Statistics and Its Interface Volume 11 (2018) 425–428

Discussion on "Doubly sparsity kernel learning with automatic variable selection and data extraction"

Hao Helen Zhang*

Kernel methods provide powerful and flexible tools for nonlinear learning in high dimensional data analysis, but feature selection remains a challenge in kernel learning. The proposed DOSK method provides a new unified framework to implement kernel methods while automatically selecting important variables and identifying a subset of parsimonious knots at the same time. A double penalty is employed to encourage sparsity in both feature weights and representer coefficients. The authors have presented the computational algorithm and as well as theoretical properties of the DOSK method. In this discussion, we first highlight the DOSK's major contributions to the machine learning toolbox. Then we discuss its connections to other nonparametric methods in the literature and point out some possible future research directions.

AMS 2000 subject classifications: Primary 62H20, 62F07; secondary 62J05.
Keywords and phrases: Reproducing kernel Hilbert space (RKHS), Kernel methods, Variable selection, High dimensional data analysis, Penalty.

1. INTRODUCTION

When the data dimension is high and not all features are informative to describe the authors are to be congratulated with their excellent contribution to kernel learning, an important area in statistical machine learning for high dimensional data analysis. Kernel machines provide a unified and powerful framework to achieve nonlinear learning for various tasks such as support vector machines for classification, kernel logistic regression, Gaussian processes, spectral clustering, and kernel principal component analysis for dimension reduction ([23, 24, 18, 3]). Kernel methods can be applied to various input domains, including real-valued vectors, categorical data, sequence data, text, and images. Due to their high flexibility, computational efficiency, and the ability to discover complex patterns and extract nonlinear features from data, kernel methods are widely used in many real applications such as medicine, climate studies, imaging sciences, and deliver state-of-the-art performance. Considering the relationship between the input and the output, it is vital to identify important features and perform dimension reduction during the learning process in order to improve interpretability and prediction accuracy of the decision rule obtained from data. This is the called variable selection, which plays a key role to facilitate computation and statistical inferences in high dimensional data analysis. However, standard kernel machines can not select important variables automatically, due to the complex form of the estimated multivariate function. In this article, the authors have proposed a novel learning method, DOSK, to achieve sparse learning in kernel methods. By solving a regularization problem with a double-penalty, the DOCK estimator can achieve two types of sparsity simultaneously, i.e., the solution has a sparse representation in terms of both variables and data points. Furthermore, the DOSK estimator is shown to enjoy nice theoretical properties – consistent in both variable selection and function estimation asymptotically, under certain regularity conditions.

We would like to supplement this article with three points. First, we share our view on important contributions of this work to the area of kernel learning. Second, we point out the connection between the data extraction idea proposed by the authors with the parsimonious knot selection commonly used in splines. Third, we suggest some future directions to improve the DOSK method.

*Department of Mathematics, University of Arizona, Tucson AZ 85721.

2. MAIN CONTRIBUTIONS OF THIS WORK

Consider a supervised learning problem with the input $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ and an output Y, which can be real-valued or take discrete values. The goal of machine learning is to estimate the relationship between $\mathbf{X}$ and Y, denoted by a function f, from the observations $(\mathbf{x}_i, y_i), i = 1, \cdots, n$. The learned function f can be used to make future predictions by $\hat{y}_{new} = f(\mathbf{x}_{new})$, when a new input $\mathbf{x}_{new}$ is given. The function f can be linear or nonlinear in $\mathbf{X}$, depending on the complexity of the underlying true function. Kernel methods provide is a unified way to achieve nonlinear learning. Define a map $\phi : \mathcal{X} \to \mathcal{F}$, from $\mathcal{X}$ to the feature space $\mathcal{F}$, where ϕ is a linear or nonlinear feature map. A kernel function can be defined as an inner product in the feature space $\mathcal{F}$,

$$K(\mathbf{x}, \mathbf{x}') =< \phi(\mathbf{x}), \phi(\mathbf{x}') >_{\mathcal{F}}, \ \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

For example, if $\mathcal{X} = \mathbb{R}^3$ with $\mathbf{x} = (1, x_1, x_2)$, then a nonlinear map $\phi : \mathbb{R}^3 \to \mathbb{R}^5$ defined as $\phi(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$ defines a kernel $K(\mathbf{x}, \mathbf{x}') = <\phi(\mathbf{x}), \phi(\mathbf{x}')>_{\mathcal{F}} = (1+ < \mathbf{x}, \mathbf{x}' >)^2_{\mathcal{X}}$, which is known as the second-order polynomial kernel. One important property of kernel functions is that they can be computed without explicitly computing ϕ, which enables them to operation in a high-dimensional feature space; this is called "kernel trick". Kernel methods are widely used in nonlinear classification and nonlinear function estimation problems. For real-valued inputs in $\mathbb{R}^p$, popular choices of kernels include linear kernel $K(\mathbf{x}, \mathbf{x}') =< \mathbf{x}, \mathbf{x}' >$, dth-order polynomial kernel $K(\mathbf{x}, \mathbf{x}') = (1+ < \mathbf{x}, \mathbf{x}' >)^d$, Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp\{-\gamma\|\mathbf{x} - \mathbf{x}'\|^2\}$, and the Laplacian kernel $K(\mathbf{x}, \mathbf{x}') = \exp\{-\gamma \sum_{j=1}^p |x_j - x_j'|\}$.

Kernel methods are closely connected to nonparametric function estimation in statistics through the reproducing kernel Hilbert space (RKHS); see [22, 8, 26]. Based on [2], if the kernel function K is positive definite on $\mathcal{X} \times \mathcal{X}$, then there exists a unique RKHS of real-valued functions on $\mathcal{X}$, with K as its reproducing kernel. We denote the RKHS associated to K by $\mathcal{H}_K$. For the standard nonparametric regression model $y_i = f(\mathbf{x}_i) + \epsilon_i, i = 1, \ldots, n$, where ϵ_i is the error term with mean zero and constant variance, the function $f \in \mathcal{H}_K$ can be estimated by solving a regularization problem

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n L\{y_i, f(\mathbf{x}_i)\} + \lambda \|f\|^2_{\mathcal{H}_K}, \tag{1}$$

where L is the loss function and $J(f)$ is the penalty function to control the complexity of f and avoid over-fitting. By the representer theorem of [14], the solution to (1) can be expressed as a finite combination of kernel representers

$$\hat{f}(\mathbf{x}) = b + \sum_{i=1} \alpha_i K(\mathbf{x}, \mathbf{x}_i). \tag{2}$$

The form (2) suggests that the solution $\hat{f}$ typically depends on all the data points through α's and as well as all the inputs through the kernel function K. This is not necessarily desired for high dimensional data analysis, as some variables may be uninformative. The DOSK method proposed in this article is produce a sparse solution $\hat{f}$, which is sparse in terms of both data points and variables. In the following, we comment on two major contributions of the DOSK to kernel learning and its novelty compared to other existing methods.

Variable selection is a challenging issue for kernel learning due to the complex form of $\hat{f}$. If the true function f is not additive, then the effects of individual variables can not be easily separated and evaluated to determine their importance. For example, when using the Gaussian kernel $K(\mathbf{x}, \mathbf{x}')$, the estimated function is expressed as $\hat{f}(\mathbf{x}) = b + \sum_{j=1}^p \alpha_i e^{-\gamma \sum_{j=1}^p (x_j - x_{ij})^2}$, where effects of all the variables are tangled together in a nonlinear fashion, then it would be difficult to separate them out and determine contributions for individual variables to $\hat{f}$. This is mainly why many nonparametric variable selection methods, including COSSO ([15]), SpAM ([16]), and [12], conduct variable selection by assuming that f has an additive structure as in [11]. For example, the COSSO method assumes that the kernel K has the structure $K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^p \theta_j K_j(x_j, x_j')$, where $K_j(x_j, x_j')$ is the kernel associated with variable X_j, and select variables by identifying nonzero θ_j's. The SpAM method assumes the additive model $f(\mathbf{x}) = b + \sum_{j=1}^p f_j(x_j)$ and select important f_j's by imposing shrinkage-type penalty onto f_j's. In contrast to these methods, the DOSK method does not require the additive assumption on f to perform variable selection automatically. This is done by adopting the KNIFE technique suggested by [1], which associates each variable X_j with a weight parameter $w_j \in (0, 1)$ and uses the "variable-weighted" kernel (called the "feature-weighted" kernel in [1]) $K_{\mathbf{w}}(\mathbf{x}, \mathbf{x}') = K(\mathbf{w} \odot \mathbf{x}, \mathbf{w} \odot \mathbf{x}')$ in kernel learning, where $\mathbf{w} = (w_1, \ldots, w_p)^T$. The nice property of $K_{\mathbf{w}}$ is that it controls the significance of the effect of X_j by a scalar $w_j \in (0, 1)$; and if $w_j = 0$, then the effect of X_j will vanish in the function $\hat{f}$. In this way, the DOSK method converts the problem of variable selection selection to the problem of selecting nonzero w_j's, which can be easily done by imposing a shrinkage penalty like the L_1 penalty on w_j's as in [21]. With regard to this, the DOSK method is more flexible than other methods by relaxing the additive assumption on the underlying true function f.

The second contribution of this work is to the established theory for the DOSK estimator. Under certain regularity conditions, the DOSK estimator is shown to be consistent in both variable selection and nonparametric function estimation asymptotically. Interestingly, the authors have shown that the convergence rate of the DOSK estimator is very close to the parametric rate, which is quite different from other nonparametric regression estimators. It would be great if the authors can provide more explanations and share their insight on the results in the article.

In summary, the DOSK method provides a theoretically-justified, computationally efficient, and flexible tool for nonparametric variable selection in the context of kernel learning. The numerical studies in the paper also show that the empirical performance of the DOSK method is superior to other methods in the settings considered by the authors. Overall, the DOSK contribute a valuable and useful sparse estimation tool for non-linear kernel learning.

3. DATA EXTRACTION VS PARSIMONIOUS KNOT SELECTION

One main feature of the DOSK method is its ability to extract data, which is done by imposing the L_1 penalty on the kernel coefficient α's in $\hat{f}$. As a result, the DOSK estimator can be expressed by a combination of a subset of

kernel representers as

$$\hat{f}(\mathbf{x}) = b + \sum_{\alpha_i \neq 0} \alpha_i K(\mathbf{x}, \mathbf{x}_i).$$

As pointed out by the authors, the DOSK method automatically selects important data points from the entire training set. In the following, we point out the close connection between data extraction to parsimonious knot selection, which is commonly used in computing the spline function in nonparametric multivariate regression. Both of them aim to achieve the sparsity in α's, but they use different techniques.

For smoothing splines, the solution $\hat{f}$ resides in a finite-dimensional model space

$$\mathcal{H}_n = \{1\} \oplus \text{span}\{K(\mathbf{x}, \mathbf{x}_i),\ i = 1, \cdots, n\},$$

where $\{1\}$ is the constant space and K is the reproducing kernel. As pointed by [9], the effective dimension of the model space $\mathcal{H}_n$ is actually much lower than n due to the penalty term $\lambda J(f)$. Therefore, it is not necessary to use all $1 + n$ dimensions to obtain $\hat{f}$, and instead one can approximate $\hat{f}$ within a low-dimensional subspace of $\mathcal{H}_n$,

$$\mathcal{H}_N = \{1\} \oplus \text{span}\{K(\mathbf{x}_i, \mathbf{x}^*),\ i = 1, \cdots, N\}$$

where $\{\mathbf{x}_i^*, i = 1, \ldots, N\}$ are a random subset of $\{\mathbf{x}_i, i = 1, \ldots, n\}$. It can be shown, as $N \to \infty$ sufficiently fast (but slower than n), the approximate solution has the same convergence rate as the original solution $\hat{f}$; see [13] for details. This is known as the parsimonious knot selection, which is commonly employed in smoothing splines and regression splines (for example, [4, 29, 17, 30]). To select the knots, one can either randomly sample a subset of N points $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*\}$ from the training data, or use selection techniques such as stepwise regression and penalized least squares (for example, [6, 5, 19, 20]), to select the knots. For smoothing splines, the computation of $\hat{f}$ is generally of the order $O(n^3)$ in multivariate settings, while that for computing the approximate solution is $O(nN^2)$. Furthermore, [9] showed that N be can be much smaller than n without degrading the function estimation. Generally speaking, when N is chosen properly and much smaller than n, the computational saving can be substantial for large datasets,

Based on the above discussions, the DOSK method can be also regarded as a new knot-selection technique for nonparametric estimation. Here, we would like to point out two main differences between the DOSK method and other knot-selection techniques. The first difference is in their knot selection mechanism: the DOSK selects knots by retaining large $|\alpha_i|$'s with the help of a shrinkage penalty, while the latter involves a random sub-sampling process. The second key difference is on their choice of N and knot locations. In theory, efficient approximation of the solution can be achieved with a small N, but in practice, the selection of knots can be delicate and data-dependent. Most existing knot selection techniques select N in some ad hoc fashion, or use some empirical formula like $N = 10n^{2/9}$ as in cubic splines suggested by [13]. By contrast, the DOSK procedure does not need to specify N or the knot locations in advance, and instead it lets data decide the number of knots and their locations automatically by solving a penalization problem. In this sense, the DOSK is more convenient for implementation. It would be interesting to make empirical comparisons of them under some numerical studies.

4. SUGGESTIONS FOR FUTURE WORKS

The DOSK method involves three tuning parameters, λ_1, λ_2, λ_3, and possibly additional parameters in the kernel (for example, γ in the Gaussian kernel). The performance of the DOSK estimator hinges on the proper selection of these smoothing parameters via some data-adaptive procedures. In this article, $\lambda_3 = 0.5$ is just used for convenience. However, it is known that the tuning process can be expensive for multiple tuning parameters. In order to speed up the tuning process, the authors may explore the possibility of building a solution-path or solution-surface over the tuning parameters or some of them. It is observed that, when $\mathbf{w}$ is fixed, the DOSK method essentially solves the elastic net penalty problem for α's, then it is possible to compute the entire regularization solution path of α's by adopting the LARS-EN algorithm in [31]. If this can be done, it will help to save the tuning cost for λ_1 and λ_2. In the context of smoothing splines, a commonly used selection criterion for λ_3 is the generalized cross validation ([25, 26, 27, 28]), which has been shown to perform well in practice. The authors may consider deriving a GCV-type tuning criteria for tuning λ_3, as an alternative to the five-fold cross validation.

Interaction selection is another important yet challenging topic for nonlinear learning problems. Though there are some recent developments for interaction selection in high-dimensional linear models such as [10], the problem of selecting nonlinear interactions is much less studied in the literature. When the dimension p is smaller than the sample size n, [15] employed the COSSO for interaction selection by using the tensor product kernel. It would be interesting to study interaction selection for nonparametric models with p much larger than n. Since the DOSK method does not make any special structure assumption on $\hat{f}$, it may be generalized to identify important two-way or multi-way nonlinear interactions between variables for complex problems. The authors intend to explore this in future research.

ACKNOWLEDGEMENTS

This research is supported in part by National Science Foundation grants 1740858 and DMS-1418172, National Institute of Health grant R01 DA044985, and Natural Science Foundation of China grant NSFC-11571009. The authors thank the Editor for the invitation on providing this discussion.

Received 9 June 2018

REFERENCES

[1] Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, **22** 284–299. MR3173715

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68** 337–404. MR0051437

[3] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*, Cambridge University Press, Cambridge.

[4] de Boor, C. (1978). *A Practical Guide to Splines*, Springer, New York. MR0507062

[5] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19** 1–141. MR1091842

[6] Friedman, J. H. and Silverman, B. (1989). Flexible parsimonious smoothing and adaptive modeling (with discussion). *Technometrics*, **31** 3–39. MR0997668

[7] Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, Boca Raton. MR1270012

[8] Gu, C. (2002). *Smoothing spline ANOVA models*, Springer-Verlag. MR1876599

[9] Gu, C. and Kim, Y. J. (2002). Penalized likelihood regression: general formulation and efficient approximation. *Canadian Journal of Statistics*, **30** 619–628. MR1964431

[10] Hao, N. and Zhang, H. H. (2014). Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association*, **109** 1285–1301. MR3265697

[11] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*, Chapman and Hall. MR1082147

[12] Huang, J., Horovitz, J. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38** 2282–2313. MR2676890

[13] Kim, Y-J. and Gu, C. (2002). Penalized least square regression: fast computation via efficient approximation. *Technical report*, Department of Statistics, Purdue University, West Lafayette, Indiana. MR1964431

[14] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline function. *Journal of Mathematical Analysis and Applications*, **33** 82–85. MR0290013

[15] Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, **34** 2272–2297. MR2291500

[16] Ravikumar, P, Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society, Series B*, **71** 1009–1030. MR2750255

[17] Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, **45** 205–223.

[18] Schölkopf, B. and Smola, A. (1998). *Learning with kernels*, The MIT Press.

[19] Stone, C., Buja, A., and Hastie, T. (1994). The use of polynomial splines and their tensor-products in multivariate function estimation. *Annals of Statistics*, **22** 118–184. MR1272079

[20] Stone, C., Hansen, M., Kooperberg, C., and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* **25**, 1371–1425. MR1463561

[21] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society, Series B.*, **58** 147–169. MR1379242

[22] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[23] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. MR1367965

[24] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York. MR1641250

[25] Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problems. *Annals of Statistics*, **13** 1378–1402. MR0811498

[26] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics, volume 59. MR1045442

[27] Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics*, **23** 1865–1895. MR1389856

[28] Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, **108** 1122–1145.

[29] Xiang, D. and Wahba, G. (1998). Approximate smoothing spline methods for large data sets in the binary case. *Proceedings of ASA Joint Statistical Meetings, Biometrics Section*, 94–98.

[30] Yau, P., Kohn, R., and Wood, S. (2002). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, **12** 23–54. MR1965210

[31] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*, **67** 301–320. MR2137327

Hao Helen Zhang
617 N. Santa Rita Ave.
Department of Mathematics
University of Arizona
Tucson, AZ 85721
USA
E-mail address: hzhang@math.arizona.edu
url: www.math.arizona.edu/~hzhang