Statistics and Its Interface Volume 11 (2018) 423–424

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

Meimei Liu* and Guang Cheng†,‡

DOSK proposed in [2] aims to perform both variable selection and data extraction at the same time under the "finite sparsity" assumption. In this short note, we propose two alternative approaches based on random projection and importance sampling without such an assumption. Furthermore, we compare these two methods with DOSK empirically in terms of statistical accuracy and computing efficiency.

Keywords and phrases: Data extraction, Importance sampling, Kernel regression, Reproducing kernel Hilbert space, Random projection, Variable selection.

We congratulate the authors on an inspiring piece of work. The authors propose a compelling double sparsity constraint to perform variable selection and data extraction at the same time. An interesting aspect of their method is to automatically reduce the dimension of the kernel matrix by penalizing the related coefficients. On the other hand, the theoretical validity of the proposed DOSK method relies on the "finite sparsity" assumption of the true function, i.e., Assumption 4. This discussion note presents two methods not relying on such an assumption, and compares them with DOSK in terms of statistical accuracy and computing time.

One dimension reduction method is through random projection; see [5], [8] and references therein. Random projection is a probabilistic data compression technique that projects the original dataset to a smaller surrogate dataset in a random manner. By introducing a random matrix $R \in \mathbb{R}^{n\times s}$, we approximate the coefficients of kernel matrix $\alpha \in \mathbb{R}^n$ by $R\tilde{\alpha}$, where $\tilde{\alpha} \in \mathbb{R}^s$. Specifically, equation (6) in [2] can be written as

$$(1)\quad \min_{\tilde{\alpha},b,\omega} \frac{1}{n}||\mathbf{y} - K_{\boldsymbol{\omega}}R\tilde{\alpha} - b||_2^2 + \lambda_2||\boldsymbol{\omega}||_1 + \lambda_3\tilde{\alpha}^T R^T K_{\boldsymbol{\omega}} R\tilde{\alpha}.$$

Commonly used random matrices include those with independent sub-Gaussian entries and randomized orthogonal system projections. Another way is to subsample s knots from the span$\{K_{\boldsymbol{\omega}}(\mathbf{x}_1,\cdot),\cdots,K_{\boldsymbol{\omega}}(\mathbf{x}_n,\cdot)\}$, where $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^p$; that is, R has columns of the form $r_i = \sqrt{\frac{n}{s}}p_i$. A representative example is uniform subsampling, where $\{p_1,\cdots,p_s\}$ are drawn uniformly at random without replacement from the n dimensional identity matrix, e.g., [4], [7].

In the above method, R is chosen as data-independent. Rather, we can choose to construct R in a data-dependent way, such as "importance sampling". Examples include sampling via the statistical leverage score ([3]), the $\lambda-$ridge leverage score ([1]), or adaptive basis selection ([6]). To be more concrete, we take the $\lambda-$ridge leverage score[1] as an example. In this case, equation (6) in [2] can be written as

$$(2)\quad \min_{\alpha,b,\omega} \frac{1}{n}||\mathbf{y} - \widetilde{K}_{\boldsymbol{\omega}}\alpha - b||_2^2 + \lambda_2||\boldsymbol{\omega}||_1 + \lambda_3\alpha^T\widetilde{K}_{\boldsymbol{\omega}}\alpha,$$

where $\widetilde{K}_{\boldsymbol{\omega}}$[2] is the Nyström approximation of the kernel matrix using the $\lambda-$ridge leverage score sampling; see [1] for details. Then, to obtain the solution in equation (2), we only need to simplify the α step in Algorithms 1 and 2 in [2], and update $\alpha^{(t)}$ by $\alpha^{(t)} = (\frac{1}{n}\widetilde{K}_w + \lambda_3 I_n)^{-1}y$.

In the end, we empirically compare the DOSK, random projection using Gaussian matrix (GP) and uniform subsampling (US) in equation (1), and $\lambda-$leverage score (LS) in equation (2), in terms of mean prediction error (MPE) and computing time. Data were generated following Regression Example 1 in [2] with $p_0 = 2$. In Figure 1 (a) and (b), we set $f_0(x_{i1}) = 10\sin(x_{i1})I_{(0<x_{i1}<2\pi)}$; in Figure 1 (c) and (d), the true function was generated with a finite sparse structure such as $f_0(x) = \sum_{i=1}^5 \gamma_i K(x_i,x)$, where $\gamma_1 = \cdots = \gamma_5 = 1/5$. Note that the former function does not satisfy the "sparsity" condition, while the latter does. In the former case, the MPE of DOSK is still comparable to the other three methods but with a large computational cost, as shown in Figure 1 (a) – (b). In the latter case, it is

*PhD student, Department of Statistics, Purdue University.
†Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: chengg@purdue.edu. Research Sponsored by NSF DMS-1712907 and Office of Naval Research (ONR N00014-15-1-2331).
‡Corresponding author.

[1] The $\lambda-$ridge leverage score is defined as $l_i(\lambda_3) = diag(K_{\boldsymbol{\omega}}(K_{\boldsymbol{\omega}} + n\lambda_3 I_n)^{-1})_i$ for $i \in \{1,\cdots,n\}$.
[2] $\widetilde{K}_{\boldsymbol{\omega}}$ is constructed by choosing s columns randomly according to a probability distribution $(p_i)_{1\le i\le n}$ defined by the $\lambda-$ridge leverage score.

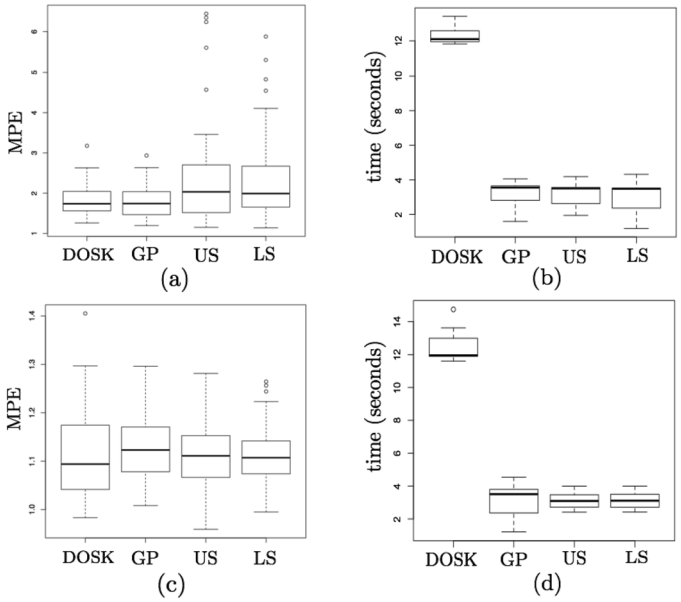

Figure 1. Consider the size of the training and testing data sets to be 100 and 1000, respectively. We set the projection dimension of GP and US to be 20, the number of basis in LS to be 20, and replicated the experiments 50 times.

not surprising that DOSK performs better than the other three methods but still with a large computational cost, as shown in Figure 1 (c) – (d). The computational burden of DOSK comes from cubic running time in n and an additional tuning step related to λ_1. A general direction of interest is to study an optimal trade-off between statistical accuracy and computation efficiency in these procedures.

Received 12 March 2018

REFERENCES

[1] Alaoui, A. E. and Mahoney, M. W., 2015. Fast randomized kernel methods with statistical guarantees. In Advances in neural information processing systems, pp. 775–783.

[2] Chen, J., Zhang, C., Kosorok, M. R. and Liu, Y., 2018, Double sparsity kernel learning with automatic variable selection and data extraction. Statistics and its inference.

[3] Gittens, A. and Mahoney, M. W., 2016. Revisiting the Nyström method for improved large-scale machine learning. The Journal of Machine Learning Research, 17(1), pp. 3977–4041. MR3543523

[4] Kim, Y. J. 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), pp. 337–356. MR2062380

[5] Liu, M., Shang, Z. and Cheng, G., 2018. Nonparametric Testing under Random Projection. arXiv preprint arXiv:1802.06308.

[6] Ma, P., Zhang, N., Huang, J. Z. and Zhong, W., 2017. Adaptive Basis Selection for Exponential Family Smoothing Splines with Application in Joint Modeling of Multiple Sequencing Samples. Statistica Sinica, 27, pp. 1757–1777. MR3726764

[7] Williams, C. K. and Seeger, M., 2001. Using the Nyström method to speed up kernel machines. In Advances in neural information processing systems, pp. 682–688.

[8] Yang, Y., Pilanci, M. and Wainwright, M. J., 2017. Randomized sketches for kernels: Fast and optimal nonparametric regression. The Annals of Statistics, 45(3), pp. 991–1023. MR3662446

Meimei Liu
Department of Statistics
Purdue University
West Lafayette, IN 47906
USA
E-mail address: liu1197@purdue.edu

Guang Cheng
Department of Statistics
Purdue University
West Lafayette, IN 47906
USA
E-mail address: chengg@purdue.edu