Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 411~418

A Note on Support Vector Density Estimation with Wavelets[1)]

Sungho Lee[2)]

Abstract

We review support vector and wavelet density estimation. The relationship between support vector and wavelet density estimation in reproducing kernel Hilbert space (RKHS) is investigated in order to use wavelets as a variety of support vector kernels in support vector density estimation.

Keywords : Reproducing kernel Hilbert space, RKHS, Support vector, Wavelets

1. Introduction and Preliminaries

The support vector method is a tool for solving multidimensional function estimation problems. It was developed in Russia in the sixties by Vapnik and co-workers(Vapnik and Lerner(1963), Vapnik and Chervonenkis(1964)). It was initially designed to solve pattern recognition problems, where one selects some (small) subset of the training data, called the support vectors, to find a decision rule with good generalization ability. Later the support vector method was extended to regression and real-valued function estimation. The support vector method is a very powerful method in a wide variety of applications and gives a new opportunity for solving probability density function estimation problem.

Let us review the support vector regression algorithm for nonlinear function estimation(see, for example, Vapnik(1995), Vapnik et al(1997), Smola and Schölkopf(1998)). The algorithm can be directly applied to support vector methods for probability density function estimation. The support vector regression algorithm

---

1) This research was supported by the Daegu University Research Grant, 2004

2) Professor, Department of Information Statistics, Daegu University, Kyungsan, 712-714, Korea.
E-mail : shlee1@daegu.ac.kr

computes a nonlinear function in the space of the input data $\mathbb{R}^m$ by using a linear function in high dimensional feature space $\mathcal{F}$ with a dot product. The functions take the form $f(x) = \omega \cdot \Phi(x) + b$ with $\Phi: \mathbb{R}^m \to \mathcal{F}$ and $\omega \in \mathcal{F}$. In order to estimate $f(x)$ from a training set

$\{ (x_i, y_i) \mid i = 1, \cdots, n,\ x_i \in \mathbb{R}^m, y_i \in \mathbb{R} \}$, one tries to minimize the empirical risk function $R_{emp}(f)$ together with a complexity term $\|\omega\|^2$, i.e. to minimize

$$R_{reg}(f) = R_{emp}(f) + \lambda \|\omega\|^2 = \frac{1}{n} \sum_{i=1}^{n} c(f(x_i), y_i) + \lambda \|\omega\|^2 \tag{1.1}$$

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

$$c(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

equation (1.1) can be minimized by solving quadratic programming problem formulated in terms of dot products in $\mathcal{F}$. It turns out that the solution can be expressed in terms of support vectors,

$$\omega = \sum_{i=1}^{n} \alpha_i \Phi(x_i), \tag{1.3}$$

and hence

$$f(x) = \sum_{i=1}^{n} \alpha_i (\Phi(x_i) \cdot \Phi(x)) + b = \sum_{i=1}^{n} \alpha_i k(x_i, x) + b$$

where $k(x_i, x)$ is a kernel function to compute a dot product in feature space(see Vapnik(1995)). The coefficients α_i's can be found by solving a quadratic programming problem (with $K_{ij} = k(x_i, x_j)$ and $\alpha_i = \beta_i^* - \beta_i$):

maximize $-\frac{1}{2} \sum_{i,j=1}^{n} (\beta_i^* - \beta_i)(\beta_j^* - \beta_j) K_{ij} + \sum_{i=1}^{n} (\beta_i^* - \beta_i) y_i - \sum_{i=1}^{n} (\beta_i^* + \beta_i) \varepsilon$

subject to $\sum_{i=1}^{n} (\beta_i - \beta_i^*) = 0,\ \beta_i,\ \beta_i^* \in [0, \frac{1}{\lambda n}]$.

Note that (1.2) is not the only possible choice of cost functions resulting in a quadratic programming problem (see Schölkopf and Smola(2002)). The remained question is which functions $k(x, y)$ correspond to a dot product in some feature

space $\mathcal{F}$. Mercer theorem (1909) indicates that any continuous symmetric function $k(x, y)$ may be used as an admissible support vector kernel(Mercer kernel) if it satisfies Mercer's condition

$$\int \int k(x, y)\, g(x)\, g(y)\, dx\, dy \geq 0 \quad \text{for all} \quad g \in L_2(\mathbb{R}^m).$$

2. Support vector density estimation

Weston et al(1999) proposed support vector method for probability density function estimation. The method is briefly introduced for later use. Consider the following linear operation equations

$$Ap(t) = \int_{-\infty}^{x} p(t)\, dt = F(x), \tag{2.1}$$

where operator A is a linear mapping from a Hilbert space function $p(t)$ to a Hilbert space of function $F(x)$. They used a regression problem in the image space $(F(x, \omega))$to estimate $p(t)$. Choose a set of density functions $p(t, \omega)$ to solve the problem in the pre-image space that are linear in the flattening space as follows:

$$p(t, \omega) = \sum_{r=0}^{\infty} \omega_r\, \phi_r(t) = (\omega \cdot \Phi(t)), \text{ where } \Phi(t) = (\phi_0(t), \cdots, \phi_m(t), \cdots).$$

Each $p(t, \omega)$ can be thought of as a hyperplane in this flattening space, where $\omega = (\omega_0, \cdots, \omega_m, \cdots)$ are the coefficients to the hyperplane. Then $F(x, \omega)$ can be expressed as a linear combination of functions in the image Hilbert space as follows:

$$F(x, \omega) = Ap(t, \omega) = \sum_{r=0}^{\infty} \omega_r \psi_r(x) = \omega \cdot \Psi(x),$$

where $\Psi(x) = (\psi_0(x), \cdots, \psi_m(x), \cdots)$ and $\psi_r(x) = A\phi_r(x)$. Thus the probability density function estimation is equivalent to estimating coefficients vector ω in the image space. Let $\{ (x_i, y_i) \mid i = 1, \cdots, n,\ x_i \in \mathbb{R},\ y_i \in \mathbb{R} \}$ be a training set with $y_i = F_n(x_i)$(=empirical distribution function) and

$\varepsilon_i = \lambda\,\hat{\sigma}_i = \lambda\sqrt{\frac{1}{n}F_n(x_i)(1-F_n(x_i))}$ where λ is usually chosen to be 1. Finally they used the support vector regression method in section 1 as follows :

maximize $-\frac{1}{2}\sum_{i,j=1}^{n}(\beta_i^*-\beta_i)(\beta_j^*-\beta_j)\,k(x_i,x_j)+\sum_{i=1}^{n}y_i(\beta_i^*-\beta_i)-\sum_{i=1}^{n}\varepsilon_i(\beta_i^*+\beta_i)$

subject to the constraints

$$\sum_{i=1}^{n}(\beta_i^*-\beta_i)=0,\quad 0\le\beta_i^*,\beta_i\le C,\qquad i=1,\cdots,n;\ \alpha_i=\beta_i^*-\beta_i.$$

These coefficients define the estimator to the density

$$\hat{p}(t)=\sum_{i=1}^{n^0}\alpha_i^0\,(\Psi(x_i^0)\cdot\Phi(t))$$

where, by equation (1.3), $\omega=\sum_{i=1}^{n}\alpha_i\Psi(x_i)$ and x_i^0 are the $n^0\le n$ support vectors with corresponding non-zero coefficients α_i^0.

3. Support vector density estimation by wavelets.

Wavelet density estimation methods have been introduced by Doukan and Leon(1990), Kerkyacharian and Picard(1992), and Donoho et al(1996). These authors have demonstrated the virtues of wavelet methods in the context of the achievability of very good convergence rates uniformly over exceptionally large function space. There are several important families of wavelets(for example, Haar's wavelets, Meyer's wavelets, Franklin's wavelets, Daubechies' compactly supported wavelets). In this section our main interests are restricted to projection kernels derived from an $L^2(R)$ multiresolution. Such kernels are reproducing kernels(see Lemma 3.1). Reproducing kernel Hilbert space(RKHS) with reproducing kernel can be used in curve fitting, function estimation, and density estimation as useful objects. It is briefly introduced for our purpose(see Walter(1994), Wahba(1990)). 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, which associates f with $f(t)$, $L_t: f\to f(t)$, 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 $f\in H$, $L_t(f)=f(t)=\langle f, k_t\rangle$. The function defined by $k(u,v)=\langle k_u, k_v\rangle$ for $u,v\in\tau$ is the reproducing kernel.

Let us review wavelet density estimation(see, Walter(1994), Kerkyacharian and Picard(1990)). We can construct a function φ(called a father wavelet) such that :

(1) The sequence $\{\varphi(x-k), k\in Z\}$ is an orthonormal family of $L^2(R)$ and $\int \varphi = 1$. Let us call V_0 the subspace spanned by this sequence.

(2) $\forall j \in Z$, $V_j \subset V_{j+1}$, if V_j denotes the subspaces spanned by the sequence $\{\varphi_{j,k},\ k\in Z\}$, $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$.

Then we have $\bigcap_{j\in z} V_j = \{0\}$ and $L^2(R) = \overline{\bigcup_{j\in z} V_j}$(= closure of $\bigcup_{j\in z} V_j$).

It is possible to require in addition that one of the following conditions holds :

(a) φ is r times differentiable and its derivatives are continuous, $\varphi \in C^r$, and φ and all its derivatives up to the order r are rapidly decreasing.

(b) φ is of class C^r compactly supported(Daubechies's wavelet, Daubechies(1992)).

Let us define the space W_j by $V_{j+1} = V_j \oplus W_j$ in this conditions, where W_j is the orthogonal complement of V_j in V_{j+1} and $\oplus$ represents the orthogonal sum of two subspaces. Then there exists a function ψ(called a mother wavelet) such that :

(1) $\{\psi(x-k),\ k\in Z\}$ is an orthonormal basis of W_0,

(2) The family $\{\psi_{j,k},\ k, j\in Z\}$ is an orthonormal basis of $L^2(R)$ if $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$.

Then $L^2(R) = V_{j0} \oplus W_{j0} \oplus W_{j0+1} \oplus \cdots$, and

$$\forall f \in L^2(R),\ f = \sum_{k\in Z} \alpha_{j_0,k}\varphi_{j_0,k} + \sum_{j\geq j_0}\sum_{k\in Z} \beta_{j,k}\psi_{j,k}$$

where $\alpha_{j_0,k} = \int f(x)\,\varphi_{j_0,k}(x)\,dx$, $\beta_{j,k} = \int f(x)\,\psi_{j,k}(x)\,dx$, $j \geq j_0$.

Let us introduce the following projectors and their associated kernels :

$$\begin{aligned} f\in L^2(R) \rightarrow P_j f (&= \text{ projection of } f \text{ onto } V_j\ ) \\ &= \sum_k \langle f, \varphi_{j,k}\rangle\, \varphi_{j,k}(x) \\ &= \sum_k \left(\int f(y)\, 2^{j/2}\, \varphi(2^j y - k)\, dy\right) \times \varphi_{j,k}(x) \\ &= \int \{2^j \sum_k \varphi(2^j x - k)\, \varphi(2^j y - k)\,\} f(y)\, dy \\ &= \int K_j(x, y) f(y)\, dy, \quad \text{where } K_j(x, y) = 2^j \sum_k \varphi(2^j x - k)\, \varphi(2^j y - k). \end{aligned}$$

Then from the above facts we can obtain the following important lemma for RKHS.

Lemma 3.1. V_j and W_j is reproducing kernel Hilbert spaces with reproducing

kernels $K_j(x, y)$ and $2^j \sum_k \psi(2^j x - k)\, \psi(2^j y - k)$ respectively.

proof. Let $f \in V_j$. Then,

$$\langle f(x),\ 2^j \sum_k \psi(2^j x - k)\, \psi(2^j t - k) \rangle$$

$$= \langle 2^{j/2} \sum_k c_k \psi(2^j x - k),\ 2^j \sum_k \psi(2^j x - k)\, \psi(2^j t - k) \rangle$$

where $c_k = \langle f(x),\ 2^{j/2} \sum_k \psi(2^j x - k) \rangle$

$$= \sum_k \langle c_k 2^{j/2} \psi(2^j x - k),\ 2^j \psi(2^j x - k)\, \psi(2^j t - k) \rangle$$

$$= \sum_k c_k 2^{j/2} \psi(2^j t - k)$$

$$= f(t).$$

Similarly it can be proved for W_j.

Notice that reproducing kernel $K_j(x, y)$ can be used as a kernel function $k(x_i, x_j)$ in Section 1 and 2 in order to compute a dot product in feature space. Hence probability density function estimation in support vector method can be considered as an optimization problem in a RKHS. The following theorem is obtained as a special case of the representer theorem in Kimeldorf and Wahba(1971).

Theorem 3.2. Let $f \in V_j$ and let $\{ (x_i, y_i) \mid i = 1, \cdots, n,\ x_i \in \mathbb{R},\ y_i \in \mathbb{R} \}$ be a training set. Then any solution to the problem : find f to minimize

$$R_{reg}(f) = R_{emp}(f) + \lambda \|\omega\|^2 = \frac{1}{n} \sum_{i=1}^{n} c(f(x_i), y_i) + \lambda \|\omega\|^2$$

has a representation of the form

$$f(\cdot) = \sum_{i=1}^{n} d_i K_j(x_i, \cdot), \quad K_j(x_i, \cdot) = 2^j \sum_k \varphi(2^j x_i - k)\, \varphi(2^j \cdot - k),$$

where

$$f(t) = \sum_k \omega_k \varphi_{j,k} \text{ and } c(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

and d_i can be found by applying the SV method in Section 1.

proof. By Applying Lemma 3.1 and support vector regression method in Section 1 to the representer theorem in Kimeldorf and Wahba(1971), the theorem is proved.

Let $x_1, x_2, \cdots, x_n$ be a random sample from probability density function $p(t)$

and distribution function $F(t)=\int_{-\infty}^{t} p(x)\,dx$ in the subspace V_j. Let $y_i = F_n(x_i)$(=empirical distribution function) and $\varepsilon_i = \lambda\,\hat{\sigma}_i = \lambda\sqrt{\frac{1}{n}F_n(x_i)(1-F_n(x_i))}$. Then the above theorem indicates that $\hat{F}(t)$ has a representation of the form $\hat{F}(t)=\sum_{i=1}^{n} d_i K_j(x_i,t)$ and $\hat{p}(t)$ has a representation of the form $\hat{p}(t)=\sum_{i=1}^{n} d_i \frac{d}{dt}K_j(x_i,t)$. Since $V_j \subset L^2(R)$ and $L^2(R)=V_j \oplus W_j \oplus W_{j+1} \oplus \cdots$, $\hat{p}(t)$ can be a good estimator of probability density function with an appropriate selection of subspace V_j. In the view of the mean integrated squared error of an estimator, $\hat{p}(t)=\frac{1}{n}\sum_{i=1}^{n} K_j(x_i,t)$ is a good estimator in some function spaces(see, for example, Kerkyacharian and Picard(1992), Donoho et al(1996)).

4. Concluding remarks

In support vector methods an important choice is support kernels. Weston et al(1999) considered the set of constant splines with infinite number of nodes in density estimation. Vapnik et al(1997) considered the set of splines of order d with an infinite number of nodes in function estimation. As Theorem 3.2 indicates, a variety of wavelets can be used as support kernels in support vector density estimation. Wavelets have many practical and theoretical advantages over the classical systems. For large sample it is known to outperform classical density estimators in representing discontinuities and local oscillations. It gives better localization properties as well as better convergence properties. Thus we can expect the same results in support vector density estimation. There is much work on this topic. Simulation studies for a variety of wavelet-based kernels and probability density functions is left for further work.

References

1. Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D.(1996). Density stimation by wavelet thresholding. *Ann. Statist.* 24, 508 - 539.
2. Daubechies, I.(1992). *Ten Lectures on Wavelet*s. CBMS-MSF, SIAM, Philadelphia.
3. Doukhan P. and Leon J.(1990). Deviation quadratique d'estimateurs d'une densite par projection orthogonale. *C.R. Acad . Sci. Paris Ser. I*

Math 310. 425 - 430.
4. Kerkyacharian, G. and Picard, D.(1992). Density estimation in Besov spaces. *Statist. Probab. Lett*. 13, 15 - 24.
5. Kimeldorf, G. and Wahba, G.(1971). Some results on Tchebycheffian spline functions. *J. Math. and Anal. Applic.*, 33,82-95.
6. Mercer, J.(1909). Functions of positive and negative type and their connection with the theory of integral equation. *Philos. Trans. Roy. Soc. London, A* 209, 415-446.
7. Scholkopf, B. and Smola, A.(2002). *Learning with Kernels*. MIT Press, Cambridge, MA.
8. Smola, A. and Scholkopf, B. (1998). From regularization operators to support vector kernels. In *Advances in Neural Information Processing Systems*, 10, 343-349, San Mateo, CA.
9. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.
10. Vapnik, V. and Chervonenkis, A.(1964). A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.
11. Vapnik, V., Golowich, S. and Smola, A.(1997). Support vector method for function approximation, regression estimation, and signal processing. In Mozer, M., Jordan, M. and Petsche, editors, *Advances in Neural Information Processing Systems* 9, 281-287, MIT Press.
12. Vapnik, V. and Lerner. L. (1963). Pattern Recognition using generalized portrait method. *Automation and Remote Control*, 24,1963.
13. Walter, G.(1994). *Wavelets and Other Orthogonal Systems with Applications*. CRC Press, Inc.
14. Wahba, G.(1990). *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v.59.
15. 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.

[received date : Dec. 2004, accepted date : May. 2005]