MR4664703 62G08 62G05 62J07

Liu, Zejian (1-RICE-S)**; Li, Meng** (1-RICE-S)

On the estimation of derivatives using plug-in kernel ridge regression estimators. (English. English summary)

J. Mach. Learn. Res. **24** (2023), *Paper No.* [266], 37 *pp.*

Consider a nonparametric regression model $y = f_0(\mathbf{x}) + \epsilon$, where $(y, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$, the regression function f_0 is left unspecified, and the regression error ϵ is sub-Gaussian with mean zero. In this article, the problem of estimating derivatives of f_0, denoted as $\partial^\beta f_0$ with $\partial^\beta = \partial_{x_1}^{\beta_1} \cdots \partial_{x_d}^{\beta_d}$ for a multi-index $\beta = (\beta_1, \ldots, \beta_d)$, is studied.

As an estimator of $\partial^\beta f_0$, a plug-in kernel ridge regression (KRR) estimator is proposed. The KRR estimator has a closed-form expression that depends on a Mercer kernel $K(\cdot, \cdot)$, i.e., a continuous, symmetric, and positive definite bivariate function, and a regularization parameter $\lambda > 0$.

Statistical properties of the KRR estimator are explored in a non-asymptotic framework. First, two error bounds for the KRR estimator are derived under the L_∞ norm, where one is built on Mercer kernels with uniformly bounded eigenfunctions and the other on general Mercer kernels. Second, when f_0 belongs to a Hölder or Sobolev class and a kernel K with polynomially decaying eigenvalues is employed, the KRR estimator is shown to be near minimax optimal up to a logarithmic factor under the L_2 norm. An interesting feature in KRR is the *plug-in property*, i.e., the same choice of λ can attain the minimax optimal rate, regardless of the order of the derivative to be estimated. Other derivative estimators, such as local polynomial regression and smoothing spline estimators, do not share this property.

Choices of the kernel K and the regularization parameter λ are discussed, and the computational complexity of KRR, including the time to find the optimal λ, is also mentioned. Finally, two Monte Carlo simulations are conducted. One of these is intended to confirm the minimax optimal rate numerically; this type of design is seldom seen in the related literature. Monte Carlo results indicate that the KRR estimator using the Matérn kernel outperforms other existing methods and tends to achieve the theoretical minimax rate in finite samples. *Masayuki Hirukawa*

[References]

1. Arash A Amini and Martin J Wainwright. Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *The Annals of Statistics*, 40(5):2483–2510, 2012. MR3097610
2. Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209. PMLR, 2013.
3. Sudipto Banerjee, Alan E Gelfand, and CF Sirmans. Directional rates of change under spatial process models. *Journal of the American Statistical Association*, 98(464):946–954, 2003. MR2041483
4. Peter J Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003. MR2001641
5. Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Matematicheskii Sbornik*, 115(3):331–355, 1967. MR0217487
6. Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.
7. Jorge Luis Ojeda Cabrera. *locpol: Kernel local polynomial regression*, 2018. URL https://CRAN.R-project.org/package=locpol. R package version 0.7-0.

8. Niamh Cahill, Andrew C Kemp, Benjamin P Horton, and Andrew C Parnell. Modeling sea-level change using errors-in-variables integrated Gaussian processes. *The Annals of Applied Statistics*, 9(2):547–571, 2015. MR3371325
9. Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. MR2335249
10. Richard Charnigo, Benjamin Hall, and Cidambi Srinivasan. A generalized C_p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011. MR2857702
11. Jie Chen, Lingfei Wu, Kartik Audhkhasi, Brian Kingsbury, and Bhuvana Ramabhadrari. Efficient one-vs-one kernel ridge regression for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2454–2458, 2016.
12. Yu Cheng, Zhigang Jin, Tao Gao, Hongcai Chen, and Nikola Kasabov. An improved collaborative representation based classification with regularized least square (CRC–RLS) method for robust face recognition. *Neurocomputing*, 215:250–259, 2016.
13. Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: On the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002. MR1930945
14. Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007. MR2354721
15. Kevin Stephen Stotter Cuddy. Convergence of Fourier series. 2012. URL http://math.uchicago.edu/~may/REU2012/REUPapers/Cuddy.pdf.
16. Wenlin Dai, Tiejun Tong, and Marc G Genton. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(1):5700–5724, 2016. MR3555052
17. Xiongtao Dai, Hans-Georg Müller, and Wenwen Tao. Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statistica Sinica*, 28(3):1583–1609, 2018. MR3821019
18. Kris De Brabanter, Jos De Brabanter, Bart De Moor, and Irene Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013. MR3033332
19. M Delecroix and AC Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382, 1996. MR1386346
20. Peter Exterkate, Patrick JF Groenen, Christiaan Heij, and Dick van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, 32(3):736–753, 2016. MR3137689
21. Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press, 1996. MR1383587
22. José C. Ferreira and Valdir A. Menegatto. Reproducing properties of differentiable Mercerlike kernels. *Mathematische Nachrichten*, 285(8-9):959–973, 2012. ISSN 0025584X. doi: 10.1002/mana.201100072. MR2928393
23. Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020. MR4209491
24. Mark S Gockenbach. *Partial Differential Equations: Analytical and Numerical Methods*, volume 122. Siam, 2005. MR2743564
25. Chong Gu. *Smoothing Spline ANOVA Models*, volume 297. Springer Science & Business Media, 2013. MR3025869
26. Zheng-Chu Guo and Ding-Xuan Zhou. Concentration estimates for learning with un-

bounded sampling. *Advances in Computational Mathematics*, 38(1):207–223, 2013. MR3011339

27. László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006. MR1920390
28. Tracy Holsclaw, Bruno Sansó, Herbert KH Lee, Katrin Heitmann, Salman Habib, David Higdon, and Ujjaini Alam. Gaussian process modeling of derivative curves. *Technometrics*, 55(1):57–67, 2013. MR3038485
29. Wolfgang Härdle. *Applied Nonparametric Regression*. Number 19 in Econometric Society Monographs. Cambridge University Press, 1990. MR1161622
30. Meng Li and Subhashis Ghosal. Bayesian detection of image boundaries. *The Annals of Statistics*, 45(5):2190–2217, 2017. MR3718166
31. Meng Li, Zejian Liu, Cheng-Han Yu, and Marina Vannucci. Semiparametric Bayesian inference for local extrema of functions in the presence of noise. *arXiv preprint arXiv:2103.10606*, 2021.
32. Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020. MR4068943
33. Yu Liu and Kris De Brabanter. Derivative estimation in random design. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3449–3458, 2018.
34. Yu Liu and Kris De Brabanter. Smoothed nonparametric derivative estimation using weighted difference quotients. *Journal of Machine Learning Research*, 21(65):1–45, 2020. MR4095344
35. Zejian Liu and Meng Li. Equivalence of convergence rates of posterior distributions and Bayes estimators for functions and nonparametric functionals. *arXiv preprint arXiv:2011.13967*, 2020.
36. Zejian Liu and Meng Li. Optimal plug-in Gaussian processes for modelling derivatives. *arXiv preprint arXiv:2210.11626*, 2022.
37. Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through selfconcordance. In *Conference on Learning Theory*, pages 2294–2340. PMLR, 2019.
38. Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010. MR2590050
39. P Mohapatra, Sreejit Chakravarty, and PK Dash. Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm and Evolutionary Computation*, 28:144–160, 2016.
40. Hans-Georg Müller, Ulrich Stadtmüller, and Thomas Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987. MR0919842
41. Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Process for Machine Learning*. The MIT Press, 2006. MR2514435
42. Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 645–652, 2010.
43. Brian Ripley. *pspline: Penalized smoothing splines*, 2017. URL https://CRAN.R-project.org/package=pspline. R package version 1.0-18.
44. Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14(1):1665–1714, 2013. MR3104492
45. Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian

concentration. *Electronic Communications in Probability*, 18, 2013. MR3125258
46. Steve Smale and Ding-Xuan Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005. MR2186447
47. Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007. MR2327597
48. Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl E Rasmussen. Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems*, pages 1057–1064, 2003.
49. Peter Sollich and Christopher Williams. Using the equivalent kernel to understand Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 1313–1320, 2005.
50. Peter X-K Song, Xin Gao, Rui Liu, and Wen Le. Nonparametric inference for local extrema with application to oligonucleotide microarray data in yeast genome. *Biometrics*, 62(2):545–554, 2006. MR2236836
51. Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 1999. MR1697409
52. Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008. MR2450103
53. Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.
54. Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. MR0673642
55. Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985. MR0790566
56. Grace Wahba. *Spline Models for Observational Data*, volume 59. Siam, 1990. MR1045442
57. Grace Wahba and Yonghua Wang. When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics-Theory and Methods*, 19 (5):1685–1700, 1990. MR1075497
58. Cheng Wang and Ding-Xuan Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011. MR2745300
59. Wen Wu Wang and Lu Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16(1):2617–2641, 2015. MR3450519
60. Wen Wu Wang, Ping Yu, Lu Lin, and Tiejun Tong. Robust estimation of derivatives using locally weighted least absolute deviation regression. *Journal of Machine Learning Research*, 20(1):2157–2205, 2019. MR3960914
61. Xiaojing Wang and James O Berger. Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25, 2016. MR3452261
62. Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006. MR2172729
63. Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.
64. Yannis G Yatracos. On the estimation of the derivatives of a function with the

derivatives of an estimate. *Journal of Multivariate Analysis*, 28(1):172–175, 1989. MR0996989

65. Yannis G Yatracos. Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture. *Statistics*, 53(6): 1251–1268, 2019. MR4034861
66. Cheng-Han Yu, Meng Li, Colin Noe, Simon Fischer-Baum, and Marina Vannucci. Bayesian inference for stationary points in Gaussian process regression models for event-related potentials analysis. *Biometrics*, 79(2):629–641, 2023. MR4606303
67. Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005. MR2175849
68. Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015. MR3450540
69. Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002. MR1928805
70. Shanggang Zhou and Douglas A Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10:93–108, 2000. MR1742102

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