Sherwood, Ben; **Wang, Lan**

Partially linear additive quantile regression in ultra-high dimension. (English) Zbl 1331.62264

Ann. Stat. 44, No. 1, 288-317 (2016).

Summary: We consider a flexible semiparametric quantile regression model for analyzing high dimensional heterogeneous data. This model has several appealing features: (1) By considering different conditional quantiles, we may obtain a more complete picture of the conditional distribution of a response variable given high dimensional covariates. (2) The sparsity level is allowed to be different at different quantile levels. (3) The partially linear additive structure accommodates nonlinearity and circumvents the curse of dimensionality. (4) It is naturally robust to heavy-tailed distributions. In this paper, we approximate the nonlinear components using B-spline basis functions. We first study estimation under this model when the nonzero components are known in advance and the number of covariates in the linear part diverges. We then investigate a nonconvex penalized estimator for simultaneous variable selection and estimation. We derive its oracle property for a general class of nonconvex penalty functions in the presence of ultra-high dimensional covariates under relaxed conditions. To tackle the challenges of nonsmooth loss function, nonconvex penalty function and the presence of nonlinear components, we combine a recently developed convex-differencing method with modern empirical process techniques. Monte Carlo simulations and an application to a microarray study demonstrate the effectiveness of the proposed method. We also discuss how the method for a single quantile of interest can be extended to simultaneous variable selection and estimation at multiple quantiles.

MSC:

62G35 Nonparametric robustness
62G20 Asymptotic properties of nonparametric inference

Keywords:

quantile regression; high dimensional data; nonconvex penalty; partial linear; variable selection

References:

[1] Bai, Z. D. and Wu, Y. (1994). Limiting behavior of \(M\)-estimators of regression coefficients in high-dimensional linear models. I. Scale-dependent case. J. Multivariate Anal. 51 211-239. · Zbl 0816.62025 · doi:10.1006/jmva.1994.1059

[2] Belloni, A. and Chernozhukov, V. (2011). \(\ell_{1}\)-penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39 82-130. · Zbl 1209.62064 · doi:10.1214/10-AOS827 · arxiv:0904.2931

[3] Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. Ann. Statist. 32 898-927. · Zbl 1092.62045 · doi:10.1214/009053604000000247 · arxiv:math/0406465

[4] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348-1360. · Zbl 1073.62547 · doi:10.1198/016214501753382273

[5] Gilliam, M., Rifas-Shiman, S., Berkey, C., Field, A. and Colditz, G. (2003). Maternal gestational diabetes, birth weight and adolescent obesity. Pediatrics 111 221-226.

[6] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over-parametrization. Bernoulli 10 971-988. · Zbl 1055.62078 · doi:10.3150/bj/1106314846 ·

[7] He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. J. Multivariate Anal. 73 120-135. · Zbl 0948.62013 · doi:10.1006/jmva.1999.1873

[8] He, X. and Shi, P. (1996). Bivariate tensor-product \(B\)-splines in a partly linear model. J. Multivariate Anal. 58 162-181. · Zbl 0865.62027 · doi:10.1006/jmva.1996.0045

[9] He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann. Statist. 41 342-369. · Zbl 1295.62053 · doi:10.1214/13-AOS1087 · euclid:aos/1364302746 · arxiv:1304.2186

[10] He, X., Zhu, Z.-Y. and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika 89 579-590. · Zbl 1036.62035 · doi:10.1093/biomet/89.3.579

[11] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. Statist. Sci. 27 481-499. · Zbl 1331.62347 · doi:10.1214/12-STS392 · euclid:ss/1356098552

[12] Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. Ann. Statist. 38 2282-2313. · Zbl 1202.62051 · doi:10.1214/09-AOS781 · arxiv:1010.4115

[13] Huang, J., Wei, F. and Ma, S. (2012). Semiparametric regression pursuit. Statist. Sinica 22 1403-1426. · Zbl 1253.62024 · www3.stat.sinica.edu.tw

[14] Ishida, M., Monk, D., Duncan, A. J., Abu-Amero, S., Chong, J., Ring, S. M., Pembrey, M. E., Hindmarsh, P. C., Whittaker, J. C., Stanier, P. and Moore, G. E. (2012). Maternal inheritance of a promoter variant in the imprinted PHLDA2 gene significantly increases birth weight. Am. J. Hum. Genet. 90 715-719.

[15] Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. Ann. Statist. 39 305-332. · Zbl 1209.62074 · doi:10.1214/10-AOS842 · arxiv:1103.1525

[16] Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. Ann. Statist. 36 2232-2260. · Zbl 1209.62064 · doi:10.1214/10-AOS827 · arxiv:0904.2931

[17] Lee, E. R., Noh, H. and Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. J. Amer. Statist. Assoc. 109 216-229. · Zbl 1367.62122 · doi:10.1080/01621459.2013.836975

[18] Li, G., Xue, L. and Lian, H. (2011). Semi-varying coefficient models with a diverging number of components. J. Multivariate Anal. 102 1166-1174. · Zbl 1216.62060 · doi:10.1016/j.jmva.2011.03.010

[19] Lian, H., Liang, H. and Ruppert, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. Statist. Sinica 25 591-607. · Zbl 06503812

[20] Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. J. Amer. Statist. Assoc. 104 234-248. · Zbl 1388.62208 · doi:10.1198/jasa.2009.0127

[21] Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. Statist. Sinica 21 1225-1248. · Zbl 1223.62020 · doi:10.5705/ss.2009.140

[22] Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. J. Nonparametr. Stat. 23 415-437. · Zbl 1359.62108 · doi:10.1080/10485252.2010.537336

[23] Schumaker, L. L. (1981). Spline Functions : Basic Theory . Wiley, New York. · Zbl 0449.41004

[24] Sherwood, B. and Wang, L. (2015). Supplement to "Partially linear additive quantile regression in ultra-high dimension." . · Zbl 1331.62264 · doi:10.1214/15-AOS1367 · euclid:aos/1449755964

[25] Stone, C. J. (1985). Additive regression and other nonparametric models. Ann. Statist. 13 689-705. · Zbl 0605.62065 · doi:10.1214/aos/1176349548 ·

[26] Tang, Y., Song, X., Wang, H. J. and Zhu, Z. (2013). Variable selection in high-dimensional quantile varying coefficient models. J. Multivariate Anal. 122 115-132. · Zbl 1279.62049 · doi:10.1016/j.jmva.2013.07.015

[27] Tao, P. D. and An, L. T. H. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. Acta Math. Vietnam. 22 289-355. · Zbl 0895.90152

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58 267-288. · Zbl 0850.62538

[29] Turan, N., Ghalwash, M., Katari, S., Coutifaris, C., Obradovic, Z. and Sapienza, C. (2012). DNA methylation differences at growth related genes correlate with birth weight: A molecular signature linked to developmental origins of adult disease? BMC Medical Genomics 5 10.

[30] Votavova, H., Dostalova Merkerova, M., Fejglova, K., Vasikova, A., Krejcik, Z., Pastorkova, A., Tabashidze, N., Topinka, J., Veleminsky, M., Jr., Sram, R. J. and Brdicka, R. (2011). Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. Placenta 32 763-770.

[31] Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. J. Amer. Statist. Assoc. 107 214-222. · Zbl 1274.62289 · doi:10.1214/07-AOS544

[32] Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. J. Amer. Statist. Assoc. 104 747-757. · Zbl 1388.62213 · doi:10.1198/jasa.2009.0138

[33] Wang, H. J., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. Ann. Statist. 37 3841-3866. · Zbl 1191.62077 · doi:10.1214/09-AOS695 · arxiv:0911.3501

[34] Wang, L., Liu, X., Liang, H. and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. Ann. Statist. 39 1827-1851. · Zbl 1227.62053 · doi:10.1214/11-AOS885 · arxiv:1112.2502

[35] Wei, Y. and He, X. (2006). Conditional growth charts. Ann. Statist. 34 2069-2131. With discussions and a rejoinder by the authors. · Zbl 1106.62049 · doi:10.1214/009053606000000623 · arxiv:math/0702634

[36] Welsh, A. H. (1989). On \(M\)-processes and \(M\)-estimation. Ann. Statist. 17 337-361. · Zbl 0701.62074 · doi:10.1214/aos/1176347021 ·

[37] Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. Ann. Statist. 37 673-696. · Zbl 1162.62037 · doi:10.1214/07-AOS580 · arxiv:0903.5474

[38] Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. Statist. Sinica 16 1423-1446. · Zbl 1109.62030

[39] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B. Stat. Methodol. 68 49-67. · Zbl 1141.62030 · doi:10.1111/j.1467-9868.2005.00532.x

[40] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 894-942. · Zbl 1183.62120 · doi:10.1214/09-AOS729 · arxiv:1002.4734

[41] Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. J. Amer. Statist. Assoc. 106 1099-1112. · Zbl 1229.62051 · doi:10.1198/jasa.2011.tm10281

[42] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36 1509-1533. · Zbl 1142.62027 · doi:10.1214/009053607000000802 · arxiv:0808.1012

[43] Zou, H. and Yuan, M. (2008). Regularized simultaneous model selection in multiple quantiles regression. Comput. Statist. Data Anal. 52 5296-5304. · Zbl 1452.62301