Bandwidth Selection for Kernel Density Estimation Using Total Variation with Fourier Domain Constraints

Alexander Suhre, Orhan Arikan, *Member, IEEE,* and A. Enis Cetin, *Fellow, IEEE*

Abstract

Kernel density estimation (KDE) is a popular approach for non-parametric estimation of an underlying density from data. The performance of KDE is mainly dependent on the bandwidth choice of the very kernel. This article presents various methods of estimating the bandwidth using sparsity in the Fourier transform domain. It uses the Total Variation (TV) and Filtered Variation (FV) cost functions to estimate the bandwidth. Simulation results indicate that, over a set of distributions of interest, the presented approaches are able to outperform classical approaches.

Index Terms

KDE, Bandwidth, Fourier Domain, Total Variation.

I. Introduction

Estimating an underlying distribution from data is a widely studied problem [1]. Probability density function (PDF) estimation approaches can broadly be divided into two classes: parametric estimation and non-parametric estimation. This study deals with non-parametric estimation. An important branch of non-parametric estimation is the kernel-based approach, which is frequently referenced as kernel density estimation (KDE) [2]. In this approach, an estimate for the underlying density $g_X(x)$ is given by

$$\hat{g}_X(x;\sigma) = \frac{1}{N}\sum_{i=0}^{N} k_\sigma(x - v_i), \tag{1}$$

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey e-mail: (suhre@ee.bilkent.edu.tr).

where N is the number of data points, v_i denotes the observed data and σ is called the bandwidth of the kernel k_σ, in case of the Gaussian function, the standard deviation.

The performance of KDE depends largely on the choice of bandwidth of the kernel, which, if not chosen appropriately, can result in an over-smoothed estimate in case of a large bandwidth with respect to the underlying density or an under-smoothed estimate in case of a small bandwidth. An approximate bandwidth should decrease the mean integrated square error (MISE) [3]. In [4] several techniques for bandwidth estimation have been evaluated and it has been concluded that the most efficient method is the "plug-into-equation" approach of Sheather et al. [5], [6]. All of these methods are proposed to find a bandwidth σ that minimizes the MISE [3]. For better mathematical tractability, minimization of the approximate MISE (AMISE) is carried out. The optimal σ is dependent on the second derivative of $g_X(x)$. For this purpose, first a kernel family is chosen, after which its second derivative is estimated from the data.

For $g_X(x)$ which has a sparse representation in the Fourier domain, alternative techniques are proposed here. In the proposed approaches, the Total Variation (TV) [7] and Filtered Variation (FV) [8] cost functions are used to estimate the bandwidth. Simulation results indicate that the presented approaches are able to outperform classical approaches under certain conditions.

The computational cost of implementing the estimator in Eq. 1, is $\mathcal{O}(N^2)$ multiplications, which for large datasets may be prohibitive. A significant number of studies have been devoted to KDE, especially with the goal of reducing its computational burden [9], [6] and [10]. In [10], Eq. (1) is implemented in the Fourier domain via multiplication. The order of solving Eq. (1) in the Fourier domain is then $\mathcal{O}(Nlog(N))$. Another feature of the proposed methods is that they take advantage of the sparsity of the data in the Fourier domain.

II. Bandwidth Estimation Methods

In the following, the problem is stated in the Fourier domain: Let $g_{\mathbf{x}}(x)$ denote the original distribution from which N samples are drawn independently. Eq. (1) can be written as a convolution as follows:

$$\hat{g}_{\mathbf{x}}(x;\sigma) = k_\sigma(x) * \frac{1}{N}\sum_{i=0}^{N-1}\delta(x - v_i), \tag{2}$$

It is straight forward to see that the Fourier transform of Eq. (2) is

$$\hat{G}_{\mathbf{x}}(\omega;\sigma) = K_\sigma(\omega)\cdot\frac{1}{N}\sum_{i=0}^{N-1}e^{-j\omega v_i} = K_\sigma(\omega)\cdot\hat{H}(\omega), \tag{3}$$

where K_σ and $\hat{H}(\omega)$ are the Fourier transforms of the kernel k_σ and the data, respectively. Implementation of Eq. (3) is carried out using the Discrete Fourier Transform. A discrete estimate $\hat{H}[k]$ of $\hat{H}(\omega)$ can easily be obtained by using uniform binning of the data in the interval $[-L, L]$ into N intervals and computing the Fast Fourier transform (FFT) of the binned data, which is the histogram $\hat{h}[i]$. If the kernel $k_\sigma(x)$ is chosen to be a Gaussian function with variance σ, its Fourier transform $K_\sigma(\omega)$ is again a Gaussian and can be written as

$$K_\sigma(\omega) = e^{\frac{-\sigma^2\omega^2}{2}}. \tag{4}$$

which is also discretized in the DFT based implementation. In the proposed approaches, σ will be estimated as the minimizer of the TV or FV of $\hat{g}_{\mathbf{x}}(x;\sigma)$.

A. Total Variation (TV) based KDE

TV is successfully used in many sparse signal processing applications including compressice sensing and deconvolution [7], [11]. In this article, the following minimization problem is solved to estimate σ:

$$\begin{aligned} \min_{\hat{\sigma}} \quad & ||\hat{g}_{\mathbf{x}}(x;\hat{\sigma})||_{TV} \\ \text{subject to} \quad & ||\hat{g}_{\mathbf{x}}(x;\hat{\sigma}) - \hat{h}(x)||_1 \leq \varepsilon. \end{aligned} \tag{5}$$

where ε is the error tolerance and $||f||_{TV}$ of a given function f in the discrete domain is defined as follows:

$$||f||_{TV} = \sum_{i=0}^{N-2} |f[i] - f[i+1]| \tag{6}$$

The minimization problem (5) can also be expressed using the Lagrangian:

$$\min_{\hat{\sigma}} ||\hat{g}_X(x;\hat{\sigma}) - \hat{h}(x)||_1 + \lambda||\hat{g}_X(x;\hat{\sigma})||_{TV} \tag{7}$$

where λ is the Lagrange multiplier. There exists an ε corresponding to each λ such that both optimization problems result in the same solution [12], [13]. Note that a small ϵ in (6) results in a PDF estimate which is very close to the binned histogram $\hat{h}[i]$. The choice of λ or ϵ, respectively, determines the smoothness of the PDF estimate. For example, a large λ gives more emphasis to the second term in (7) and results in very smooth PDF estimates. This behavior will be more closely investigated in Section III, where different λ values are used.

The cost function given in (5) - (7) is minimized in the Fourier domain in an iterative manner by using projection onto convex sets techniques [14]. The iterative algorithm starts with a $\hat{\sigma_0}$ value in Fourier

domain and the Inverse FT (IFT) of $K_{\hat{\sigma_0}}(\omega) \cdot \hat{H}(\omega)$ is computed. The cost value of (5) is computed for $\hat{\sigma_0}$. Then $\hat{\sigma_0}$ is increased (or decreased) to get $\hat{\sigma_1}$ and the IFT of $K_{\hat{\sigma_1}}(\omega) \cdot \hat{H}(\omega)$ is computed again. This process is repeated in a greedy manner until a satisfactory $\hat{\sigma}$ value producing the lowest cost is determined in the search space. The Fourier domain implementation is faster than convolution based implementations, because each iteration is of order $\mathcal{O}(Nlog(N))$.

The set $C_1 = \{g_X(x;\sigma)|\ ||g_X(x;\sigma)||_{TV} \leq \epsilon_1\}$, i.e., the set of functions whose TV cost is less than or equal to a threshod is a convex set [11]. Similarily, the set $C_2 = \{g_X(x;\sigma)|\ ||\hat{g}_X(x;\hat{\sigma}) - \hat{h}(x)||_1 \leq \epsilon_2\}$ is also a convex set. So the iterative algorithm always converges when $C_1 \cap C_2$ is non-empty [15]. If the sets do not intersect, then either ϵ_1 or ϵ_2 or both are slowly increased so that they have a non-empty intersection. Since the optimization problems in (5) and (7) are equivalent, the iterative procedure described in the previous paragraph leads to computationally efficient solutions. It is also possible to solve (5) and (7) using brute force search methods because the PDFs are single-variate functions in this paper. For higher-dimensional problems, the iterative algorithm can be used.

B. Filtered Variation (FV) based KDE

Recently, the FV concept was introduced to denoising [8] by generalising the TV cost function. (6) can be rewritten as

$$||f||_{FV} = \sum_{i=0}^{N-1} |(f * w)[i]|, \tag{8}$$

where $w = \{-1,\ \ 1\}$ is a discrete highpass filter and $*$ denotes the discrete convolution operator. The filter $w = \{-1,\ \ 1\}$ is a simple Haar-type half band high-pass filter with ideal cut-off frequency $\pi/2$ [16]. However, Haar wavelets or Haar filters are poor high-pass filters [17]. It is possible to replace the high pass filter in Eq. (8) with longer filters. When different filters are used for w the framework is called the filtered variation (FV) [8]. For completeness, the considered minimization is given as follows

$$\min_{\hat{\sigma}} ||\hat{g}_{\mathbf{x}}(x;\hat{\sigma}) - \hat{h}(x)||_1 + \lambda||\hat{g}_{\mathbf{x}}(x;\hat{\sigma})||_{FV}, \tag{9}$$

The cut-off frequency of the high-pass filter is related with the smoothness of the estimated PDF. If a smooth PDF is desired, the high-pass filter should have a low cut-off frequency. If it is known that the PDF is multi-modal, a high cut-off frequency should be picked. In this article, the following longer Lagrange filter is chosen instead of the Haar filter: $w = \{\frac{1}{32},\ 0,\ -\frac{9}{32},\ \frac{1}{2},\ -\frac{9}{32},\ 0,\ \frac{1}{32}\}$, which is a also a halfband filter as the Haar filter but it has a more desirable frequency response [16]. This filter is chosen because we have both multi-modal PDFs and smooth PDFs in our data set in Section III. This filter was

used by Daubechies and for wavelet construction [16]. If there is no prior knowledge about the estimated PDF, the cut-off frequency can be chosen as $\pi/2$.

The cost function (9) is minimized as the TV cost function (7) in an iterative manner. After finding a suitable $\hat{\sigma}$ by one of the methods described above, KDE is carried out by computing Eq. (3). The resultant estimate of the distribution is the inverse Fourier transform of this product.

III. Simulation Results

The performance of the proposed methods are evaluated by using 15 test distributions from [3]. These are mixtures of Gaussians of several flavors. Some of the original PDFs are smooth and unimodal, some of them are multimodal and have sharp peaks as shown in Fig. 1. N random variates were independently drawn from each of the distributions and were used to compute an N-bin histogram. This histogram was used as the initial estimate of the desired distribution and was used as input to the methods from Section II and Sheather's method. The performances of these methods were measured against the original test distributions under two different error criteria: Kullback-Leibler (KL) divergence and Mean Integral Square Error (MISE).

Distribution Number	Sheather	TV	FV
1	4.86E-03	1.01E-02	1.40E-2
2	5.91E-03	1.07E-02	1.26E-2
3	2.52E-02	2.29E-02	2.27E-02
4	1.95E-02	1.87E-02	1.89E-02
5	2.65E-02	2.46E-02	2.38E-02
6	5.26E-03	8.68E-03	1.26E-02
7	6.34E-03	1.14E-02	1.45E-02
8	7.06E-03	1.01E-02	1.37E-02
9	6.39E-03	8.67E-03	1.22E-02
10	2.11E-02	2.34E-02	1.87E-02
11	7.78E-03	1.07E-02	1.45E-02
12	3.21E-02	2.51E-02	2.42E-02
13	1.19E-02	1.17E-02	1.45E-02
14	2.14E-01	7.26E-02	8.05E-02
15	1.18E-01	3.64E-02	3.93E-02
Average	3.41E-02	2.04E-02	2.25E-02

TABLE I
KL-divergence values for Sheather's method and the four proposed methods and all 15 distributions of the dataset. λ was chosen as 3 for TV and as 391 for FV.

In the first set of experiments N was chosen to be 1024. The experiments were carried out 500 times and the averaged results can be seen in Tables I-III. TV and FV are clearly the best performing algorithms, with each of them outperforming Sheather's method on average for both KL divergence and MISE. For

example, the average KL divergence values for Sheather, TV (λ=3) and FV (λ=391) are $3.41 \cdot 10^{-2}$, $2.04 \cdot 10^{-2}$ and $2.25 \cdot 10^{-2}$, respectively. Similarily, for MISE, the TV and FV methods produce better results on average as shown in Table III.

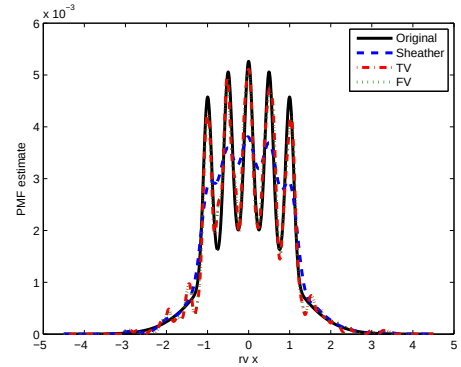
Distribution Number	Sheather	TV	FV
1	9.14E-06	2.14E-05	3.10E-05
2	1.39E-05	2.54E-05	3.08E-05
3	9.57E-05	7.85E-05	7.93E-05
4	7.01E-05	8.85E-05	7.59E-05
5	9.05E-05	1.14E-04	1.69E-04
6	1.10E-05	2.04E-05	3.08E-05
7	1.17E-05	2.38E-05	3.12E-05
8	1.58E-05	2.19E-05	3.12E-05
9	1.43E-05	2.02E-05	2.97E-05
10	7.63E-05	9.64E-05	5.28E-05
11	2.38E-05	3.16E-05	4.13E-05
12	7.50E-05	5.22E-05	5.09E-05
13	4.57E-05	3.72E-05	4.04E-05
14	1.38E-04	8.46E-05	8.70E-05
15	1.53E-04	7.14E-05	7.45E-05
Average	5.63E-05	5.25E-05	5.70E-05

TABLE II
MISE VALUES FOR SHEATHER'S METHOD AND THE FOUR PROPOSED METHODS AND ALL 15 DISTRIBUTIONS OF THE DATASET. λ WAS CHOSEN AS 3 FOR TV AND AS 391 FOR FV.

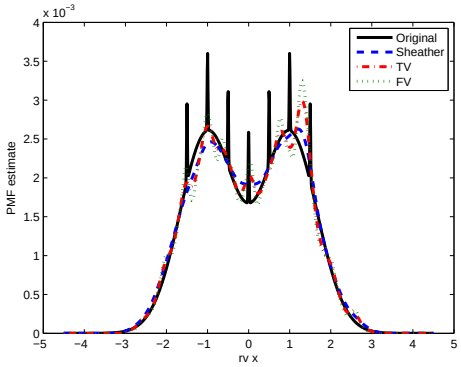
In case of KL divergence, TV and FV outperform Sheather's method in 7 out of 15 cases. All of these 7 distributions (namely 3, 4, 5, 10, 12, 14, 15) have significant support in the Fourier domain, i.e., they show a lot of peaks or sharp modalities. The presented algorithms are outperformed by Sheather's method when the distribution under test is low-pass in nature, e.g., distributions 1, 2, 6, 7, 8, 9.

Examples estimates for four different distributions are given in Figure 1. The three chosen distributions represent the different categories described above, namely distribution 10 and 15 (high-pass) and distribution 11 (high-pass with spikes). The different bandwidths of the investigated methods for all distributions are shown in Figure 2. For distributions 11 and 13, Sheather's method does not detect any of the sharp peaks as shown in Figure 1 b). In fact its estimates are rather poor. However, the TV and FV methods perform better when judged on visual inspection. In Figure 1 b), it detects four out of seven peaks.

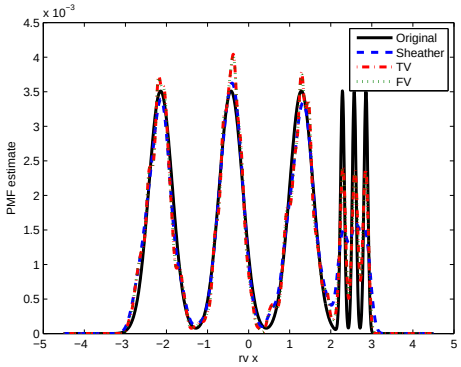
It is possible to increase the performance of the TV and FV methods by including some more obvious a priori information. One may divide the 15 example distributions used in this study into two classes based on their level of smoothness and spikiness: The smooth low-pass group consists of PDFs 1, 2, 6, 7, 8, 9 and the spiky high-pass group consists of PDFs 3, 4, 5, 10, 12, 14, 15. It is possible to guess the nature of the PDF by simply examining the histogram or the FFT of the histogram. For N=1024, λ was



(a) Distribution 10 from [3]



(b) Distribution 11 from [3]



(c) Distribution 15 from [3]

Fig. 1. KDE estimates using the proposed methods and Sheather's method for 3 out of 15 example distributions. N was chosen as 2^{10}.

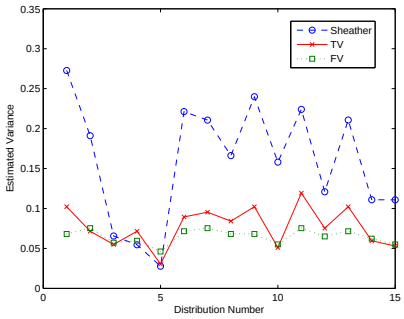


Fig. 2. Estimated $\hat{\sigma}$ for Sheather's method and the proposed methods.

increased to 7.2 for the low-pass group in TV and decreased to 2.1 for the high-pass group and resulting in an average KL divergence decrease of 10.8% to $1.82 \cdot 10^{-2}$. Similarily, the MISE was decreased on average by 15.5% to $4.43 \cdot 10^{-5}$. Using FV, λ was increased to 19550 for the low-pass group and decreased to 117.3 for the high-pass group resulting in an average KL divergence decrease of 19.1% to $1.82 \cdot 10^{-2}$. Similarily, the MISE was decreased on average by 20.9% to $4.51 \cdot 10^{-5}$. For TV, 5 out of 15 distributions and for FV, 11 out of 15 distributions yielded a smaller KL divergence than Sheather's method. Conversely, using TV, 7 out of 15 distributions and using FV, 9 out of 15 distributions yielded a smaller MISE than Sheather's method. The FV method provides better results than the TV method and Sheather's method, because the Lagrangian filter used for FV has sharper transitions from stopband to passband than the Haar filter used in TV, although both have the same cut-off frequency.

In the second set of experiments, the influence of different choices of N on the performance of the proposed methods was investigated. N was varied between $[2^5, 2^6, ..., 2^{12}]$. For each N, the experiments were again carried out 500 times for each distribution and results were averaged. The results for KL divergence can be seen in Figure 3. For illustration purposes, the natural logarithm of the average values was used for the plot. The TV and FV methods perform better than Sheather's method for N larger than 512. This is because it is difficult to estimate sharp peaks in the PDF with a low N value and Sheather's method usually provides smooth PDF estimates.

IV. Conclusion

This letter has shown alternative ways to compute the bandwidth of the kernel for KDE. The proposed methods have been compared with the commonly used Sheather method. On average, the total variation and filtered variation methods were found to result in higher fidelity than Sheather's method when measured under MISE and KL divergence. The proposed methods seem to be especially effective when

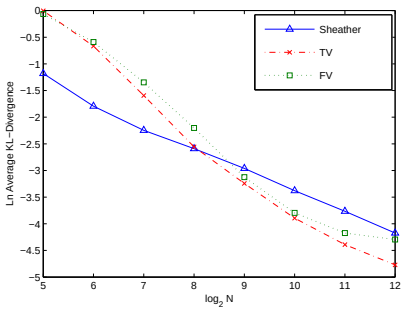


Fig. 3. The natural logarithm of the average performance of the compared methods versus N under KL divergence

the distribution to estimate is strongly multimodal. Since they utilize FFT and fast converging convex set iterations, the proposed methods are of low algorithmic complexity.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986, vol. 37, no. 1.

[3] J. S. Marron and M. P. Wand, "Exact mean integrated squared error," *Annals of Statistics*, vol. 20, no. 2, pp. 712–736, 1992.

[4] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. pp. 401–407, 1996.

[5] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, no. 3, pp. pp. 683–690, 1991.

[6] V. C. Raykar, R. Duraiswami, and L. H. Zhao, "Fast computation of kernel estimators," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 205–220, March 2010.

[7] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992.

[8] K. Kose, V. Cevher, and A. E. Cetin, "Filtered variation method for denoising and sparse signal processing," in *Acoustics, Speech, and Signal Processing, 2012. ICASSP-12 ., IEEE Intern. Conf. on*, 2012.

[9] M. P. Wand, "Fast computation of multivariate kernel estimators," *Journal of Computational and Graphical Statistics*, vol. 3, no. 4, pp. pp. 433–445, 1994.

[10] B. W. Silverman, "Algorithm as 176: Kernel density estimation using the fast fourier transform," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, pp. pp. 93–99, 1982.

[11] P. L. Combettes and J.-C. Pesquet, "Image restoration subject to a total variation constraint," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1213–1222, 2004.

[12] J. A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[13] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *Information Theory, IEEE Trans. on*, vol. 50, no. 6, pp. 1341–1344, 2004.

[14] P. L. Combettes, "The foundations of set theoretic estimation," *Proceeding of the IEEE*, vol. 81, no. 2, pp. 182–208, 1993.

[15] A. Cetin, O. Gerek, and Y. Yardimci, "Equiripple fir filter design by the fft algorithm," *IEEE Signal Proc. Magazine*, vol. 14, no. 2, pp. 60 –64, Mar 1997.

[16] C. Kim, A. Ansari, and A. E. Cetin, "A class of linear-phase regular biorthogonal wavelets," in *Acoustics, Speech, and Signal Processing ICASSP-92 ., IEEE Intern. Conf. on*, 1992, pp. 673–676.

[17] S. Mallat, *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications)*. Academic Press, Sep. 1999.