New Methods and Theory for the Comparison of Nonparametric Curves

A General information

1 Applicant(s)

PI
date of birth
address
telephone
e-mail

B Project description

1 State of the art and preliminary work

The comparison of trend curves is an important topic in many statistical applications. Economists, for example, are interested in comparing the trends of long-term interest rates for different countries. Moreover, they may want to assess whether the trends in real GDP growth differ across countries. In finance, massive amounts of data on thousands of stocks are available today. One question of interest is to compare how the volatility of different stocks evolves over time. Finally, in climatotology, large spatial data sets have been collected which comprise long temperature time series for many different locations. Climatologists are very much interested in analyzing the trending behaviour of these time series. In particular, they would like to know how the temperature trend varies across locations.

The main aim of this project is to develop new methods and theory for the comparison of nonparametric trend curves. Classically, time trends are modelled stochastically in econometrics, e.g. by a unit root model [see ??]. Recently, there has been a growing interest in models with deterministic time trends [see ??]. An interesting modelling framework considered in ?? among others is as follows: Suppose we observe a number of time series $\mathcal{Y}_i = \{Y_{it} : 1 \le t \le T\}$ for $1 \le i \le n$. Each time series $\mathcal{Y}_i$ is modelled by the equation

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \beta_i^\top X_{it} + \alpha_i + \varepsilon_{it} \qquad (1)$$

for $1 \leq t \leq T$, where m_i is a nonparametric trend curve, $X_{it} = (X_{it,1}, \ldots, X_{it,d})$ is a d-dimensional vector of regressors or controls and β_i is the corresponding parameter vector, α_i are so-called fixed effect error terms and ε_{it} are standard regression errors with $\mathbb{E}[\varepsilon_{it}] = 0$ for all t. Within model (1), one may approach several interesting statistical questions.

(a) Testing for equality of nonparametric trend curves.

The first question is the following: Are all time trends m_i the same? That is, do all time series in the sample exhibit the same trending behaviour? This question can be approached formally by means of a statistical test. The null hypothesis can be formulated as $H_0 : m_1 = \ldots = m_n$.

- Stock and Watson (1988) is one of the first papers to compare trend curves in a multiple time series. However, the focus of the author's attention is rather on stochastic trends than deterministic ones. The authors develop two tests for detecting common stochastic trends in a number of time series. They apply these tests to the economic time series, in particularly, the postwar U.S. data on the federal funds rate and the three- and twelve-month treasury bill rates. All three time series on inerest rates appear to share a common stochastic trend.

- Vogelsang and Franses (2005) consider a simple linear model

  $$Y_{it} = \alpha_i + \beta_i t + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{2}$$

  where $\beta_i t$ is a linear trend function, α_i are so-called fixed effect error terms and ε_{it} are standard regression errors such that a functional central limit theorem is applicable to $\{\varepsilon_{it}\}$. This model can be considered as a special case for our model (1) with nonparametric time curve $m_i(t)$ being linear in t. The authors propose two $F-$tests and a t-test to test the null hypothesis $H_0 : R\beta = r$, where R is $q \times n$ known deterministic matrix and r is $q \times 1$ known deterministic vector. They derive an asymptotic theory for these tests and provide relevant critical values. As an empirical application, the authors compare the postwar European time series of gross domestic product (GDP) to the time series of GDP in Italy. They reject the null hypothesis that the rates of growth between Italy and 6 other European countries in the years 1950 to 1992 are the same.

- Park and Kang (2008), Park, Hannig and Kang (2009a) and Park, Vaughan, Hannig and Kang (2009b) extend the well-known SiZer method for analyzing one time series, which was originally proposed by Chaudhuri and Marron (1999), and advanced as a procedure to analyze time series Rondonotti et al. (2007). In Park and Kang (2008), only the model with independent idiosyncratic errors $\{\varepsilon_{it}\}$ is considered, whereas in Park, Hannig and Kang (2009a) and Park, Vaughan, Hannig and Kang (2009b) the errors are allowed to be dependent across t. The authors use a regression function estimation to fit a local linear function to obtain a kernel etimate. This kernel estimate depends on a location x and a bandwidth h. The proposed SiZer method then uses the color map to display the significance of differences between two regression functions for a range of locations x and locations h. This is a useful graphical tool that can indicate the regions where the differences between trend curves should be investigated further. However, this method has its limitations since the comparison should be done only pairwise

- Degras, Xu, Zhang and Wu (2012) is a seminal paper on the parallelism between the deterministic trends in multiple time series. The model considered is

$$Y_{it} = m_i\left(\frac{t}{T}\right) + \varepsilon_{it} \tag{3}$$

  for $1 \leq t \leq T$ and $1 \leq n \leq N$, where m_i is a nonparametric trend curve and ε_{it} are standard regression errors with zero mean for all t. Furthermore, ε_{it} are allowed to be weakly dependent in the terms of Wu (2005). This dependence is non-stationary and it generalizes the stationary assumptions on the error process used previously in the literature. In our proposed project we also intend to follow this model (and this non-stationary error process), further including the fixed effect α_i and the covariates $\{X_{it}\}$.

  The authors test the parrallel hypothesis, specifically, $H_0 : m_i(u) = c_i + m(u)$ for all $i = 1, \ldots, N$ and $u \in [0, 1]$. In this setting c_i are considered to be vertical shifts between the reference curve $m(u)$ and the i-specific curve $m_i(u)$. In order to do so, the authors first estimate the individual trends $m_i(u) - c_i$ and the common trend $m(u)$ by the means of local linear smoothing procedure and then they develop a test based on the L_2-distances between the estimators. They derive the asymptotic theory for the proposed test and decise a clustering algorithm based on the test statistic. As an illustration of the use of the proposed method, the authors analyze download trends (up to a scale) in the time series that consist of cell phone download activity in different areas in the United States. This is an interesting question for developing region-specific advertising strategies for cell phone companies.

- Sun (2011) add covariates to a simple linear trend model regression.

- Xu (2012).

- Zhang, Su and Phillips (2012).

- Hidalgo and Lee (2014).

Most tests of the hypothesis $H_0 : m_1 = \ldots = m_n$ existing in the literature proceed in two steps: They first estimate the curves of interest by nonparametric methods and then construct a distance measure between the estimated curves which serves as a test statistic. By construction, these tests depend on one or several smoothing parameters which are needed to estimate the curves m_i. However, there is no theory available for a proper choice of the bandwidth/smoothing parameter. In particular, the optimal (MSE minimizing) bandwidth used for curve fitting is usually not optimal for testing. A classical way to get a bandwidth-free test statistic is to use empirical process theory and partial sum processes [cp. Hidalgo and Lee (2014)]. A more modern way which is related to these partial sum processes are so-called multiscale tests. The idea is as follows: ??

Multiscale tests for the comparison of nonparametric curves under general conditions are not available to the best of our knowledge. One aim of the project is to develop such a test for nonparametric regression curves. Multiscale tests do not only have the advantage of being bandwidth-free. They also are much more informative compared to other tests. They do not only allow to test whether the curves m_i are all the same or not; they also allow to say, with a pre-specified statistical confidence, which curves are different and in which regions they differ.

(b) Clustering of nonparametric trend curves.

When the number of curves is large, classical tests for the comparison of nonparametric curves are not fully appropriate as a statistical tool. The issue is the following: In most applications where the number of curves is large, one can expect that not all curves are exactly the same. Hence, a test of the null that all curves are the same is quite uninformative. Most frequently, the hypothesis will be rejected. A more interesting question is the following: Are there groups of curves that are the same? This question leads to the problem of curve clustering. Clustering of coefficient or functions in panel data models is a relative young emerging field in econometrics:

– Bonhomme and Manresa (2015). Grouped patterns of heterogeneity in panel data.

– Su, Shi and Phillips (2016). Identifying latent structures in panel data.

– Su and Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects.

– Wang et al. (2018). Homogeneity pursuit in panel data models: theory and application.

In the statistics literature, there is also a literature on curve clustering (functional and longitudinal data):

– Abraham, Cornillon, Matzner-Løber and Molinari (2003). Unsupervised curve clustering using B-splines.

– James and Sugar (2003). Clustering for sparsely sampled functional data.

– Tarpey and Kinateder (2003). Clustering functional data.

– Ray and Mallick (2006). Functional clustering by Bayesian wavelet methods.

– Chiou and Li (2007). Functional clustering and identifying substructures of longitudinal data.

– Degras, Xu, Zhang and Wu (2012). Testing for parallelism among trends in multiple time series.

Most of the clustering procedures in the literature depend on a number of smoothing parameters. Multiscale approaches do not.

1.1 Project-related publications

1.1.1 Articles published by outlets with scientific quality assurance, book publications, and works accepted for publication but not yet published

1.1.2 Other publications

2 Objectives and work programme

2.1 Anticipated total duration of the project

2 years from 01.10.2019 to 30.09.2021

2.2 Objectives

The main aim of the project is to develop new methods and theory for the comparison and clustering of nonparametric curves. We intend to consider the following model framework: Suppose we observe a number of time series $\mathcal{Y}_i = \{Y_{it} : 1 \le t \le T\}$ for $1 \le i \le n$. Each time series $\mathcal{Y}_i$ is modelled by the equation

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \beta_i^\top X_{it} + \alpha_i + \varepsilon_{it} \tag{4}$$

for $1 \le t \le T$, where m_i is a nonparametric trend curve, $X_{it} = (X_{it,1}, \dots, X_{it,d})$ is a d-dimensional vector of regressors or controls, α_i are so-called fixed effect error terms and ε_{it} are standard regression errors with $\mathbb{E}[\varepsilon_{it}] = 0$ for all t. As usual in nonparametric regression, we let m_i depend on rescaled time t/T rather than real time t; compare ??, ?? and ?? among many others for the use of the rescaled time argument. For simplicity, the controls X_{it} are assumed to enter the model equation linearly with β_i being the corresponding parameter vector. However, it is possible to extend the model to allow for nonlinear parametric and even nonparametric specifications of X_{it}. [Conditions on the fixed effects and the error terms.]

The first main contribution of the project is to contruct a novel multiscale test for the comparison of the trend curves m_i $(1 \le i \le n)$. Compared to existing methods, the approach has the following main advantages:

(1)

(2)

To the best of our knoweledge, there is no other multiscale method available in the literature. The only exception is ?? who have developed theory for the case $n = 2$. However, the theory is developed under severe restrictions: ??. We do not only aim to develop methodology but also derive a complete asymptotic theory for the proposed multiscale test. In particular, we will derive the limit distribution and analyse the behaviour under (local) alternatives.

The second main contribution is to develop a clustering approach which is based on the multiscale test from the first main part of the project. The only multiscale clustering method available in the literature is Vogt and Linton (2018). They consider a very general nonparametric regression model but only derive consistency results for the clustering method. We consider a somewhat simpler model but will derive a complete distribution theory for the clustering method (which in particular allows to make not only converge statements but also confidence statements about the estimated groups and their number).

Model (4) and the proposed testing/clustering method are useful in a number of application contexts which we aim to explore. We here give some examples:

Example 1. *Short-term risk-free interest rates are one of the main topics of interest in the financial markets. For example, it is a key component of the capital asset pricing model, which describes the relationship between risk and return. Furthermore, the risk-free rate is also a required input in financial calculations regarding the pricing of bonds. There is an evergrowing amount of literature on the dynamics of interest rate. US Treasury bills are the real-world investment that serve as the proxy for these rates. Park et al. (2009b) analyze the yields of the 3-month, 6-month, and 12-month Treasury bills in the context of comparing nonparametric curves. The authors assume that the yields come from the following model:*

$$Y_{it} = m_i(t) + \sigma_i \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T, \tag{5}$$

which is a simplification of our model (4)*. Park et al. (2009b) apply SiZer method to the data and come to the conclusion that the underlying structure for different time periods is almost identical. They could not find any significant difference between any pair of the time periods, which concides with the results from applying other methods, see, for example, Fan and Yao (2008).*

Example 2. *Another example of comparison of time series with nonparametric trend functions described in Park et al. (2009b) involves the long-term rates for US, Canada, and Japan from January 1980 to December 2000. The data is assumed to follow the same model* (5)*. The authors perform pairwise comparison of the curves as well as comparison of the three time series at the same time using the proposed SiZer method. In both cases their method was able to detect significant differences nd indicate "suspicious" regions. However, since SiZer is a graphical device that is mainly designed for data exploration rather than for rigorous statistical inference, they do not make simultaneous confidence statements with a predetermined confidence level about the regions where these differences were most probable to occur. Our proposed multiscale method, in contrast, is a rigorous level-α-test of the hypothesis H_0 which is aimed specifically at that.*

Example 3. *Economic growth has been a key topic in marcoeconomics over many decades. Economists are very much interested in the question whether gross domestic product (GDP) growth has been faster in some countries than in others. One of the ways to model the source of economic growth is to incorporate a nonparametric deterministic time trend in the model. For example, Zhang et al. (2012) consider such a model for the OECD economic growth data. Specifically, they investigate the following model for growth rates:*

$$\Delta \ln GDP_{it} = \beta_1 \Delta \log L_{it} + \beta_2 \Delta \log K_{it} + \beta_3 \Delta \log H_{it} + f_i(t/T) + \alpha_i + \varepsilon_{it}, \qquad (6)$$

where $i = 1, \ldots, n$, $t = 1, \ldots, T = 140$, GDP is gross domestic product, K is capital stock, L is labour input, H is human capital, α_i is a fixed effect, $f_i(\cdot)$ is an unknown smooth time trend furnction and ε_{it} are idiosyncratic errors. The errors are allowed to be dependent cross-sectionally, but not serially over t. The data comes from $n = 16$ OECD countries.

Zhang et al. (2012) estimate the common component of time trends which appears to be significantly different from zero over a wide range its support. Moreover, they test the null hypothesis that there are no significant differences in the time trends for the 16 OECD countries. Based on the bootstrap p-values the authors are able to reject the null hypothesis of all the trends being equal at the 10% confidence level. Hence, it can be interesting to be able to further cluster the OECD countries based on their economic growth rates.

Example 4. *The issue of global warming has been a vital topic for many scientists over the last few decades. Since the late 1970, different models that describe the global temperature have been published. In the current literature it is common to assume that the temperate time series (global as well as local) follow a model that can be decomposed into a deterministic trend component and a noise component, see, for example, Ghil and Vautard (1991) and Mudelsee (2018). In order to estimate and attribute the trends in climate variables, a variety of econometric methods have been employed, starting from the simple linear (Yue et al. (2013)) and quadratic regression (??) to the empirical mode decompoistion (Wu et al. (2011)), spectrum analysis (Ghil and Vautard (1991)) and semi- and fully non-parametric methods (Gao and Hawthorne (2006)). Parametric and change points methods are mostly suited to quantify the magnitude of the warming trend or to determine the change points, whereas nonparametric methods are best designed to describe the trend over the full time interval without imposing any additional structure on it. However, most of these papers apply nonparametric methods to analyze only one time series or the authors assume that the trend function is common for different time series (Atak et al. (2011)). To our knowledge, only a few papers regarding the comparison of warming trends in different cities or countries have been published (Zhang et al. (2012)).*

Zhang et al. (2012) propose the following semiparametric panel model for unbalanced data to describe the trend in UK regional temperatures:

$$y_{it} = \beta_i^T D_t + m_i(t/T) + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T \tag{7}$$

where y_{it} are the monthly mean maximum temperature, monthly mean minimum temperature or total rainfall in millimeters at a station i in month t, D_t is a 11-dimensional vector of monthly dummy variables, α_i is the fixed effect for station i, $m_i(\cdot)$ is an unknown trend function and ε_{it} are idiosyncratic errors. The dataset used is the balanced panel data set for $n = 26$ stations in UK for $T = 382$ months from October 1978 to July 2010. This model is a special case of our proposed model (4) with dummy variables as covariates.

Zhang et al. (2012) are interested in testing the null hypothesis $m_i = m$ for all $i = 1, 2, \dots, n$. In order to do this, they apply a non-parametric R^2-based test for common trends that was developed in their paper. Based on the obtained p-values, they reject the null hypothesis of common trend at 5% level for the monthly mean maximum temperature and the monthly mean minimum temperature. However, they do not reject the null hypothesis for the total rainfall eve at the significance level of 10%. As before, it would be interesting to further cluster the UK stations based on the common trend in order to be able to detect the causes of this warming trend. Moreover, it can also be of particular interest to see in which time regions the trends are significantly different from each other.

2.3 Work programme incl. proposed research methods

All phases of the research will be conducted in close collaboration with the partners in Bonn.

Milestone	**2019**	**2020**	**2021**
	Month	Month	Month
Multiscale inference for fixed number of time series	10-12	1-9	
Multiscale inference for growing number of time series		10-12	1-9

2.4 Data handling

2.5 Other information

Please use this section for any additional information you feel is relevant which has not been provided elsewhere.

[Text]

3 Bibliography

References

ABRAHAM, C., CORNILLON, P.-A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, **30** 581–595.

ATAK, A., LINTON, O. and XIAO, Z. (2011). A semiparametric panel model for unbalanced data with application to climate change in the united kingdom. *Journal of Econometrics*, **164** 92–115.

BONHOMME, S. and MANRESA, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, **83** 1147–1184.

CHAUDHURI, P. and MARRON, J. S. (1999). SiZer for the exploration of structures in curves. *Journal of the American Statistical Association*, **94** 807–823.

CHIOU, J.-M. and LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69** 679–699.

DEGRAS, D., XU, Z., ZHANG, T. and WU, W. B. (2012). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, **60** 1087–1097.

FAN, J. and YAO, Q. (2008). *Nonlinear time series: nonparametric and parametric methods.* Springer Science & Business Media.

GAO, J. and HAWTHORNE, K. (2006). Semiparametric estimation and testing of the trend of temperature series. *The Econometrics Journal*, **9** 332–355.

Ghil, M. and Vautard, R. (1991). Interdecadal oscillations and the warming trend in global temperature time series. *Nature*, **350** 324.

Hidalgo, J. and Lee, J. (2014). A cusum test for common trends in large heterogeneous panels. In *Essays in Honor of Peter CB Phillips*. Emerald Group Publishing Limited, 303–345.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98** 397–408.

Mudelsee, M. (2018). Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*.

Park, C., Hannig, J. and Kang, K.-H. (2009a). Improved SiZer for time series. *Statistica Sinica*, **19** 1511–1530.

Park, C. and Kang, K.-H. (2008). Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, **52** 3954–3970.

Park, C., Vaughan, A., Hannig, J. and Kang, K.-H. (2009b). SiZer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, **139** 3974–3988.

Ray, S. and Mallick, B. (2006). Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68** 305–332.

Rondonotti, V., Marron, J. S. and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, **1** 268–289.

Stock, J. H. and Watson, M. W. (1988). Testing for common trends. *Journal of the American statistical Association*, **83** 1097–1107.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, **206** 554–573.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, **84** 2215–2264.

Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *Journal of Econometrics*, **164** 345–366.

Tarpey, T. and Kinateder, K. K. (2003). Clustering functional data. *Journal of classification*, **20** 093–114.

Vogelsang, T. J. and Franses, P. H. (2005). Testing for common deterministic trend slopes. *Journal of Econometrics*, **126** 1–24.

Vogt, M. and Linton, O. (2018). Multiscale clustering of nonparametric regression curves. Tech. rep., Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Wang, W., Phillips, P. C. and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, **33** 797–815.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natn. Acad.*

Sci. USA, **102** 14150–14154.

Wu, Z., Huang, N. E., Wallace, J. M., Smoliak, B. V. and Chen, X. (2011). On the time-varying trend in global-mean surface temperature. *Climate dynamics*, **37** 759.

Xu, K.-L. (2012). Robustifying multivariate trend tests to nonstationary volatility. *Journal of Econometrics*, **169** 147–154.

Yue, T.-X., Zhao, N., Ramsey, R. D., Wang, C.-L., Fan, Z.-M., Chen, C.-F., Lu, Y.-M. and Li, B.-L. (2013). Climate change trend in china, with improved accuracy. *Climatic Change*, **120** 137–151.

Zhang, Y., Su, L. and Phillips, P. C. (2012). Testing for common trends in semi-parametric panel data models with fixed effects. *The Econometrics Journal*, **15** 56–100.

4 Requested modules/funds

Explain each item for each applicant (stating last name, first name).

4.1 Basic Module

4.1.1 Funding for Staff

Nr.	Position	2019	2020	2021
1	Research staff U. Bonn (EGr. 13 TV-L 75 %)	11.869 €	47.475 €	35.606 €
2	Student Assistant Bonn	2.700 €	10.800 €	8.100€
	Required Amount	14.569€	58.275€	43.706€

Job description of staff payed from auxiliary support for the funding period requested

1. Marina Khismatullina already possesses considerable experience in the study of nonparametric models with time series error. Moreover, she is a co-author of the paper “Multiscale Inference and Long-Run Variance Estimtor in Nonparametric Regression with Time Series Friends” by Khismatullina and Vogt, which is currently submitted to JRSSB. She will be capable to develop computational software taylored to assess the empirical performance of the proposed multiscale test.

2. At the onset of the project a student assistent position should be available in order to support stuff with exploratory data analysis, data mining and organisational issues. The prerequisities are strong analytical and programming skills.

4.1.2 Direct Project Costs

[Text]

4.1.2.1 Equipment up to Euro 10,000, Software and Consumables

[Text]

4.1.2.2 Travel Expenses

[Text]

4.1.2.3 Visiting Researchers (excluding Mercator Fellows)

[Text]

4.1.2.4 Expenses for Laboratory Animals

[Text]

4.1.2.5 Other Costs

[Text]

4.1.2.6 Project-related publication expenses

[Text]

5 Project requirements

5.1 Employment status information

For each applicant, state the last name, first name, and employment status (including duration of contract and funding body, if on a fixed-term contract).

[Text]

5.2 First-time proposal data

Only if applicable: Last name, first name of first-time applicant

[Text]

5.3 Composition of the project group

List only those individuals who will work on the project but will not be paid out of the project funds. State each person's name, academic title, employment status, and type of funding.

[Text]

5.4 Cooperation with other researchers

5.4.1 Researchers with whom you have agreed to cooperate on this project

[Text]

5.4.2 Researchers with whom you have collaborated scientifically within the past three years

[Text]

5.5 Scientific equipment

The University of Bonn has a sufficient infrastructure in hard- and software. Personal computers are available and can be used within the project. Equipment like printer, fax and copier can be used as well.

6 Additional information

If applicable, please list proposals requesting major instrumentation and/or those previously submitted to a third party here.

[Text]