

# 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 there are different trends in real GDP growth 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  $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$  for  $1 \leq i \leq n$ . Each time series  $\mathcal{Y}_i$  is modelled by the equation

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

for  $1 \leq t \leq T$ , 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 and  $\beta_i$  is the corresponding parameter vector,  $\alpha_i$  are so-called fixed effect error terms and  $\varepsilon_{it}$  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  $H_0 : m_1 = \dots = m_n$ . [Literature on comparison of trend curves and more generally of regression curves.]

- ? Stock & Watson (1988). Testing for common trends.
- Vogelsang & Franses (2005). Testing for common deterministic trend slopes.
- Park, Vaughan, Hannig & Kang (2009). SiZer analysis for the comparison of time series.
- Degras, Xu, Zhang & Wu (2011). Testing for parallelism between trends in multiple time series.
- Sun (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter.
- Xu (2011). Robustifying multivariate trend tests to nonstationary volatility.
- Zhang, Su & Phillips (2012). Testing for common trends in semi-parametric panel data models with fixed effects.
- Hidalgo & Li (2014?). A CUSUM test of common trends in large heterogeneous panels.

Most tests of the hypothesis  $H_0 : m_1 = \dots = m_n$  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 & 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 & Manresa (2015). Grouped patterns of heterogeneity in panel data.
- Su, Shi & Phillips (2016). Identifying latent structures in panel data.
- Su & Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects.
- Wang, Phillips, & Su (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-Lober & Molinari (2003). Unsupervised curve clustering using B-splines.
- James & Sugar (2003). Clustering for sparsely sampled functional data.
- Tarpey & Kinateder (2003). Clustering functional data.
- Ray & Mallick (2006). Functional clustering by Bayesian wavelet methods.
- Chiou & Li (2007). Functional clustering and identifying substructures of longitudinal data.

- Degras, Xu, Zhang, & 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  $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$  for  $1 \leq i \leq n$ . Each time series  $\mathcal{Y}_i$  is modelled by the equation

$$Y_{it} = m_i\left(\frac{t}{T}\right) + \beta_i^\top X_{it} + \alpha_i + \varepsilon_{it} \quad (2)$$

for  $1 \leq t \leq T$ , 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,  $\alpha_i$  are so-called fixed effect error terms and  $\varepsilon_{it}$  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  $\beta_i$  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 construct a novel multiscale test for the comparison of the trend curves  $m_i$  ( $1 \leq i \leq n$ ). 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 cluseretering method available in the literature is Vogt & 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 (2) 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.** *Examples from Park et al. (2009), SiZer analysis for the comparison of time series and Zhang et al. (2012), Testing for common trends in semi-parametric panel data models with fixed effects. [Add that the analysis of climate data with econometric methods has attracted some attention in recent years and cite econometrics literature.]*

## 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 Month	2020 Month	2021 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

- PARK, C., VAUGHAN, A., HANNIG, J. and KANG, K.-H. (2009). SiZer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, **139** 3974–3988.
- 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 paid 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 Estimator 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 tailored to assess the empirical performance of the proposed multiscale test.
2. At the onset of the project a student assistant position should be available in order to support staff with exploratory data analysis, data mining and organisational issues. The prerequisites 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]