MR4320950 62G05 60G22 62M10
Caron, Emmanuel (F-AVIG-LM); **Dedecker, Jérôme** (F-UPARIS-LAM); **Michel, Bertrand [Michel, Bertrand[1]]** (F-NANTC-LM)
Gaussian linear model selection in a dependent context. (English summary)
Electron. J. Stat. **15** (2021), *no. 2,* 4823–4867.

This article is concerned with a nonparametric fixed-design regression model with a Gaussian dependent error. The model can be expressed as

$$Y_i = f^*\left(\frac{i}{n}\right) + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where f^* is an unknown function having support on $[0,1]$ and ε_i is a stationary Gaussian process. The aim is to estimate f^* using n observations $\{Y_i\}_{i=1}^n$.

The article focuses on fitting the regressogram of order m to f^*, where m is chosen via a model selection procedure. More specifically, the optimal order $\widehat{m}$ is determined by

$$\widehat{m} \in \underset{m \in \{1,\dots,n\}}{\arg\min} \left\{ \left\| Y - \widehat{f}_m \right\|_n^2 + \mathrm{pen}(m) \right\}, \tag{1}$$

where $\|X\|_n = \frac{1}{n}\sum_{i=1}^n X_i^2$, $\widehat{f}_m$ is the regressogram of order m, and $\mathrm{pen}(m)$ is some penalty term depending on m.

It is demonstrated that the form of $\mathrm{pen}(m)$ varies with the dependent structure of ε_i. Details are given below.

- If ε_i is a short-memory process so that its spectral radius (the largest eigenvalue of its long-run variance) is bounded, then the penalty is in the form
$$\mathrm{pen}(m) = K\left(\frac{m}{n}\right)$$
for some $K > 0$.
- If ε_i is a long-memory process so that its kth-order autocovariance satisfies $|E(\varepsilon_i \varepsilon_{i-k})| \le \kappa k^{-\gamma}$ for some $\kappa > 0$ and $\gamma \in (0,1)$, then the penalty becomes
$$\mathrm{pen}(m) = K\left(\frac{m}{n}\right)^{\gamma}$$
for some $K > 0$.
- If ε_i is anti-persistent so that $\mathrm{Var}(S_n) \le \kappa n^{2-\gamma}$ for the partial sum process $S_n = \varepsilon_1 + \cdots + \varepsilon_n$, some $\kappa > 0$ and $\gamma \in (1,2)$, then the penalty is given by
$$\mathrm{pen}(m) = K\left(\left(\frac{m}{n}\right)^{\gamma} + \frac{\log m}{n}\right)$$
for some $K > 0$.

In practice, however, making the order selection via (1) fully operational requires one to prespecify the dependent structure of ε_i and choose the constant K in a data-driven manner. Now, if $\mathrm{Var}(S_n) \sim \kappa n^{2-\gamma}$ for some $\gamma \in (0,2)$, then the leading term of $\mathrm{pen}(m)$ is of order $(m/n)^{\gamma}$. The exponent γ also relates to the Hurst index H of the partial sum process S_n so that $\gamma = 2 - 2H$. In conclusion, it is recommended that the order selection is implemented in the following two-step procedure:

1. Run the dimension jump algorithm [see J.-P. Baudry, C. Maugis-Rabusseau and

B. Michel, Stat. Comput. **22** (2012), no. 2, 455–470; MR2865029] using the penalty shape $(m/n)^{2-2\widehat{H}_1}$ to select the pre-model of order $\widehat{m}_1$, where $\widehat{H}_1$ is the Whittle estimate of the Hurst index using the Y process.

2. Rerun the dimension jump algorithm using the penalty shape $(m/n)^{2-2\widehat{H}_2}$ to select the final model of order $\widehat{m}$, where $\widehat{H}_2$ is the Whittle estimate of the Hurst index using the residual of the pre-model $\widehat{m}_1$.

Masayuki Hirukawa

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. MR0483125 MR0483125
2. Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*. MR4021408 MR4021408
3. Baraud, Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields* **117** 467–493. MR1777129 MR1777129
4. Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics* **6** 127–146. MR1918295 MR1918295
5. Baraud, Y., Comte, F. and Viennet, G. (2001). Adaptive estimation in autoregression or-mixing regression via model selection. *The Annals of Statistics* **29** 839–875. MR1865343 MR1865343
6. Baudry, J.-P., Maugis, C. and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing* **22** 455–470. MR2865029 MR2865029
7. Beran, J. (1994). *Statistics for long-memory processes. Monographs on Statistics and Applied Probability* **61**. Chapman and Hall, New York. MR1304490 MR1304490
8. Beran, J. and Feng, Y. (2002). Local polynomial fitting with long-memory, short-memory and antipersistent errors. *Ann. Inst. Statist. Math.* **54** 291–311. MR1910174 MR1910174
9. Beran, J. and Shumeyko, Y. (2012). On asymptotically optimal wavelet estimation of trend functions under long-range dependence. *Bernoulli* **18** 137–176. MR2888702 MR2888702
10. Birgé, L. and Massart, P. (2001a). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268. MR1848946 MR1848946
11. Birgé, L. and Massart, P. (2001b). A generalized Cp criterion for Gaussian model selection. *Technical report, Universités de Paris 6 et Paris 7*. MR1848946
12. Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields* **138** 33–73. MR2288064 MR2288064
13. Cirel'son, B. S., Ibragimov, I. A. and Sudakov, V. N. (1976). Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)* 20–41. Lecture Notes in Math., Vol. 550. MR0458556 MR0458556
14. Csörgő, S. and Mielniczuk, J. (1995a). Close short-range dependent sums and regression estimation. *Acta Sci. Math. (Szeged)* **60** 177–196. MR1348687 MR1348687
15. Csörgő, S. and Mielniczuk, J. (1995b). Distant long-range dependent sums and regression estimation. *Stochastic Process. Appl.* **59** 143–155. MR1350260 MR1350260
16. Csörgő, S. and Mielniczuk, J. (1995c). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.* **23** 1000–1014. MR1345211 MR1345211
17. Dedecker, J., Dehling, H. and Taqqu, M. S. (2015). Weak convergence of the empirical process of intermittent maps in L^2 under long-range dependence. *Stoch.*

Dyn. **15** 29p. MR3332268 MR3332268
18. Dedecker, J., Gouëzel, S. and Merlevède, F. (2018). Large and moderate deviations for bounded functions of slowly mixing Markov chains. *Stoch. Dyn.* **18** 1850017, 38. MR3735414 MR3735414
19. Dehling, H. and Taqqu, M. S. (1989). The empirical process of some long-range dependent sequences with an application to U-statistics. *Ann. Statist.* **17** 1767–1783. MR1026312 MR1026312
20. DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation* **303**. Springer Science & Business Media. MR1261635 MR1261635
21. Doukhan, P., Massart, P. and Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.* **30** 63–82. MR1262892 MR1262892
22. Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.
23. Gendre, X. (2008). Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electronic Journal of Statistics* **2** 1345–1372. MR2471290 MR2471290
24. Gendre, X. (2014). Model selection and estimation of a component in additive regression. *ESAIM: Probability and Statistics* **18** 77–116. MR3143734 MR3143734
25. Gerschgorin, S. (1931). Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk* **7** 749–754.
26. Giraitis, L., Koul, H. L. and Surgailis, D. (2012). *Large sample inference for long memory processes.* Imperial College Press, London. MR2977317 MR2977317
27. Giraitis, L., Robinson, P. M. and Surgailis, D. (2000). A model for long memory conditional heteroscedasticity. *Ann. Appl. Probab.* **10** 1002–1024. MR1789986 MR1789986
28. Giraud, C. (2014). *Introduction to high-dimensional statistics.* Chapman and Hall/CRC. MR3307991 MR3307991
29. Hall, P. and Hart, J. D. (1990). Nonparametric regression with long-range dependence. *Stochastic Process. Appl.* **36** 339–351. MR1084984 MR1084984
30. Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9** 33–49. MR1678880 MR1678880
31. Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.* **116** 770–799.
32. Johnstone, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9** 51–83. MR1678881 MR1678881
33. Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59** 319–351. MR1440585 MR1440585
34. Lerasle, M. (2011). Optimal model selection for density estimation of stationary data under various mixing conditions. *The Annals of Statistics* **39** 1852–1877. MR2893855 MR2893855
35. Li, L. and Xiao, Y. (2007). On the minimax optimality of block thresh-olded wavelet estimators with long memory data. *J. Statist. Plann. Inference* **137** 2850–2869. MR2323796 MR2323796
36. Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15** 661–675.
37. Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10** 422–437. MR0242239 MR0242239
38. Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin. MR2319879 MR2319879

39. Pesee, C. (2008). Long-range dependence of financial time series data. *WASET* **44** 163–167.
40. Pipiras, V. and Taqqu, M. S. (2017). *Long-range dependence and self-similarity. Cambridge Series in Statistical and Probabilistic Mathematics, [45].* Cambridge University Press, Cambridge. MR3729426 MR3729426
41. Puplinskaitė, D. and Surgailis, D. (2010). Aggregation of a random-coefficient AR(1) process with infinite variance and idiosyncratic innova-tions. *Adv. in Appl. Probab.* **42** 509–527. MR2675114 MR2675114
42. Robinson, P. M. (1991). Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *J. Econometrics* **47** 67–84. MR1087207 MR1087207
43. Robinson, P. M. (1997). Large-sample inference for nonparametric regression with dependent errors. *Ann. Statist.* **25** 2054–2083. MR1474083 MR1474083
44. Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable non-Gaussian random processes. Stochastic Modeling.* Chapman & Hall, New York Stochastic models with infinite variance. MR1280932 MR1280932
45. Stephenson, D. B., Pavan, V. and Bojariu, R. (2000). Is the North Atlantic Oscillation a random walk? *International Journal of Climatology* **20** 1–18.
46. Toussoun, O. (1925). Mémoire sur L'Histoire du Nil. 3 vols. *Cairo, L'Institut Français D'Archéologie Orientale.*
47. Tran, L., Roussas, G., Yakowitz, S. and Truong Van, B. (1996). Fixed-design regression for linear time series. *Ann. Statist.* **24** 975–991. MR1401833 MR1401833
48. Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *The Annals of Statistics* **24** 466–484. MR1394972 MR1394972
49. Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik* **2** 423–434. MR0060797 MR0060797

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