

An adaptive approximation for Gaussian wavelet kernel

Young-Mok Ha*, Ji Won Yoon**

*Electronics and Telecommunications Research Institute (ETRI), Republic of Korea

**School of Information Security, Korea University, Republic of Korea

**Corresponding author's Email: jiwon_yoon@korea.ac.kr

Abstract—Kernel machine plays a critical role in science community since temporal data become more important and popular with rapidly increasing big data analysis. A major problem for the machine is difficulty in constructing kernel function. We show that it is possible to adaptively estimate the parameters of Gaussian wavelet kernel in Laplace method. Our approach is constructed on an obvious fact that the gradient of the kernel with respect to a central variable of feature space becomes zero. It is remarkable that the complexity of our estimation method is $O(N)$ for N data. In order to validate the performance of the proposed approach, we simulate two kernel regression models which exploit the proposed approach on real electricity load data from Korea power exchange and electricity consumption data from Ireland's Commission for Energy Regulation.

Keywords—Gaussian wavelet kernel, adaptive inference, Laplace approximation, electricity load/consumption forecast.

I. INTRODUCTION

Kernel machine has been used to solve various problems such as regression, classification, clustering, and decision making for a long time. In using this kernel machine, procedure for determining model parameters including mode or covariance matrix of the kernel has to be carefully considered.

In this paper, we present an efficient estimation method to obtain the parameters of a famous kernel function, Gaussian wavelet kernel. Our approach is constructed on the obvious fact that the gradient of the Gaussian kernel with respect to a central variable in feature space becomes zero. In order to estimate them, [1] and [2] exploit cross-validation and Markov chain Monte Carlo (MCMC) sampling respectively. However, because of the high time complexity, the methods are not proper to use in adaptive parameter determination. For N data, the complexity of the k -fold cross-validation is $O(N^3/k)$. In case of MCMC, the accuracy varies depending on the number of simulations. It requires hours or even days to calculate accurate posterior so it is time-consuming. To avoid the heavy methods, Jaakkola and Haussler deal with probabilistic kernel using Bayesian sequential approach on parameter estimation [3] that allows to estimate the parameter reasonably. However, their approach does not solve the problems reflecting the sudden change of the trend. Using a recursive method is another commonly used method [4]–[7]. However, in using Newton's

method, Nelder-Mead simplex, or conjugate gradient, we also have to consider the problem of local maxima or convergence in this case.

Our proposed approach in this paper allows flexible estimation using probabilistic approach and Laplace approximation apart from problems mentioned above. The approach is constructed on an obvious fact that the gradient of the kernel with respect to a central variable of feature space becomes zero. It is remarkable that the complexity of our estimation method is $O(N)$.

The remainder of this paper consists of as follows. In the next section, we give short overviews of a Gaussian kernel and Gaussian wavelet kernel. In section III, the proposed approach is described. Data descriptions and the results of two applications are given in section IV. Finally, we present our conclusion in section V.

II. BACKGROUND KNOWLEDGE

A. Gaussian kernel

Kernel function plays a key role in kernel-based frameworks. The kernel is very useful in solving problems like non-linear problems that are difficult to deal with. In science literatures, this kernel function is generally used to measure the weight or similarity.

There are many ways to define kernel. As follows, using the projection and the inner product toward feature space is the most common way to define kernel:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \varphi(\mathbf{x}, \mathbf{x}')^T \varphi(\mathbf{x}, \mathbf{x}') \\ &= \sum_i \varphi_i(\mathbf{x}, \mathbf{x}')^T \varphi_i(\mathbf{x}, \mathbf{x}') \end{aligned}$$

We could reasonably define kernels such as Gaussian, Matérn class, γ -exponential, rational quadratic. Among them, this paper illustrates the most commonly used kernel, Gaussian kernel.

Gaussian kernel defined by equation (1)

$$k_G(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}|\mathbf{x} - \mathbf{x}'|_S^T \Sigma^{-1} |\mathbf{x} - \mathbf{x}'|_S\right) \quad (1)$$

has a very familiar form to us. The internal factors of this simple function explains what factors have to be considered in using kernel and how kernels can be used. The first factor

$|\mathbf{x} - \mathbf{x}'|_S$ is the operator that measures the distance between $\mathbf{x}$ and $\mathbf{x}'$ in feature space S . Before we use the operator, which feature space to consider and which distance measure to use have to be determined. Inappropriate and irrelevant choice ruins the machine using this function so a proper choice has to be made. The minus mark is the second factor. This mark leads to the increase in the value of equation (1) as the distance measure value decreases, so it makes equation (1) to measure similarity. In other words, as the two variables get similar, the greater the values will be. The last factor, Σ is the parameter that adjust the value of $|\mathbf{x} - \mathbf{x}'|_S^2$. By defining relation between bases which consist of S , this parameter gives weight to each elements of $|\mathbf{x} - \mathbf{x}'|_S^2$, and highly controls the performance of the kernel machine using equation (1). Therefore, to maximize the use of the advantages of kernel machine, this parameter has to be carefully determined.

B. Gaussian wavelet kernel

Gaussian wavelet kernel is a Gaussian kernel that S is defined by wavelet spaces. Consider discrete wavelet kernel with discrete wavelet transform (DWT) as an example. DWT is one of the most popular candidates among the signal processing transforms. The transform gives us independent multi-scale resolutions and analyses a discrete signal $\mathbf{x}$, $\mathbf{x} \in L^2(\mathbb{R}) = \{f \mid \sum \|f\|^2 < \infty\}$ by decomposing a local signal with scale function, φ , and mother wavelet, ψ . For example, using decomposition level $J = 3$, we can approximate a stationary discrete signal $\mathbf{x}$ of length 2^5 as

$$\mathbf{x} = \sum_{i=0}^{2^4-1} \Psi_{1,i}^{x[i]} \psi_{1,i} + \sum_{j=0}^{2^3-1} \Psi_{2,j}^{x[j]} \psi_{2,j} + \sum_{k=0}^{2^2-1} \Psi_{3,k}^{x[k]} \psi_{3,k} + \sum_{r=0}^{2^1-1} \Phi_{3,r}^{x[r]} \varphi_{3,r}$$

where $\Phi_{s,\tau}^x = \langle \mathbf{x}, \varphi_{s,\tau} \rangle$ and $\Psi_{s,\tau}^x = \langle \mathbf{x}, \psi_{s,\tau} \rangle$. The operator $\langle a, b \rangle$ denotes the sum of the elements of a times b . The parameters $s, \tau \in \mathbb{R}$, $s > 0$ are the dilation and translation factors and notations $\varphi_{s,\tau}$ and $\psi_{s,\tau}$ are translated, and scaled versions of φ and ψ , respectively. The mutual orthogonality of each wavelet and scale allows us to consider a local signal in several independent spaces simultaneously. This means that the probability of $\mathbf{x}$ in real-time space can be measured jointly:

$$p(\mathbf{x}) \propto p_{\Phi_J}(\mathbf{x}) \prod_{j=1}^J p_{\Psi_j}(\mathbf{x}).$$

Thus, a Gaussian (discrete) wavelet kernel, can be defined by

$$k_{DWT}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2} \Phi_{j,k_\Phi}^z T \Sigma_{\Phi_{j,k_\Phi}}^{-1} \Phi_{j,k_\Phi}^z\right) \times \prod_{j=1}^J \exp\left(-\frac{1}{2} \Psi_{j,k_\Psi}^z T \Sigma_{\Psi_{j,k_\Psi}}^{-1} \Psi_{j,k_\Psi}^z\right). \quad (2)$$

Here, $\Phi_{j,k_\Phi}^z = [\Phi_{j,0}^z, \Phi_{j,1}^z, \dots, \Phi_{j,k_\Phi-1}^z]$ and $\Psi_{j,k_\Psi}^z = [\Psi_{j,0}^z, \Psi_{j,1}^z, \dots, \Psi_{j,k_\Psi-1}^z]$. $\mathbf{z}$ denotes $\mathbf{x}_j - \mathbf{x}_i$.

III. PROPOSED APPROACH

In this section, for simple description, we propose our approach for (1). It is straightforward to generalize the approach for (2).

For modeling an elaborate structure of kernel based machine, it is necessary to find the parameters. The process to estimate them is very important since how we determine it decides performance. Although there are lots of solutions to estimate the parameter such as CV or MCMC, in general, they are not without time cost problem. Moreover, the time complexity of those approaches are sometimes exponentially increasing as the dimension of the parameter become large.

In this paper, we propose an efficient estimation method based on Laplace approximation. In general, the approximation consists of two steps; finding mode and estimating covariance matrix by calculating Hessian matrix at the given mode. The most frequently used approach for the first step is recursive method such as Nelder-Mead simplex. However, for accuracy and computational efficiency, we avoid to use it. The major problem of it is local maxima problem. In addition, it requires un-predictable memory or time consumption. To reduce loss that stems from them, we use the global maxima which is given at every moment. For the work, we focus on the following important properties of normalized similarities

$$k_{nor}(\mathbf{x}_i, \mathbf{x}_n) = \frac{k_G(\mathbf{x}_i, \mathbf{x}_n)}{\sum_{j=1} k_G(\mathbf{x}_j, \mathbf{x}_n)} \quad (3)$$

between $\mathbf{x}_n$ and $\mathbf{x}_{1:n-1}$ defined by equation (3):

- A feature space of a kernel machine which uses equation (3) is defined by S (see II-A Gaussian kernel function) and $\mathbf{x}_n$.
- If $\mathbf{x}_n$ changes, the location of S where the machine focuses changes.
- The range is from zero to one, and the total sum is one.
- For a central variable in the feature space, $\mathbf{x}_i = \mathbf{x}_n$, equation (3) becomes one. This point is a global maxima in similarity domain.

The third property allows equation (3) to become probability in the feature space or the weight for weighted averaging process. And the last property implies the fact that the kernel is smooth in distance domain and the gradient of the Gaussian kernel with respect to the co-variation becomes zero. Note that the property makes the estimation avoid to find local maxima. From these properties, a $p(\mathbf{x}_i)$ in a feature space of a kernel machine expressed as

$$p_{\mathbf{x}_n}(\mathbf{x}_i) \propto k(\mathbf{x}_i, \mathbf{x}_n).$$

Moreover, with inverse Wishart distribution for a parameter sets $\{\Sigma^{-1}, \nu, \alpha\}$, we can represent a posterior over Σ by

$$p(\Sigma | \mathbf{x}_{1:n}, \Sigma^{-1}, \nu, \alpha, \theta) = p(\Sigma | \Sigma^{-1}, \nu, \alpha) p(\mathbf{x}_{1:n} | \Sigma, \theta)$$

where

$$p(\Sigma | \Sigma^{-1}, \nu, \alpha) \propto |\Sigma|^{\frac{\nu+\alpha+1}{2}} \exp\left(-\frac{1}{2} \text{Tr}(\Sigma^{-1} \Sigma^{-1})\right).$$

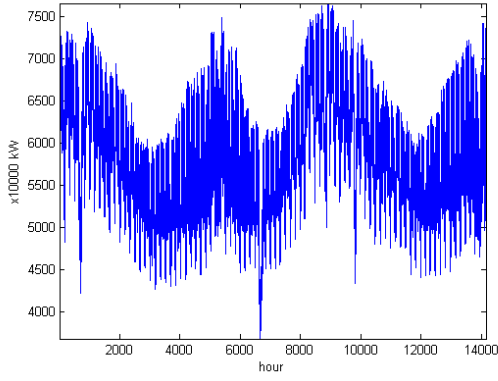


Figure 1. KPX electricity load data plot

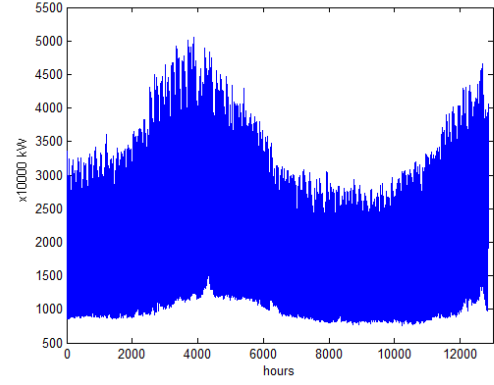


Figure 3. CER electricity demand data plot

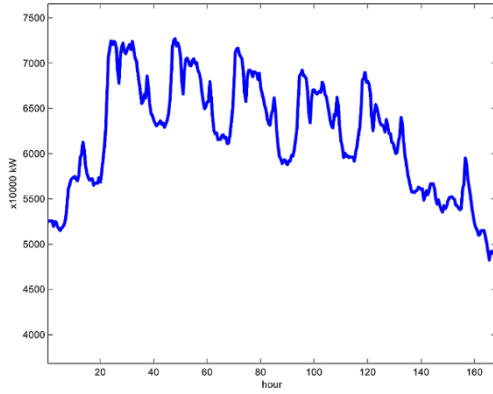


Figure 2. KPX electricity load data plot for one week

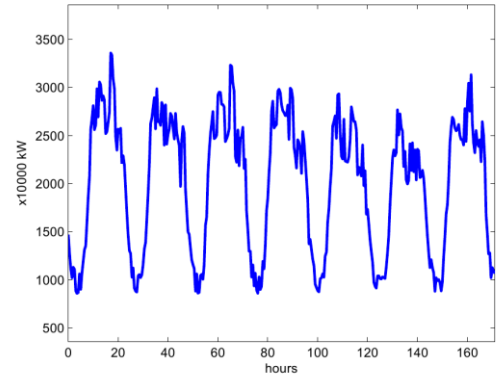


Figure 4. CER electricity demand plot for one week

The proposed approach exploits Laplace approximation at $\mathbf{x}_n$ and suggests that

$$\Sigma_{MAP} \propto \Omega^{-1} + S$$

where $S \propto \sum_{i=0}^{n-1} (\mathbf{x}_i - \mathbf{x}_n)(\mathbf{x}_i - \mathbf{x}_n)^T$ with $\nabla_x^2 p(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_n}$. Namely, S^{-1} is Hessian matrix of the kernel at $\mathbf{x}_n$. This method illustrates the Σ can be calculated through the complexity of $O(N)$. With this small complexity, we could estimate the parameter adaptively. In times when we are using MCMC, CV, and large data set, the fact that we do not estimate the parameter every time due to computational burden can be quite meaningful. Especially, it is more meaningful when several Σ s are estimated quickly in multi-dimension.

IV. APPLICATIONS

In this section, we forecast electricity load/consumption to verify the validity of the proposed approach. Electricity load/consumption forecasting is an important factor in the operation planning, maintenance, and capacity evaluation of a system. Researchers in many countries have studied and applied various techniques to forecast electricity load or demand.

For performance analysis, we implemented two main techniques for the kernel based regression; *Weighted k-nearest*

neighbors (k-NN) regression and *k-NN applied Gaussian process regression*. In addition, we used DWT among {complex, continuous, discrete}-wavelet transform since the type of the data is real time series.

A. Data description & Variable setting

We use two observation sets, shown in Figures 1-4. Figure 1 and 2 are the public on-line data of the Korea Power Exchange (KPX) of the Republic of Korea. KPX is a public company located in the Republic of Korea. The company provides power data on its Web-page, <http://www.kpx.or.kr>. In this experiment, using a python implementation, we crawled the data for the period between 00:00, Dec. 25th 2011 and 23:55, Aug. 12th 2013. In order to solve the memory problem efficiently, we pre-processed the data to provide an interval of 30 minutes. The KPX data show three intervals: a week, 3 months, and a year.

Another set of observations, the total and published data of Ireland's Commission for Energy Regulation is also used. The data were measured every 30 minutes. To solve the forecast problem, we created 30 minute intervals of the total power supply data in a time series model using the given data. The CER data has periods of a day, week and year. More detail about the data can be found at <http://www.ucd.ie/issda/data/>.

For the forecast, we refine random variables. We divide a given trajectory of discrete time stationary stochastic process, $O = \{O_n, n \in \mathbb{N}\}$, into sequential blocks. The stochastic process over the block, $X = \{x_i, i \in \mathbb{N}\}$, is expressed as

TABLE I
NUMERICAL RESULTS ON ELECTRICITY LOAD DATA OF REPUBLIC OF KOREA & ELECTRICITY CONSUMPTION DATA OF REPUBLIC OF IRELAND
WITH GASUSSIAN DISCRETE WAVELET KERNEL (GDWK). (N)RMSE : (NORMALIZED) ROOT MEAN SQUARE.

Electricity load data of Republic of Korea	RMSE	NRMSE (%)
Weighted k -NN Regression with GDWK	83.96	2.1
Weighted k-NN Regression with GDWK + Proposed approach	69.44	1.8
k -NN applied Gaussian Process with GDWK	977.47	24.6
k -NN applied Gaussian Process with GDWK + Proposed approach	969.95	24.4
Electricity consumption data of Republic of Ireland	RMSE	NRMSE
Weighted k -NN Regression with GDWK	328.25	7.7
Weighted k -NN Regression with GDWK + Proposed approach	306.07	7.1
k -NN applied Gaussian Process with GDWK	291.40	6.8
k-NN applied Gaussian Process with GDWK + Proposed approach	282.57	6.6

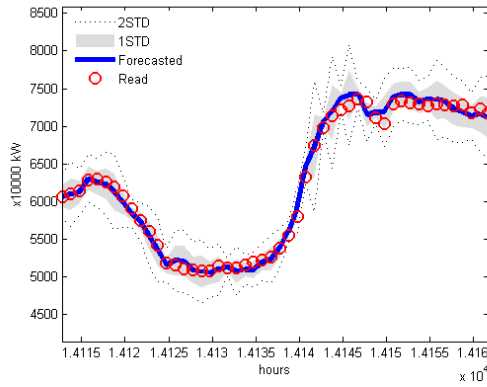


Figure 5. One day forecasting of KPX using k -NN Regression with Gaussian discrete wavelet kernel + Proposed approach ($l=64, k=1050, J=1$)

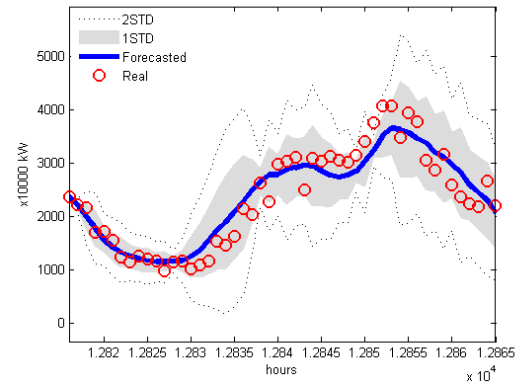


Figure 6. One day forecasting of CER using k -NN applied Gaussian process with Gaussian discrete wavelet kernel + Proposed approach ($l=56, k=100, J=1$)

$$\mathbf{x}_i = O\{k + (i - 1)l\}, \quad i, k \in \mathbb{N}, \forall k \in (0, l]$$

where l is block length. This approach is well known in the statistical community since it enables us to understand the whole progress of the stochastic process. Moreover, the index is reordered according to certain criterion such as Euclidean distance from $\mathbf{x}_n$ in the space. We denote the reordered pair by $\{\tilde{\mathbf{x}}_i, \tilde{y}_i\}$ where $y_i = O\{(i \times l) + 1\}$. Remind that our task is to estimate the $\hat{y}_n$, finding the relations between the elements of $\tilde{\mathbf{x}}_{1:n}$ ($= \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_n\}$) in the space.

B. Experimental Results

To verify the validity of the proposed approach, we forecasted the electricity load/consumption through the kernel regression models using proposed model, and same models using Nelder-Mead simplex for one day(=forty eight steps).

To find the most accurate model, we used normalized root-mean square error (NRMSE) based cross-validation on parameters $\{l, k, J\}$. Each parameters are as follows; l is block size (see IV-A Data description & Variable setting) and k is the clustering parameter of k -NN. J is decomposition level of wavelet transform. In case of Gaussian process, the rest of the

parameters are determined by the Nelder-Mead method based maximum likelihood.

Table I shows the numerical results. The graphical results of the most accurate models shown in Figures 5 and 6 have margins of 1 or 2 of standard deviation (STD) at each trial. The most accurate models in each case are the Weighted k -NN Regression and k -NN applied Gaussian process in the case of KPX and CER, respectively, and the accuracy of the proposed model is higher in *all* cases. In the first case, the accuracy differs by 0.3%, and in the second case, it differs by 0.2%. The results show that the computationally efficient approach we propose is useful for kernel machines.

V. CONCLUSIONS

In this paper, we have proposed an adaptive estimation method of $O(N)$ which is based on Laplace approximation for Gaussian wavelet kernel. Moreover, two kernel estimators were implemented in electricity load/consumption forecast to verify validation of the approach. The approaches are compared in the perspective of accuracy using NRMSE metrics. For the performance evaluation, we used public on-line data of the Korea Power Exchange (KPX) that is real electricity load data of Republic of Korea and electricity consumption data from Ireland's Commission for Energy Regulation. In the

experiments, kernel based regression approaches, which exploit the proposed approach, outperform the models which do not.

ACKNOWLEDGMENT

This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (MSIP) (NRF-2013R1A1A1012797), and by the Information and Communications technologies (ICT) R&D program of MSIP/IITP (Institute for Information & communications Technology Promotion) [B0101-15-0663, Safety-critical Distributed Modular SW Platform].

REFERENCES

- [1] M. Sinn, J. W. Yoon, F. Calabrese, and E. Bouillet, "Predicting arrival times of buses using real-time GPS measurements", in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on. IEEE, 2012*, pp. 1227-1232.
- [2] R. L. Hyndman, X. Zhang, and M. L. King, "Bandwidth selection for multivariate kernel density estimation using MCMC," in *Econometric Society 2004 Australasian Meetings*, no. 120. Econometric Society, 2004.
- [3] T. Jaakkola and D. Haussler, "Probabilistic kernel regression models," in *Proceedings of the 1999 Conference on AI and Statistics*, vol. 126. San Mateo, CA, 1999, pp. 00-04.
- [4] M. Deisenroth, D. Fox, and C. Rasmussen, "Gaussian processes for data efficient learning in robotics and control", 2013.
- [5] S. Roberts, M. Osborne, M. Ebdon, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, 2013.
- [6] C. K. Williams and C. E. Rasmussen, "Gaussian processes for regression", 1996.
- [7] C. E. Rasmussen, "Gaussian processes for machine learning", 2006.



Young-Mok, Ha received the B.Sc. degree in the college of information and communication, Korea University, Republic of Korea. He obtained M.Sc. degree in graduate school of information security, Korea University, 2014. He is currently working for Electronics and Telecommunications Research Institute (ETRI), Republic of Korea. His research interest is embedded system, operating systems, cyber security, and Bayesian statistics.



Ji Won Yoon received the B.Sc. degree in information engineering at the SungKyunKwan University, Republic of Korea. He obtained the M.Sc. degree in School of informatics at the University of Edinburgh UK in 2004 and the Ph.D. degree in signal processing group at the University of Cambridge UK in 2008 respectively. After doing post-doctoral research at robotics group, the University of Oxford and statistics group, Trinity College Dublin, he moved to IBM Research as a research scientist to do big data analysis. Since 2012 he is an assistant professor at Korea University. His current research is Bayesian signal and information processing for cyber defense.