Vilar, José A.; **Vilar, Juan M.**

Time series clustering based on nonparametric multidimensional forecast densities. (English)

Zbl 1336.62197
Electron. J. Stat. 7, 1019-1046 (2013).

Summary: A new time series clustering method based on comparing forecast densities for a sequence of $k > 1$ consecutive horizons is proposed. The unknown k-dimensional forecast densities can be nonparametrically approximated by using bootstrap procedures that mimic the generating processes without parametric restrictions. However, the difficulty of constructing accurate kernel estimators of multivariate densities is well known. To circumvent the high dimensionality problem, the bootstrap prediction vectors are projected onto a lower-dimensional space using principal components analysis, and then the densities are estimated in this new space. Proper distances between pairs of estimated densities are computed and used to generate an initial dissimilarity matrix, and hence a standard hierarchical clustering is performed. The clustering procedure is examined via simulation and is applied to a real dataset involving electricity prices series.

MSC:

62H30 Classification and discrimination; cluster analysis (statistical aspects)
62M20 Inference from stochastic processes and prediction
62M10 Time series, auto-correlation, regression, etc. in statistics (GARCH)
62G07 Density estimation
62G09 Nonparametric statistical resampling methods
62H25 Factor analysis and principal components; correspondence analysis

Keywords:

time series clustering; multidimensional forecast density; bootstrap; kernel estimation; principal components analysis

Software:

R; TRAMO; clusfind

Full Text: DOI Euclid

References:

[1] Alonso, A. M., Berrendero, J. R., Hernandez, A. and Justel, A. (2006). Time series clustering based on forecast densities. Comput. Statist. Data Anal. 51 762-776. · Zbl 1157.62484 · doi:10.1016/j.csda.2006.04.035

[2] Alonso, A. M., Peña, D. and Romo, J. (2002). Forecasting time series with sieve bootstrap. J. Statist. Plann. Inference 100 1-11. · Zbl 1007.62077 · doi:10.1016/S0378-3758(01)00092-1

[3] Boets, J., De Cock, K., Espinoza, M. and De Moor, B. (2005). Clustering time series, subspace identification and cepstral distances. Commun. Inf. Syst. 5 69-96. · Zbl 1089.62103 · doi:10.4310/CIS.2005.v5.n1.a3

[4] Bühlmann, P. (1997). Sieve bootstrap for time series. Bernouilli 3 123-148. · Zbl 0874.62102 · doi:10.2307/3318584

[5] Caiado, J., Crato, N. and Peña, D. (2006). A periodogram-based metric for time series classification. Comput. Statist. Data Anal. 50 2668-2684. · Zbl 1445.62222

[6] Cao, R., Febrero-Bande, M., Gonzalez-Manteiga, W., Prada-Sánchez, J. M. and García-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. Commun. Stat., Simulation Comput. 26 961-978.

[7] Chouakria-Douzal, A. and Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. Adv. Data Anal. Classif. 1 5-21. · Zbl 1131.62078 · doi:10.1007/s11634-006-0004-6

[8] Conejo, A. J., Plazas, M. A., Espínola, R. and B., M. (2005). Day-ahead electricity price forecasting using the wavelet transformand ARIMA models. IEEE Trans. Power Syst. 20 1035-1042.

[9] Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. Comput. Statist. Data Anal. 52 1860-1872. · Zbl 1452.62624

[10] Franke, J., Kreiss, J.-P. and Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. Bernouilli 8 1-37. ·

Zbl 1006.62038

[11] Galeano, P. and Peña, D. (2000). Multivariate analysis in vector time series. Resenhas 4 383-403. · Zbl 1098.62558

[12] García-Martos, C., Rodríguez, J. and Sánchez, M. J. (2007). Mixed models for short-run forecasting of electricity prices: Application for the Spanish market. IEEE Trans. Power Syst. 22 544-552.

[13] Gavrilov, M., Anguelov, D., Indyk, P. and Motwani, R. (2000). Mining the stock market (extended abstract): which measure is best? In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining . KDD'00 487-496. ACM, New York, USA.

[14] Gómez, V. and Maravall, A. (1996). Programs TRAMO (Times Series Regression with ARIMA noise, Missing observations and Outliers) and SEATS (Signal Extraction in ARIMA Time Series). Instructions for the user. Working paper 9628, Bank of Spain, Madrid.

[15] Hart, J. D. (1994). Automated Kernel Smoothing of Dependent Data by Using Time Series Cross- Validation. Journal of the Royal Statistical Society. Series B (Methodological) 56 529-542. · Zbl 0800.62224

[16] Hubert, L. and Arabie, P. (1985). Comparing partitions. J. Classification 2 193-218. · Zbl 0587.62128

[17] Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data . Prentice-Hall, Inc., Upper Saddle River, NJ, USA. · Zbl 0665.62061

[18] Jolliffe, I. T. (2002). Principal component analysis , second ed. Springer Series in Statistics . Springer-Verlag, New York. · Zbl 1011.62064

[19] Kakizawa, Y., Shumway, R. H. and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. J. Amer. Statist. Assoc. 93 328-340. · Zbl 0906.62060 · doi:10.2307/2669629

[20] Kalpakis, K., Gada, D. and Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time-series. In Proceedings 2001 IEEE International Conference on Data Mining (N. Cercone, T. Y. Lin and X. Wu, eds.) 273-280. IEEE Comput. Soc.

[21] Kaufman, L. and Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis . John Wiley and Sons, New York. · Zbl 1345.62009

[22] Kim, C.-I., Yu, I.-K. and Song, Y. H. (2002). Prediction of system marginal price of electricity using wavelet transform analysis. Energy Conv. Manag. 43 1839 - 1851.

[23] Kovačić, Z. J. (1998). Classification of time series with applications to the leading indicator selection. In Data science, classification, and related methods. Proceedings of the fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27-30, 1996 204-207. Springer.

[24] Liao, T. W. (2005). Clustering of time series data: a survey. Pattern Recognition 38 1857-1874. · Zbl 1077.68803 · doi:10.1016/j.patcog.2005.01.025

[25] Maharaj, E. A. (1996). A significance test for classifying ARMA models. J. Statist. Comput. Simulation 54 305-331. · Zbl 0899.62116 · doi:10.1080/00949659608811737

[26] Maharaj, E. A. (2002). Comparison of non-stationary time series in the frequency domain. Comput. Statist. Data Anal. 40 131-141. · Zbl 0990.62078 · doi:10.1016/S0167-9473(01)00100-1

[27] Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. Comm. Statist. A-Theory Methods 7 1233-1241. · Zbl 0403.62033 · doi:10.1080/03610927808827707

[28] Pértega, S. and Vilar, J. A. (2010). Comparing Several Parametric and Nonparametric Approaches to Time Series Clustering: A Simulation Study. J. Classification 27 333-362. · Zbl 1337.62137

[29] Piccolo, D. (1990). A distance measure for classifying ARIMA models. J. Time Series Anal. 11 153-164. · Zbl 0691.62083 · doi:10.1111/j.1467-9892.1990.tb00048.x

[30] Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. J. Amer. Statist. Assoc. 66 846-850.

[31] Samé, A., Chamroukhi, F., Govaert, G. and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. Adv. Data Anal. Classif. 5 301-321. · Zbl 1274.62427

[32] Struzik, Z. R. and Siebes, A. (1999). The Haar wavelet in the time series similarity paradigm. In Principles of Data Mining and Knowledge Discovery. Proceedings of the third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999 12-22. Springer.

[33] R Core Team (2012). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

[34] Vilar, J. A., Alonso, A. M. and Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. Comput. Statist. Data Anal. 54 2850-2865. · Zbl 1284.62575

[35] Vilar, J. A. and Pértega, S. (2004). Discriminant and cluster analysis for Gaussian stationary processes: local linear fitting approach. J. Nonparametr. Stat. 16 443-462. · Zbl 1076.62063 · doi:10.1080/10485250410001656453

[36] Vilar, J. M., Cao, R. and Aneiros, G. (2012). Forecasting next-day electricity demand and price using nonparametric functional methods. Int. J. Electr. Power Energy Syst. 39 48-55.

[37] Vilar, J. M., Vilar, J. A. and Pértega, S. (2009). Classifying Time Series Data: A Nonparametric Approach. J. Classification 26 3-28. · Zbl 1276.62042

[38] Wand, M. P. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. Comput. Statist. 9 97-116. · Zbl 0937.62055

[39] Weron, R. (2006). Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach . HSC Books . Hugo Steinhaus Center, Wroclaw University of Technology.

matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.