MR4764692 62G08 62R07

Dai, Chi-Shian (1-WI-S)**; Shao, Jun** (1-WI-S)

Kernel regression utilizing external information as constraints. (English. English summary)

Statist. Sinica **34** (2024), *no.* 3, 1675–1697.

This article is another work by the authors on kernel regression estimation when two cross-sectional datasets—namely, the primary (or internal) and secondary (or external) ones—are combined. The internal dataset typically consists of responses to detailed scientific questions. While this dataset may include many additional covariates relative to the external one, the cost of data collection is expensive. The external dataset is less expensive to collect, and as a result its sample size is typically larger than that of the internal one. However, this dataset may only contain crude information such as summary statistics.

More concretely, suppose that the internal dataset includes $(Y, \mathbf{U})$, where $Y \in \mathbb{R}$ is a response and $\mathbf{U} \in \mathbb{R}^p$ is a vector of covariates. Also suppose that the external dataset contains $(Y, \mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^q$ is a sub-vector of $\mathbf{U}$ with $q < p$. This setup is the same as the one in Section 3 of [C.-S. Dai and J. Shao, Stat. Theory Relat. Fields **8** (2024), no. 1, 51–68; MR4718776]. Although both this article and [op. cit.] investigate how external information can improve efficiency in nonparametric kernel estimation for the conditional expectation $\mu(\mathbf{u}) = \mathbb{E}(Y|\mathbf{U} = \mathbf{u})$, there is a crucial difference. This article in principle considers the case in which two datasets belong to the same population. In contrast, [op. cit.] dealt with the case of heterogenous populations, i.e., the case in which two datasets are drawn from different populations, arising from difference in collection methods and/or time periods. Therefore, [op. cit.] can be viewed as a natural extension of this article; indeed, the research extension in this direction is discussed in Section 2.5 of this article.

This article focuses on the local constant or Nadaraya-Watson (NW) regression estimation combining n internal observations $\{(Y_i, \mathbf{U}_i)\}_{i=1}^n$ and m external observations $\{(Y_j, \mathbf{X}_j)\}_{j=1}^m$. The proposed estimation method, called the constrained kernel (CK) regression method, takes two steps. In the first step, the fitted values $\{\widehat{\mu}_i\}_{i=1}^n = \{\widehat{\mu}(\mathbf{U}_i)\}_{i=1}^n$ can be obtained through a constrained optimization under the constraints implied by summary information from the external dataset. In the second step, the CK estimator at a given design point $\mathbf{U} = \mathbf{u}$ can be computed as the NW estimator with responses $\{Y_i\}_{i=1}^n$ replaced by their fitted values $\{\widehat{\mu}_i\}_{i=1}^n$.

It is demonstrated that under some regularity conditions, the asymptotic mean integrated squared error (AMISE) of the CK estimator is smaller than that of the NW estimator with no external information. Methods of choosing the bandwidth and constructing confidence intervals are also proposed as practical considerations. Simulation results confirm improvement in the AMISE of the CK estimator. *Masayuki Hirukawa*

[References]

1. Bierens, H. J. (1987). Kernel estimators of regression functions. In *Advances in Econometrics: Fifth World Congress Vol. 1* (Edited by T. F. Bewley), 99–144. Cambridge University Press MR1117039
2. Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **59**, 447–461. MR1440590
3. Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**,

107–117. MR3494641

4. Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**, 449–460. MR1772408
5. Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86**, 328–332. MR1137117
6. Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382. MR1173804
7. Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. 2nd Edition. CRC Press. MR1680784
8. Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics* **49**, 79–99. MR1450693
9. Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20**, 2008–2036. MR1193323
10. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Routledge. MR1383587
11. Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390
12. Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics* **20**, 675–694. MR1165587
13. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M. et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035.
14. Kim, H. J., Wang, Z. and Kim, J. K. (2021). Survey data integration for regression analysis using model calibration. *arXiv 2107.06448*.
15. Lawless, J., Kalbfleisch, J. and Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**, 413–438. MR1680310
16. Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008. MR2354409
17. Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327. MR1137117
18. Liu, D., Görges, M. and Jenkins, S. A. (2012). University of queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research. *Anesthesia & Analgesia* **114**, 584–589.
19. Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science* **32**, 293–312. MR3648961
20. Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179. MR2949349
21. Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association* **99**, 1131–1139. MR2109501
22. Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* **73**, 166–179. MR1763322
23. Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102**, 169–180. MR3335103
24. Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* **83**, 242–272. MR4256318

25. Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71. MR1450191
26. Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94**, 285–296. MR2331487
27. Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall, Boca Raton.
28. Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729
29. Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96**, 185–193. MR1952731
30. Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 363–410. MR1924297
31. Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science* **3**, 625–650. MR4181993
32. Zhang, Y., Ouyang, Z. and Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics* **11**, 161–184. MR3634319
33. Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association* **85**, 986–1001.

Note: This list reflects references listed in the original paper as accurately as possible with no attempt to correct errors.