Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 939–946

A Note on Nonparametric Density Estimation for the Deconvolution Problem†

Sungho Lee[1)]

Abstract

In this paper the support vector method is presented for the probability density function estimation when the sample observations are contaminated with random noise. The performance of the procedure is compared to kernel density estimates by the simulation study.

Keywords: Nonparametric density estimation; deconvolution; kernel estimator; support vector; reproducing kernel Hilbert space(RKHS).

1. Introduction

Let X and Z be independent random variables with density functions $f(x)$ and $q(z)$, respectively, where $f(x)$ is unknown and $q(z)$ is known. One observes a sample of random variables $Y_i = X_i + Z_i,\ i = 1, 2, \ldots, n$. The objective is to estimate the density function $f(x)$ where $g(y)$ is the convolution of $f(x)$ and $q(z)$, $g(y) = (f * q)(y) = \int_{-\infty}^{\infty} f(y - z)q(z)dz$. The problem of measurements being contaminated with noise exists in many different fields (see, for example, Louis, 1991; Zhang, 1992). The most popular approach to the problem was to estimate $f(x)$ by a kernel estimator and Fourier transform (see, for example, Carroll and Hall, 1988; Liu and Taylor, 1989; Fan, 1991). Fan (1991) proved that the estimators of $f(x)$ are asymptotically optimal pointwise and globally if the Fourier transform of the kernel has bounded support. A further approach is based on wavelet (see, for example, Pensky and Vidakovic, 1999; Walter, 1999; Lee, 2001; Lee, 2002; Lee and Hong, 2002).

Recently the support vector method has drawn much attention on classification and regression problem. It was developed in Russia in the sixties by Vapnik and co-workers (Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1964). Weston *et al.* (1999) proposed the support vector method for probability density function estimation using support vector regression algorithm. In Lee and Taylor (2008) the support vector method was applied to estimate probability density function when the sample observations are contaminated with random noise. However, the simulation study in the paper doesn't show any practical merit when it is compared to classical kernel estimates. Hence, in this paper we mention some modifications to the procedure in Lee and Taylor (2008) and introduce another method of estimation of a deconvolution density using support vector regression method based on a reproducing kernel Hilbert space(RKHS).

† This research was supported by the Daegu University Research Grant, 2007.
1) Professor, Department of Statistics, Daegu University, Kyungsan, Kyungbuk 712-714, Korea.
E-mail: shlee1@daegu.ac.kr

2. Support Vector Method for the Deconvolution Problem

In this section, we will introduce two different methods of estimation of a deconvolution density using support vector regression method. First, we will briefly introduce the method of estimating a deconvolution density proposed by Lee and Taylor (2008).

Let

$$f(x,\,\omega)=\sum_{r=0}^{\infty}\omega_r\phi_r(x)=\omega\cdot\Phi(x),\quad \omega=(\omega_0,\ldots,\omega_m,\ldots),\ \Phi(x)=(\phi_0(x),\ldots,\phi_m(x),\ldots).$$

Then,

$$\begin{aligned}G(y)&=\int_{-\infty}^{\infty}F(y-u)q(u)du\\&=\omega\cdot\Theta(y),\quad \Theta(y)=(\theta_0(y),\,\theta_1(y),\ldots),\ \theta_r(y)=(q*\psi_r')(y),\quad \psi_r'=\phi_r\end{aligned}$$

and

$$\begin{aligned}g(y)&=\int_{-\infty}^{\infty}f(y-u)q(u)du\\&=\omega\cdot\Theta'(y),\quad \Theta'(y)=({\theta_0}'(y),\,{\theta_1}'(y),\ldots),\ {\theta_r}'(y)=(q*\phi_r)(y).\end{aligned}$$

In order to estimate $f(x)$ from a training set $\{(y_i,\,G_n(y_i))\,|\,y_i\in\mathbb{R},\,G_n(y_i)\in\mathbb{R},\,i=1,2,\ldots,n\}$, we try to minimize the empirical risk function $R_{emp}(G)$ with a complexity term $||\omega||^2$.

$$\text{minimize } R_{reg}(G)=R_{emp}(G)+\lambda||\omega||^2=\frac{1}{n}\sum_{i=1}^{n}c(G(y_i),\,G_n(y_i))+\lambda||\omega||^2 \tag{2.1}$$

with $c(G(y_i),\,G_n(y_i))$ being the cost function and λ being a regularization constant. For the ε-insensitive cost functions (see, Vapnik, 1995)

$$c(G(y),\,G_n(y))=\begin{cases}|G(y)-G_n(y)|-\epsilon, & \text{for } |G(y)-G_n(y)|\geq\epsilon,\\ 0, & \text{otherwise,}\end{cases}$$

the equation (2.1) can be minimized by solving quadratic programming problem formulated in terms of dot products in $\mathscr{F}$. Then

$$\omega=\sum_{i=1}^{n}\beta_i\,\Theta(y_i) \tag{2.2}$$

and hence

$$\begin{aligned}\hat{G}(y)&=\sum_{i=1}^{n}\beta_i(\Theta(y_i),\,\Theta(y))=\sum_{i=1}^{n}\beta_ik(y_i,\,y),\qquad k(y_i,\,y)=(\Theta(y_i),\,\Theta(y)),\\ \hat{g}(y)&=\sum_{i=1}^{n}\beta_i(\Theta(y_i),\,\Theta'(y))=\sum_{i=1}^{n}\beta_ik'(y_i,\,y),\quad k'(y_i,\,y)=(\Theta(y_i),\Theta'(y)),\\ \hat{f}(x)&=\sum_{i=1}^{n}\beta_i(\Theta(y_i),\,\Phi(x)),\end{aligned}$$

where $k(y_i, y)$ is a kernel function to compute a dot product in feature space. Now the Fourier transform of $\Theta'(y) = ({\theta_0}'(y), {\theta_1}'(y), \dots)$ can be defined as

$$\tilde{\Theta}'(\omega) = \left(\tilde{\theta_0}'(\omega), \tilde{\theta_1}'(\omega), \dots\right) = \left(\tilde{q}(\omega)\tilde{\phi_0}(\omega),\ \tilde{q}(\omega)\tilde{\phi_1}(\omega), \dots\right).$$

Let $\tilde{\Phi} = (\tilde{\phi_0}(\omega), \tilde{\phi_1}(\omega), \dots)$. Then $\tilde{\Phi}(\omega) = \tilde{\Theta}'(\omega)/\tilde{q}(\omega)$ and by applying the Fourier inversion formula, we can obtain a support vector density estimator for the deconvolution problem

$$\hat{f}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{\hat{g}}(\omega)}{\tilde{q}(\omega)}\, e^{i\omega x} d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \sum_{j=1}^{n} \beta_j \tilde{k}'(y_j, \omega)\frac{e^{i\omega x}}{\tilde{q}(\omega)} d\omega, \tag{2.3}$$

when $\tilde{\hat{f}}(\omega)$ is absolutely integrable.

Now we will introduce the other method of estimation of a deconvolution density using the support vector regression method based on a reproducing kernel Hilbert space (RKHS). A (real) RKHS H is a Hilbert space of real-valued functions f on an interval τ with the property that, for each $t \in \tau$, the evaluation functional L_t, $L_t : f \to f(t)$, is a bounded linear functional. Then, by Riesz representation theorem, for each $t \in \tau$ there exists a unique element $k_t \in H$ such that for each $f \in H$, $L_t(f) = f(t) = (k_t, f)$. The function defined by $k_u(v) = k(u, v) = (k_u, k_v)$ for $u, v \in \tau$ is the reproducing kernel. Then, by the Moore - Aronszajn theorem (Aronszajn, 1950), to every positive definite function k on $\tau \times \tau$ there corresponds a unique RKHS H_k of real valued functions on τ with k as its reproducing kernel. Note that any positive definite function $k(u, v)$ has an expansion $k(u, v) = \sum_{i=0}^{\infty} \lambda_i \phi_i(u)\phi_i(v)$. Let us consider the set of functions, $f(x, \omega) = \sum_{r=0}^{\infty} \omega_r \phi_r(x)$ and define the inner product as $(f(x, \omega), f(x, \omega^*)) = \sum_{i=0}^{\infty} \omega_i \omega_i^* / \lambda_i$. Then we have a RKHS H_k with its reproducing kernel k and will apply these properties of RKHS to the estimation of a deconvolution density. In order to estimate $f(x)$, first, $g(y)$ will be estimated using the reproducing kernel of RKHS (see, Mukherjee and Vapnik, 1999). Let

$$\hat{g}(y, \omega) = \sum_{i=1}^{n} \omega_i k(y_i, y)$$

and

$$\text{minimize } \Omega(\hat{g}, \hat{g}) = (\hat{g}, \hat{g})_H = \sum_{i,j=1}^{n} \omega_i \omega_j k(y_i, y_j)$$

$$\text{subject to } \max_i \mid G_n(y) - \int_{-\infty}^{y} \sum_{j=1}^{n} \omega_j k(y_j, y')dy'|_{y=y_i} = \epsilon. \tag{2.4}$$

This optimization problem is closely related to the support vector regression problem with an ϵ-insensitive loss function and hence the coefficients ω_i's can be found by solving the following quadratic programming problem:

$$(\alpha_0\ \alpha_0^*) = \arg\max_{\alpha, \alpha^*} -\frac{1}{2}(\alpha^* - \alpha)R'K^{-1}R(\alpha^* - \alpha) + y(\alpha_i^* - \alpha_i) - \epsilon(\alpha^* + \alpha)$$

$$0 \le \alpha_i^*, \quad \alpha_i \le C, \quad i = 1, \dots, n,$$

where $K = [k_{ij}]_{n\times n}$, $k_{ij} = k(y_i, y_j)$ and $R = [r_{ij}]_{n\times n}$, $r_{ij} = \int_{-\infty}^{y_i} k(y_j, y)dy$. Then $\hat{g}(y, \omega) = \sum_{i=1}^{n} \omega_i k(y_i, y)$, $\omega = K^{-1}R(\alpha_0^* - \alpha_0)$. Finally, applying the Fourier inversion formula, $f(x)$ will be estimated. That is,

$$\hat{f}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{\hat{g}}(\omega)}{\tilde{q}(\omega)} e^{i\omega x} d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \sum_{j=1}^{n} \omega_j \tilde{k}(y_j, \omega)\frac{e^{i\omega x}}{\tilde{q}(\omega)} d\omega, \tag{2.5}$$

when $\tilde{\hat{f}}(\omega)$ is absolutely integrable.

3. Simulation and Discussion

In this section we have tried to compare the performance of two methods introduced in Section 2 using Gaussian kernel and the following three different estimators of continuous distribution function, (3.1)–(3.3). However, the singular problem in K^{-1} was occurred when we used an estimator (2.5) with standard Gaussian kernel. Thus we fail to obtain simulation results with regard to an estimator (2.5) and hence the simulation study in this section is limited to the support vector method proposed by Lee and Taylor (2008) with three different estimators of continuous distribution function. The singular problem in the procedure of estimation is expected to be solved later.

The empirical distribution function was used in Lee and Taylor (2008)

$$G_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y) \tag{3.1}$$

as an estimator of the unknown distribution function $G(y)$, which didn't show good approximation for the Parzen's kernel estimator in the simulation study. For the deconvolution problem, we deal with continuous distribution function only. Thus it make sense to construct the empirical distribution function as a monotone increasing continuous function. In this section two alternative estimators are considered in addition to the empirical distribution function $G_n(y)$. Let $(y_{(1)}, y_{(2)}, \cdots, y_{(n)})$ be an ordered data set. As an alternative estimator, we can consider the following linear empirical distribution function (see, Mukherjee and Vapnik, 1999)

$$G_n^1(y) = \begin{cases} \dfrac{k}{n} + \dfrac{1}{n}\dfrac{y - y_{(k)} - \tau_k/2}{\tau_k}, & \text{if } y \in \left[y_{(k)} - \dfrac{\tau_k}{2}, y_{(k)} + \dfrac{\tau_k}{2}\right), \\ \dfrac{k}{n}, & \text{if } y \in \left[y_{(k)}, y_{(k+1)}\right) \text{ and } y \notin \left[y_{(k)} - \dfrac{\tau_k}{2}, y_{(k)} + \dfrac{\tau_k}{2}\right), \end{cases} \tag{3.2}$$

where τ_k is the distance between the k^{th} data point and it's nearest neighbor. As another continuous distribution function estimator, using $G_n(y)$, we can consider the following kernel regression estimator proposed by Nadaraya-Watson (Nadaraya, 1964; Watson, 1964)

$$G^2{}_n(y) = \frac{\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) G_n(Y_i)}{\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right)}. \tag{3.3}$$

Now, using three different estimators of distribution function (3.1)–(3.3), we compare the support vector method with Parzen's methods. A numerical study of the deconvolution density is constructed when $q(z) = 1/(2\gamma)e^{-|z|/\gamma}$, $\gamma = 0.1$ and $f(x)$ is the standard normal probability distribution as shown in Section 5, Pensky and Vidakovic (1999). The Gaussian Parzen estimator used in this section $\hat{f}(x)$ (see, Liu and Taylor, 1989) is

$$\hat{f}(x) = \frac{1}{2\pi n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} e^{i\omega(x-y_j)}\frac{\tilde{K}(h\omega)}{\tilde{q}(\omega)}\, d\omega,$$

where $\tilde{K}(\omega) = e^{-\omega^2/2}$ is the Fourier transform of the standard normal probability distribution. For given $q(z) = 1/2\gamma^{-1}e^{-|z|/\gamma}$, $\gamma = 0.1$, the Gaussian Parzen estimator $\hat{f}(x)$ (see, Pensky and Vidakovic, 1999) is

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}hn}\sum_{j=1}^{n} e^{-0.5\left(\frac{x-y_j}{h}\right)^2}\left[1 - \frac{\gamma^2}{h^2}\left\{\left(\frac{x-y_j}{h}\right)^2 - 1\right\}\right]$$

and the support vector density estimator $\hat{f}(x)$ (see, Lee and Taylor, 2008) is

$$\begin{aligned}\hat{f}(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\sum_{j=1}^{n}\beta_j\tilde{k}'(y_j, \omega)\frac{e^{i\omega x}}{\tilde{q}(\omega)}\, d\omega \\ &= -\left(\frac{1}{\sigma^2} + 3\frac{\gamma^2}{\sigma^4}\right)\sum_{j=1}^{n}\beta_j(x-y_j)e^{-\frac{(x-y_j)^2}{2\sigma^2}} + \frac{\gamma^2}{\sigma^6}\sum_{j=1}^{n}\beta_j(x-y_j)^3 e^{-\frac{(x-y_j)^2}{2\sigma^2}},\end{aligned}$$

where $k(x, y) = e^{-(x-y)^2/2\sigma^2}$ and $\tilde{q}(\omega) = (1+\gamma^2\omega^2)^{-1}$.

Figure 3.1 and 3.2 show plots of the Parzen and support vector deconvolution estimates for the Gaussian kernel when 20 and 100 points are randomly generated respectively from the standard normal probability distribution $f(x)$ and double exponential probability distribution $q(z)$. The exact probability density function $f(x)$ is shown in bold line and the support vector deconvolution estimate is shown in dashed lines. For the support vector deconvolution estimates Gunn's program and MatLab 6.5 was used (see, Gunn, 1998). Each estimate was picked with the best possible value of parameters based on the exact probability density function $f(x)$. Figure 3.1 presents the simulation study when a random sample of size 20 is generated from a target distribution, standard normal probability distribution $f(x)$ and a noise distribution, double exponential probability distributions $q(z)$. The parameters, $h = 0.7$, $\sigma = 1.2$, $C = \infty$, are chosen for the best possible estimates with ϵ-insensitive loss function($\epsilon = 0.05$). The support vector deconvolution estimate uses five points in the approximation. Figure 3.2 presents the simulation study when a random sample of size 100 is generated and parameters, $h = 0.6$, $\sigma = 1.1$, $C = \infty$, are chosen for the best possible estimates with ϵ-insensitive loss function($\epsilon = 0.05$). The support vector deconvolution estimate uses five points in the approximation, *i.e.*, the number of support vectors is five.

Figure 3.1 and 3.2 show that the performance of the Parzen deconvoluion estimates is better than the support vector deconvolution estimates. In particular, support vector deconvolution estimates do not yield good estimates on the right tail part.

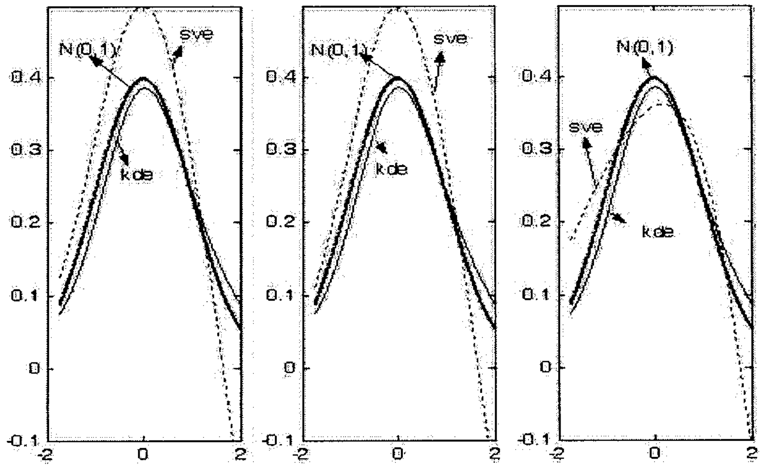
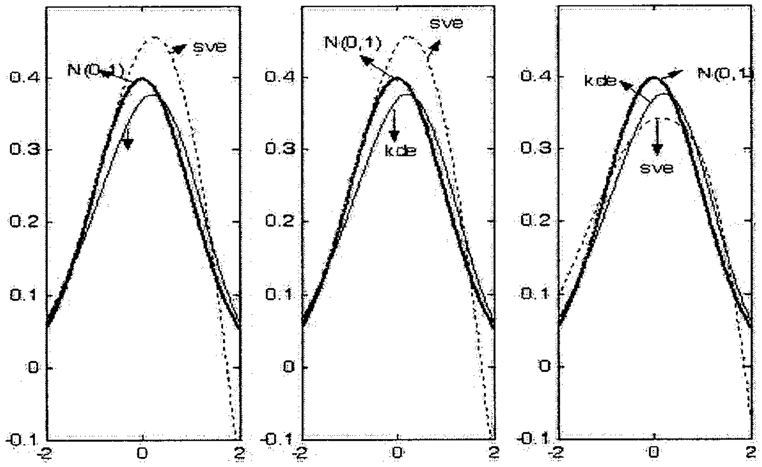


Figure 3.1: The simulation study when a random sample is of size $n = 20$ ((Left) Estimation with $G_n(y)$, (Middle) Estimation with $G_n^1(y)$, (Right) Estimation with $G_n^2(y)$)



abc

Figure 3.2: The simulation study when a random sample is of size $n = 100$ ((Left) Estimation with $G_n(y)$, (Middle) Estimation with $G_n^1(y)$, (Right) Estimation with $G_n^2(y)$)

4. Concluding Remarks

In this paper two different methods of estimation of a deconvolution density using support vector regression algorithm were introduced. A simulation study was limited to the first method because the singularity problem was occurred in the procedure of the other proposed method. The performance of support vector deconvolution estimates conducted doesn't yield good estimates than Parzen deconvolution estimates. A new support vector method is needed to minimize drawbacks of the support vector deconvolution estimates as the figures indicate. Thus another method proposed in this paper is expected to be a solution to this problem when the singular problem is solved in the future.

References

Aronszajn, N. (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 337–404.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density, *Journal of the American Statistical Association*, **83**, 1184–1886.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems, *The Annals of Statistics*, **19**, 1257–1272.

Gunn, S. R. (1998). Support vector machines for classification and regression. *Technical report*. University of Southampton.

Lee, S. (2001). A note on strongly consistent wavelet density estimator for the deconvolution problem, *The Korean Communications in Statistics*, **8**, 859–866.

Lee, S. (2002). A note on Central limit theorem for deconvolution wavelet density estimator, *The Korean Communications in Statistics*, **9**, 241–248.

Lee, S. and Hong, D. H. (2002). On a strongly consistent wavelet density estimator for the deconvolution problem, *Communications in Statistics - Theory and Methods*, **31**, 1259–1272.

Lee, S. and Taylor, R. L. (2008). A note on support vector density estimation for the deconvolution problem, *Communications in Statistics - Theory and Methods*, **37**, 328–336.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem, *The Canadian Journal of Statistics*, **17**, 427–438.

Louis, T. A. (1991). Using empirical Bayes methods in biopharmaceutical research, *Statistics in Medicine*, **10**, 811-827.

Mukherjee, S. and Vapnik, V. (1999). Support vector method for multivariate density estimation. *Technical Report*. A.I. Memo no. 1653, MIT AI Lab.

Nadaraya, E. (1964). On regression estimators, *Theory of Probability and It's Application*, **9**, 157–159.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolutoin, *The Annals of Statistics*, **27**, 2033–2053.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer Verlag, New York.

Vapnik, V. and Chervonenkis, A. (1964). A note on one class of perceptrons, *Automation and Remote Control*, **25**.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method, *Automation and Remote Control*, **24**.

Walter, G. G. (1999). Density estimation in the presence of noise, *Statistics & Probability Letters*, **41**, 237–246.

Watson, G. S. (1964). Smooth regression analysis, *Sankhyā: The Indian Journal of Statistics*, Ser. A, **26**, 359–372.

Weston, J., Gammerman, A., Stitson, M., Vapnik, V., Vovk, V. and Watkins, C. (1999). Support vector density estimation, In Scholkopf, B. and Smola, A., editors, *Advances in Kernel Methods-Suppot Vector Learning*, 293–306, MIT Press, Cambridge, MA.

Zhang, H. P. (1992). On deconvolution using time of flight information in positron emission tomography, *Statistica Sinica,* **2**, 553–575.

[Received May 2008, Accepted September 2008]