Statistics and Its Interface Volume 11 (2018) 429–431

Rejoinder of “Double sparsity kernel learning with automatic variable selection and data extraction”

Jingxiang Chen, Chong Zhang,
Michael R. Kosorok, and Yufeng Liu*

We would like to thank the editor Professor Heping Zhang for organizing this discussion, and the discussants for their thoughtful insights. These discussions help us to understand DOSK further and inspire interesting connections with existing methods as well as future research directions.

The main goal of the proposed DOSK technique is to perform flexible high dimensional nonlinear learning. In particular, both variable selection and data extraction are built in the regularization terms for DOSK to achieve the flexible learning goal.

Meimei Liu and Guang Cheng point out that DOSK relies on the finite sparsity assumption of the true underlying function, i.e., Assumption 4. As a potential solution to remove this assumption, they propose two methods to achieve the goal of sparsity, and compare them with DOSK using simulation studies. Their first idea is to use a random matrix to project the original data into a smaller surrogate data matrix in a random manner. In general, we agree that the computation can be more efficient by reducing the size of data. Under supervised learning, however, we suspect it may be difficult to justify which random matrix should be the optimal one to use if the selection is not fully data dependent. In contrast, DOSK makes use of the training data, and automatically selects the most important observations and variables through sparse regularization terms. This can be a potential reason that DOSK can perform well in numerical studies.

In contrast to the random projection in the first approach suggested by Liu and Cheng, their second approach considers data-dependent projection matrices. We find this second approach to be very appealing. It is analogous to the prescreening step in variable selection before applying sparse regularization (Fan and Lv [3]). By performing the proposed data reduction using sampling techniques (Ma *et al.* [9]), the corresponding computation can be much more efficient as shown by their numerical comparisons. One caution we would like to make is the potential danger of losing some useful data points by using two stages of data extraction. It will be interesting to establish some theoretical guarantee for the dimension reduction step of the kernel matrix.

Yuan Huang and Shuangge Ma discuss the concept of data extraction and types of problems it may benefit. They focus on the setting of high dimensional data with possibly very small sample sizes, commonly seen in bioinformatics. They believe that with very limited data, it may not be a good idea to remove part of data as the data source is very scarce. We would like to clarify that although DOSK performs data extraction, the observations being removed still contribute to the final model. A simple analogy is the calculation of mean versus median for us to estimate the center of a distribution. Although the mean uses all data points while the median only uses the middle one or two numbers for calculation, the mean is not necessarily always better than median. Furthermore, although the final calculation of the median only involves the middle one or two numbers, the other observations help to decide which observations are the middle ones. This relates to the issue of efficiency and robustness in estimation (Huber and Ronchetti [5]). Similarly, DOSK automatically identifies the subset of observations for the final model and all observations contribute to the model. Another interesting analogy is the Support Vector Machine (SVM). By design of the SVM algorithm, the final classifier will only use the set of support vectors for calculation. The non-support vector observations contribute to the classifier in helping identify the support vectors. Despite the potential usage of a small subset of observations for the final SVM classifier, SVM is one of the most competitive classifiers in practice (Schölkopf and Smola [10], Blanchard *et al.* [2]). Similarly, the data extraction property of DOSK helps to provide a more accurate prediction.

Huang and Ma suggest that one of “the most appropriate scenario for the proposed approach may be the one with a large sample size and ultrahigh-dimensional covariates”. We agree that this can be a very promising application field in the era of big data. One potential direction is to use the “divide and conquer” idea as briefly mentioned in the paper. In particular, one can split the predictors and observations into multiple chunks, and learn on each part with double sparsity to find a representative subset. Then we can combine the selected predictors and observations to train a global model. Such an approach can be computed efficiently. Some recent developments along this line (Zhang *et al.* [15]) can provide useful insights for these potential extensions.

Since data extraction may not be beneficial for all applications, Huang and Ma suggest developing some “diagnostic tools” to decide whether data extraction is necessary.

*Corresponding author.

Indeed, such diagnostic tools can potentially be very useful. Based on our empirical observations, we find that the amount of data extraction depends on the complexity of the underlying function. In general, the more complicated the nonlinear function, the more data observations needed for model construction. Such a phenomenon exists for SVM models. In particular, nonlinear SVM boundaries often use more support vectors than simple linear boundaries (Wu and Liu [12]). One possible diagnose tool is to examine the complexity of the underlying function to decide how much data extraction is needed. This is also related to tuning parameter selection. The current DOSK controls variable selection and data extraction by choosing appropriate tuning parameters. With three tuning parameters, the computation cost can be large as mentioned by Huang and Ma. Our current strategy is somewhat ad hoc. One potential improvement is to increase the computational efficiency using more recent optimization tools. For example, the idea of FISTA (Fast Iterative Shrinkage-Thresholding Algorithm, Beck and Teboulle [1]) may be introduced to increase the convergence rate of our algorithm.

Hao Helen Zhang points out very interesting connections between our proposed data extraction using regularization with the literature on parsimonious knot selection. A similar connection was also briefly mentioned by Zhang *et al.* [13]. Interestingly, although smoothing spline solutions involve all data points as knots in the kernel representation of the function, the effective dimension of the model space is shown to be much less than n due to the use of regularization (Kim and Gu [6]). Knot selection is a traditional problem in nonparametric regression (Wahba [11]). This connection can help to further justify the usefulness of our proposed data extraction.

In terms of tuning parameter selection and computation involved in DOSK, Zhang points out the possibility of developing a solution path algorithm for DOSK. The existing literature on path algorithms such as Li *et al.* [7] can be helpful. Zhang also mentions about the possible extension of interaction selection on the KNIFE related formulation in DOSK. In contrast to the additive form used in other methods such as COSSO (Lin and Zhang [8]), the function representation in DOSK is more flexible. However, how to effectively identify interactions may require more detailed structure identification (Zhang *et al.* [14], Hao and Zhang [4]). More exploration is needed along this direction. Zhang also points out the connection and differences of the convergence rate in the paper with existing parametric and nonparametric rates in the literature. In general, these results heavily depend on the assumptions. As pointed out by Liu and Cheng, the current assumptions used can be strong. It will be interesting to systematically examine the theoretical results and compare them with other existing ones in terms of both assumptions and resulting rates.

Finally, our sincere thanks go to the editor and discussants for their helpful comments and inspiring remarks. These discussions will help lead fruitful research related to DOSK, in particular, simultaneous variable selection and data extraction in kernel learning.

Received 16 June 2018

REFERENCES

[1] Beck, A. and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**(1), 183–202. MR2486527

[2] Blanchard, G., Bousquet, O., and Massart, P. (2008). Statistical Performance of Support Vector Machines. *Annals of Statistics*, **36**(2), 489–531. MR2396805

[3] Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society: Series B*, **70**(5), 849–911. MR2530322

[4] Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, **109**(507), 1285–1301. MR3265697

[5] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley. MR2488795

[6] Kim, Y. and Gu, C. (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(2), 337–356. MR2062380

[7] Li, Y., Liu, Y., and Zhu, J. (2014). Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, **102**(477), 255–268. MR2293307

[8] Lin, Y. and Zhang, H. H. (2006). Component Selection and Smoothing in Multivariate Nonparametric Regression. *The Annals of Statistics*, **34**(5), 2272–2297. MR2291500

[9] Ma, P., Mahoney, M., and Yu, B. (2014). A statistical perspective on algorithmic leveraging. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 91–99, Bejing, China. PMLR.

[10] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press.

[11] Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics. MR1045442

[12] Wu, Y. and Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, **102**(479), 974–983. MR2411659

[13] Zhang, C., Liu, Y., and Wu, Y. (2015). On Quantile Regression in Reproducing Kernel Hilbert Spaces with Data Sparsity Constraint. *Journal of Machine Learning Research*. In press. MR3491134

[14] Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models. *Journal of the American Statistical Association*, **106**(495), 1099–1112. MR2894767

[15] Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, **16**, 3299–3340. MR3450540

Jingxiang Chen
Department of Biostatistics
University of North Carolina at Chapel Hill
USA
E-mail address: jgxchen@email.unc.edu

Chong Zhang
Department of Statistics and Actuarial Science
University of Waterloo
USA
E-mail address: zhangchong101@gmail.com

Michael R. Kosorok
Department of Biostatistics
University of North Carolina at Chapel Hill
USA
E-mail address: kosorok@bios.unc.edu

Yufeng Liu
Department of Statistics and Operations Research
Departments of Genetics and Biostatistics
Carolina Center for Genome Sciences
Lineberger Comprehensive Cancer Center
University of North Carolina at Chapel Hill
USA
E-mail address: yfliu@email.unc.edu