Generalized Kernel Regularized Least Squares (Addendum)

Qing Chang[*] and Max Goplerud[†]

October 3, 2023

This document contains additional information on updates to the `gKRLS` package after the publication of Chang and Goplerud (2023).

1 Calibrate Kernel Bandwidth

In Hainmueller and Hazlett (2014) and Chang and Goplerud (2023), the kernel $\boldsymbol{K}$ is defined as follows for kernel bandwidth P:

$$\boldsymbol{K}_{ij} = \exp\left(\frac{-||\boldsymbol{w}_i - \boldsymbol{w}_j||_2^2}{P}\right). \tag{1}$$

P is chosen after the standardization of the covariates to be the dimensionality of $\boldsymbol{w}_i$ (or the rank in the case of a non-full rank design). While this works well, a more data-driven choice is proposed by Hartman, Hazlett and Sterbenz (2021). They suggest that a good choice of P results in meaningful variability in the scaled distances between observations, i.e. $\boldsymbol{K}_{ij}$, where very large values of P make all $\boldsymbol{K}_{ij}$ approximately one and very small values make all non-diagonal elements approximately zero (Hartman, Hazlett and Sterbenz, 2021, p. 21). For a simple and reliable procedure of achieving this, they suggest maximizing the variance of $\text{vec}(\boldsymbol{K})$ over P, ignoring the diagonal. For reasons noted momentarily, we slightly adapt their proposal and do not ignore the diagonal. Their approach can be formalized below, where $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the Hadamard product:

$$P^* = \underset{P}{\text{argmax}} \left\{ \frac{1}{N^2} \mathbf{1}^T \left(\boldsymbol{K} \odot \boldsymbol{K}\right) \mathbf{1} - \left[\frac{1}{N^2}\mathbf{1}^T \boldsymbol{K} \mathbf{1}\right]^2 \right\}. \tag{2}$$

In the case of a small N, this can be evaluated efficiently using numerical optimization. However, for large N, Chang and Goplerud (2023) suggest using random sketching for

[*]PhD Candidate, Department of Political Science, University of Pittsburgh. Email: qic47@pitt.edu. Website: qingcchang.com

[†]Assistant Professor, Department of Political Science, University of Pittsburgh. Email: mgoplerud@pitt.edu. Website: mgoplerud.com

some matrix $\boldsymbol{S}$ to create a smaller problem with design $\boldsymbol{K}\boldsymbol{S}^T$. In this case, Hartman, Hazlett and Sterbenz (2021) consider a variant of sub-sampling sketching and maximize the variance of $\text{vec}(\boldsymbol{K}\boldsymbol{S}^T)$. Our initial experiments found that this sometimes resulted in degenerated estimates, so we propose the following adaptation. In Yang, Pilanci and Wainwright (2017), they note (p. 1013) that given some sketching matrix $\boldsymbol{S}$, the true kernel $\boldsymbol{K}$ is approximated by the following matrix:

$$\tilde{\boldsymbol{K}} = \boldsymbol{K}\boldsymbol{S}^T \left(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T\right)^{-1} \boldsymbol{S}\boldsymbol{K} \tag{3}$$

Thus, we propose the following optimization problem for calibrating P when using random sketching. We do not ignore the diagonal of $\tilde{\boldsymbol{K}}$ as it need not be unity for a particular $\boldsymbol{S}$.

$$P^* = \underset{P}{\text{argmax}} \left\{ \frac{1}{N^2}\mathbf{1}^T \left(\tilde{\boldsymbol{K}} \odot \tilde{\boldsymbol{K}}\right)\mathbf{1} - \left[\frac{1}{N^2}\mathbf{1}^T\tilde{\boldsymbol{K}}\mathbf{1}\right]^2 \right\}. \tag{4}$$

Naive computation of this, however, runs into the same problems of evaluating and storing an N^2 matrix. Fortunately, the objective can be simplified as follows. Define $\boldsymbol{A} = \boldsymbol{K}\boldsymbol{S}^T\left(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T\right)^{-1/2}$ and note that $\tilde{\boldsymbol{K}} = \boldsymbol{A}\boldsymbol{A}^T$.

$$P^* = \underset{P}{\text{argmax}} \left\{ \frac{1}{N^2}\mathbf{1}^T \left[\boldsymbol{A}\left[\boldsymbol{A}^T\boldsymbol{A}\right] \odot \boldsymbol{A}\right]\mathbf{1} - \left[\frac{1}{N^2}\bar{\boldsymbol{a}}^T\bar{\boldsymbol{a}}\right]^2 \right\}; \quad \bar{\boldsymbol{a}} = \boldsymbol{A}^T\mathbf{1} \tag{5}$$

This reformulation requires the creation of no matrix larger than $N \times M$ and no matrix multiplication larger than an $N \times M$ times a $M \times M$ matrix. Costs could be reduced with parallelization if desired.

References

Chang, Qing and Max Goplerud. 2023. "Generalized Kernel Regularized Least Squares." *Political Analysis* pp. 1–15.
URL: *https://doi.org/10.1017/pan.2023.27*

Hainmueller, Jens and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Hartman, Erin, Chad Hazlett and Ciara Sterbenz. 2021. "kpop: A Kernel Balancing Approach for Reducing Specification Assumptions in Survey Weighting." *arxiv preprint* .
URL: *https://arxiv.org/abs/2107.08075*

Yang, Yun, Mert Pilanci and Martin J Wainwright. 2017. "Randomized Sketches for Kernels: Fast and Optimal Nonparametric Regression." *The Annals of Statistics* 45(3):991–1023.