma$

$$p_5=\frac{1}{60}(-(a-c-2\gamma)^5+(b-c-2\gamma)^5+(a-d-2\gamma)^5-(b-d-2\gamma)^5)$$

else

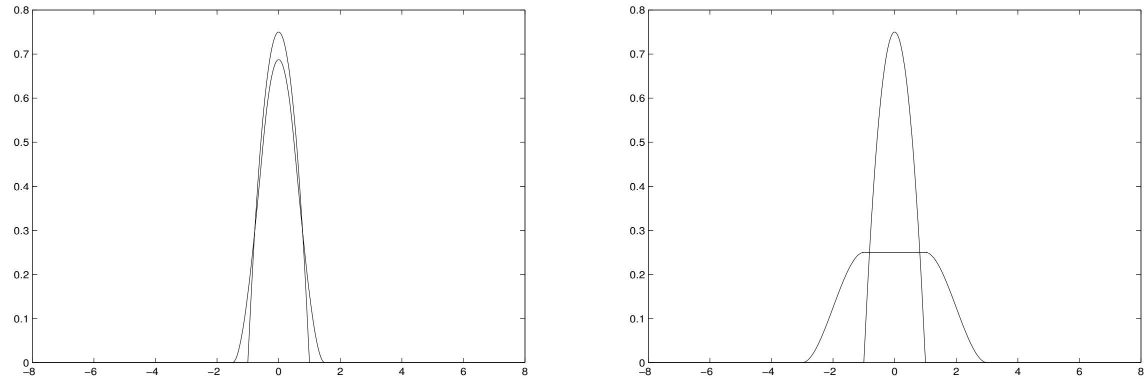
$$p_5=\frac{1}{60}((e_2-d-2\gamma)^5-(g_2-d-2\gamma)^5\Theta(e_2-g_2)-(a-c-2\gamma)^5+(e_2-c-2\gamma)^5+(a-d-2\gamma)^5-(e_2-d-2\gamma)^5),$$

where $a=x_i+\frac{1}{2}\tau_i$, $b=x_i-\frac{1}{2}\tau_i$, $c=x_j+\frac{1}{2}\tau_j$, $d=x_j-\frac{1}{2}\tau_j$, $e_1=\min(c+\gamma,a)$, $g_1=\min(b,d+\gamma)$, $e_2=\min(c+2\gamma,a)$, and $g_2=\min(b,d+2\gamma)$.



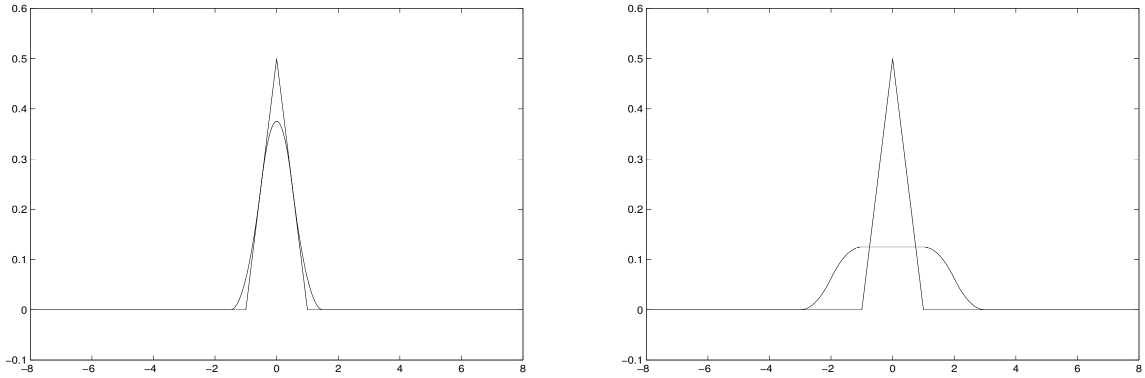
(a) (b)

Figure 9: (a) Plots of a Laplacian kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 1$. (b) Plots of a Laplacian kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 4$.



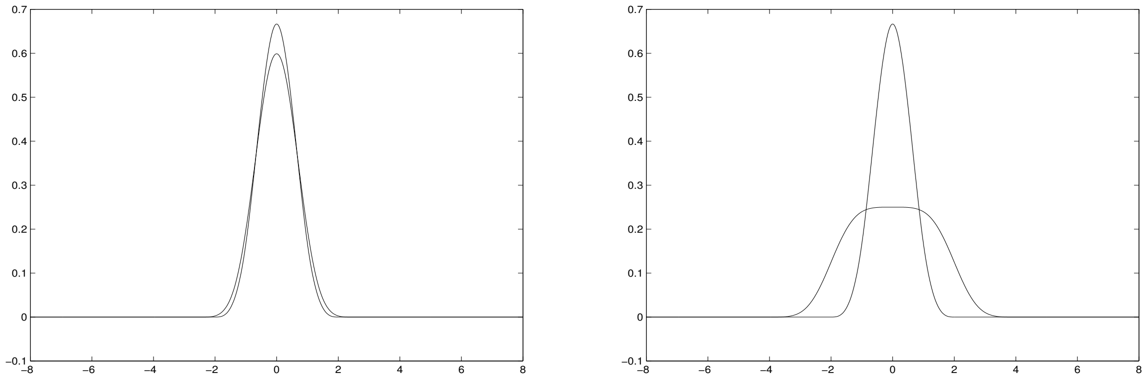
(a) (b)

Figure 10: (a) Plots of an Epanechnikov kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 1$. (b) Plots of a Epanechnikov kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 4$.



(a) (b)

Figure 11: (a) Plots of a first order β-spline kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 1$. (b) Plots of a first order β-spline kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 4$.



(a) (b)

Figure 12: (a) Plots of a third order β-spline kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 1$. (b) Plots of a thid order β-spline kernel and it's corresponding variable kernel with $\gamma = 1$ and $\tau_i = 4$.

D Tables of Estimates

1-Dimensional Example

We summarize the results of estimating the one-dimensional distribution in the experimental section using various kernels.
Tables (3) amd (4) summarizes the results: nsv stands for the number of functions in the estimate, points stands for the number of training points, Lap is an abbreviation for Laplacian, Gaus is an abbreviation for Gaussian, Epa is an abbreviation for Epanechnikov, and Var stands for variable meaning the variable versions of a particular kernel is used.

	points	nsv	σ	γ	$L1_p$
Lap Parzen	20	20	$\frac{.36}{\sqrt{20}}$	.5	3.223×10^{-2}
Lap SVM	20	13	$\frac{.36}{\sqrt{20}}$	1.25	3.331×10^{-2}
Var Lap Parzen	20	20	$\frac{.36}{\sqrt{20}}$	.5	3.232×10^{-2}
Var Lap SVM	20	13	$\frac{.36}{\sqrt{20}}$	1.25	3.324×10^{-2}
Gaus Parzen	20	20	$\frac{.36}{\sqrt{20}}$	.55	3.361×10^{-2}
Gaus SVM	20	6	$\frac{.36}{\sqrt{20}}$	1.2	3.128×10^{-2}
Epa Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.55	3.311×10^{-2}
Epa SVM	20	6	$\frac{.36}{\sqrt{20}}$	4.2	3.275×10^{-2}
Var Epa Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.65	3.290×10^{-2}
Var Epa SVM	20	6	$\frac{.36}{\sqrt{20}}$	4.2	3.164×10^{-2}
1st Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.45	3.214×10^{-2}
1st SVM	20	12	$\frac{.36}{\sqrt{20}}$	2.95	3.510×10^{-2}
Var 1st Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.45	3.237×10^{-2}
Var 1st SVM	20	4	$\frac{.36}{\sqrt{20}}$	2.95	3.534×10^{-2}
3rd Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.0	3.252×10^{-2}
3rd SVM	20	5	$\frac{.36}{\sqrt{20}}$	2.15	3.440×10^{-2}
Var 3rd Parzen	20	20	$\frac{.36}{\sqrt{20}}$	1.05	3.200×10^{-2}
Var 3rd SVM	20	7	$\frac{.36}{\sqrt{20}}$	2.1	3.050×10^{-2}
GMM	20	2			6.519×10^{-2}

Table 3: Performance of the SVM algorithm, Parzen's windows, and the GMM algorithm for various kernels and 20 training points.

2-Dimensional Example

We summarize the results of estimating the two-dimensional distribution in the experimental section using various kernels.
Table (5) summarizes the results: nsv stands for the number of functions in the estimate, points stands for the number of training points, Lap is an abbreviation for Laplacian, Gaus is an abbreviation for Gaussian, Epa is an abbreviation for Epanechnikov, and Var stands for variable meaning the variable versions of a particular kernel is used.

	points	nsv	σ	γ	$L1_p$
Lap Parzen	100	100	$\frac{.33}{\sqrt{100}}$	.5	2.423×10^{-2}
Lap SVM	100	37	$\frac{.33}{\sqrt{100}}$	1.25	2.044×10^{-2}
Var Lap Parzen	100	100	$\frac{.33}{\sqrt{100}}$	.5	2.423×10^{-2}
Var Lap SVM	100	42	$\frac{.33}{\sqrt{100}}$	1.25	1.939×10^{-2}
Gaus Parzen	100	100	$\frac{.33}{\sqrt{100}}$	.6	2.208×10^{-2}
Gaus SVM	100	5	$\frac{.33}{\sqrt{100}}$	1.25	2.165×10^{-2}
Epa Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.1	2.184×10^{-2}
Epa SVM	100	9	$\frac{.33}{\sqrt{100}}$	2.9	2.338×10^{-2}
Var Epa Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.1	2.181×10^{-2}
Var Epa SVM	100	9	$\frac{.33}{\sqrt{100}}$	2.9	2.095×10^{-2}
1st Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.4	2.154×10^{-2}
1st SVM	100	24	$\frac{.33}{\sqrt{100}}$	2.75	2.105×10^{-2}
Var 1st Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.4	2.152×10^{-2}
Var 1st SVM	100	5	$\frac{.33}{\sqrt{100}}$	2.75	1.966×10^{-2}
3rd Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.0	2.178×10^{-2}
3rd SVM	100	4	$\frac{.33}{\sqrt{100}}$	2.1	2.041×10^{-2}
Var 3rd Parzen	100	100	$\frac{.33}{\sqrt{100}}$	1.0	2.177×10^{-2}
Var 3rd SVM	100	4	$\frac{.33}{\sqrt{100}}$	2.1	2.053×10^{-2}
GMM	100	2			5.860×10^{-2}
GMM	100	4			3.806×10^{-2}

Table 4: Performance of the SVM algorithm, Parzen's windows, and the GMM algorithm for various kernels and 100 training points.

	points	nsv	σ	γ	$L1_p$
Lap Parzen	60	60		.72	8.635×10^{-3}
Lap SVM	60	30	$\frac{.36}{\sqrt{60}}$	.85	8.176×10^{-2}
Var Lap Parzen	60	60		.68	8.335×10^{-3}
Var Lap SVM	60	31		1.25	8.145×10^{-3}
Gaus Parzen	60	60		.72	7.182×10^{-3}
Gaus SVM	60	7	$\frac{.36}{\sqrt{60}}$	1.0	5.690×10^{-3}
Epa Parzen	60	60		2.24	6.997×10^{-3}
Epa SVM	60	13	$\frac{.36}{\sqrt{60}}$	3.8	8.212×10^{-3}
Var Epa Parzen	60	60		2.16	6.926×10^{-3}
Var Epa SVM	60	11	$\frac{.36}{\sqrt{60}}$	3.39	7.840×10^{-3}
1st Parzen	60	60		1.8	7.265×10^{-3}
1st SVM	60	8	$\frac{.36}{\sqrt{60}}$	2.3	5.151×10^{-3}
Var 1st Parzen	60	60		1.68	6.981×10^{-3}
Var 1st SVM	60	8	$\frac{.36}{\sqrt{60}}$	2.3	5.095×10^{-3}
3rd Parzen	60	60		1.2	7.049×10^{-3}
3rd SVM	60	6	$\frac{.36}{\sqrt{60}}$	1.8	6.076×10^{-3}
Var 3rd Parzen	60	60	$\frac{.36}{\sqrt{60}}$	1.2	7.038×10^{-3}
Var 3rd SVM	60	7	$\frac{.36}{\sqrt{60}}$	1.8	6.004×10^{-3}
GMM	60	2			2.250×10^{-2}
GMM	60	4			1.249×10^{-2}

Table 5: Performance of the SVM algorithm, Parzen's windows, and the GMM algorithm for various kernels and 60 training points.

References

[1] S. Basu and C.A. Micchelli. Parametric density estimation for the classification of acoustic feature vectors in speech recognition. In *Nonlinear Modeling, Advanced Black-Box Techniques*. Kluwer Publishers, 1998.

[2] A. Kolmogorov. Sulla determinazione empirica di una legge di diztribuzione. *Ist. Ital. Attuari*, pages 4–11, 1933.

[3] N.M. Markovich. Experimental analysis of nonparametric probability density estimates ans of methods for smoothing them. *Avtomatika i Telemekhanika*, (7):110–119, July 1989.

[4] V.A. Morozov. *Methods for solving incorrectly posed problems*. Springer-Verlag, Berlin, 1984.

[5] S. Paramasamy. On multivariate kolmogorov-smirnov distribution. *Statistics & Probability Letters*, 15:140–155, 1992.

[6] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statis.*, 33:1065–1076, 1962.

[7] D.L. Phillips. A technique for the numerical solution of integral equations of the first kind. *J.Assoc. Comput. Machinery*, 9:84–97, 1962.

[8] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans on Speech and Audio Processing*, 3(1):1–27, 1995.

[9] M. Talagrand. Sharper bounds for gaussian and empirical processes. *Ann. Probab.*, 22, 1994.

[10] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[11] V. N. Vapnik. *Statistical learning theory.* J. Wiley, 1998.

[12] V.N. Vapnik, N.M. Markovich, and A.R. Stefanyuk. Rate of convergence in L_2 of the projection estimator of the distribution density. *Avtomatika i Telemekhanika*, (5):64–74, April 1992.

[13] V.N. Vapnik and A.R. Stefanyuk. Nonparametric methods for restoring probability densities. *Avtomatika i Telemekhanika*, (8):38–52, 1978.

[14] V.V. Vasin. Relationship of several variational methods for the approximate solution of ill-posed problems. *Math Notes*, 7:161–166, 1970.