

Burdejová, Petra; Härdle, Wolfgang Karl

**Working Paper**

## Dynamic semi-parametric factor model for functional expectiles

SFB 649 Discussion Paper, No. 2017-027

**Provided in Cooperation with:**

Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

*Suggested Citation:* Burdejová, Petra; Härdle, Wolfgang Karl (2017) : Dynamic semi-parametric factor model for functional expectiles, SFB 649 Discussion Paper, No. 2017-027, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:

<https://hdl.handle.net/10419/191791>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Dynamic semi-parametric factor model for functional expectiles

Petra Burdejová\*  
Wolfgang K. Härdle\*



\* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

<http://sfb649.wiwi.hu-berlin.de>  
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin  
Spandauer Straße 1, D-10178 Berlin



# Dynamic semi-parametric factor model for functional expectiles \*

Petra Burdejová<sup>1</sup> and Wolfgang K. Härdle<sup>1,2</sup>

<sup>1</sup>Humboldt-Universität zu Berlin, Berlin, Germany.

<sup>2</sup>Sim Kee Boon Institute for Financial Economics, Singapore Management  
University, Singapore.

December 1, 2017

## Abstract

High-frequency data can provide us with a quantity of information for forecasting, help to calculate and prevent the future risk based on extremes. This tail behaviour is very often driven by exogenous components and may be modelled conditional on other variables. However, many of these phenomena are observed over time, exhibiting non-trivial dynamics and dependencies. We propose a functional dynamic factor model to study the dynamics of expectile curves. The complexity of the model and the number of dependent variables are reduced by lasso penalization. The functional factors serve as a low-dimensional representation of the conditional tail event, while the time-variation is captured by factor loadings. We illustrate the model with an application to climatology, where daily data over years on temperature, rainfalls or strength of wind are available.

**Keywords:** factor model, functional data, expectiles, extremes.

**JEL Classification:** C14, C38, C55, C61, Q54.

---

\*This research was supported by the International Research Training Group IRTG 1792 "High Dimensional Non Stationary Time Series" and the Collaborative Research Center CRC 649 "Economic Risk", Humboldt-Universität zu Berlin.

# 1 Introduction

Statistical analysis of high-dimensional data nowadays plays a crucial role in various fields. Usually, one observes a high-dimensional vector evolving in time, that can be not only correlated to other variables but hide various types of inter-dependencies. One solution on how to analyze such data for further modelling is to treat it as discrete observations of functional times series. For example, in climatology and meteorology the evolution of temperature curves as Ramsay and Silverman (2005), the wind speed as Burdejova et al. (2017), or pollution data as Ignaccolo et al. (2008), observed as a function of time over the year, can exhibit the trend or interesting periodical pattern supporting the thesis of climate changes. The same approach of functional data analysis was applied in health-care and clinical research, see e.g. Erbas et al. (2007) who tested the trend in breast-cancer mortality, or Lee and Carter (1992) who performed the population analysis for mortality and fertility curves. Countless applications can be found in financial engineering, for example yield curve modelling as Nelson and Siegel (1987) or Härdle and Majer (2016), modelling the collateralized debt obligations, see Choros-Tomczyk et al. (2016), analyzing the dynamics of limit order book or implied volatility, see e.g. Benko et al. (2009) or even the intraday price curves, see Kokoszka et al. (2014).

However, in most of the above mentioned applications, one is interested in capturing the tail behaviour of the variables rather than variation around the mean. The majority of recent research in functional data has nonetheless focused only on the variation around the mean, as can be seen in monographs of Ramsay and Silverman (2005), Ferraty and Vieu. (2006) or Horváth and Kokoszka (2012).

For that purpose, in our work we generalise one of the functional analytical models for expectiles. Expectiles, similar to quantiles, are tail measures, which uniquely characterise the conditional distribution of random variables. The same way as the quantile for the level of  $\tau = 0.5$  corresponds to the me-


dian, the  $\tau = 0.5$  expectiles correspond to the mean. In case of the conditional expectiles on the other variable, e.g. time over year for the temperature or time over day for intra-day price curves, we refer to, so-called, expectile curves.

Guo et al. (2013) modelled such expectile and quantile curves by rewriting them via Karhunen-Loève expansion. Tran et al. (2016) also presented an analogue principal components of such tail event curves in an asymmetric norm. Both of these approaches assume the observations to be the independent realizations of a stationary stochastic process. Even-though Hörmann and Kokoszka (2010) showed that Karhunen-Loève expansion approach is suitable for the temporal dependence between functional observations as well, the question of modelling strong-dependent or non-stationary functional observations of extremes remains open. Therefore, the goal of our model is three-fold: focus on the modeling of the conditional extreme events, e.g. different expectile-levels, capture the dynamics of such tail event curves and do that with respect to any hidden pattern, dependence or non-stationarity.

In this work we extend the generalized dynamic semi-parametric factor model for expectiles and provide the convergent algorithm for its estimation. This approach offers focus on modeling the time-development of tail-event expectile curves with respect to possible strong-dependency or non-stationarity as well. Our work refers to the factor models as in Park et al. (2009) and Song et al. (2014), who did similar dynamic semi-parametric factor models for the  $L_2$ -norm, which, for our model, corresponds as a specific case of expectile at 0.5-level.

As a motivation, let us assume that there is a need to estimate a collection of expectile curves, each coming from a separate data-set. Our first motivation example used in Chapter 5 regards a set of daily observations of average temperature, i.e. the data as a vector  $X_n \in \mathbb{R}^{365}$  for a year  $n = 1, \dots, N$ . In this situation, one needs to analyse jointly the time (over years for  $n = 1, \dots, N$ ) as well as space dynamics (within the year for  $t$ -th

element of  $X_n[t]$  for  $t = 1, \dots, 365$ ) by simultaneous fitting. In the factor model, functional factors serve as a low-dimensional representation of the conditional tail event, while the time-variation is captured by factor loadings. All of them we approximate by a linear combination of basis-functions. Since the temporal dependencies and non-stationarity can arise from different sources, we start with an overparametrized model, which captures almost any behaviour or trend, e.g. the cycles, the linear or quadratic trend as well. Since then we seek a sparse solution and reduce the complexity of the model and the number of dependent variables by lasso penalization. Further, we apply the proposed model for estimation and forecasting of daily temperature. Other applicable examples can be also found in the analyses of customer demand planning, e.g. forecasting of electricity consumption as seen in López-Cabrera and Schulz (2017) via VAR-model or finance, economics, climatology or neurobiology as mentioned already above.

The paper is organized as follows. After reviewing the concept of expectile curves in the first chapter we present the dynamic semi-parametric factor model for expectiles and the algorithm to estimate this model based on the iterative weighted least squares. In Chapter 3 we examine the performance of our model and algorithm in the simulation study. Finally, we apply the model to a Chinese temperature dataset in Chapter 5 and to the dataset of the wind speed of hurricanes in Chapter 6. The last section summarises our findings. All codes used to obtain the results in this paper are available at  Quantlet, see details in Borke and Härdle (2017a) and Borke and Härdle (2017b).

## 2 Expectiles and expectile curves

The concept of expectiles was first presented by Newey and Powell (1987). Expectiles have a similar interpretation as quantiles, but are more efficient, easier to compute due to the  $L_2$ -norm and they are also a coherent risk

measure. Having a random variable  $Y$  the  $\tau$ -expectile can be obtained by minimizing the expected loss:

$$e^\tau = \arg \min_{\theta} E \{ \rho_\tau(Y - \theta) \}$$

with asymmetric loss function

$$\rho_\tau(u) = |u|^\alpha \left| \tau - \mathbf{I}_{\{u < 0\}} \right|,$$

where  $\alpha = 2$ . In case of  $\alpha = 1$  we get the quantiles. By generalization we can also get M-quantiles, see Breckling and Chambers (1988) or Jones (1994), who also showed that expectiles can be expressed as quantiles.

Expectiles can be understood intuitively in a similar way as quantiles. Though the  $\tau$ -quantile can be defined as a value above the  $\tau \cdot 100\%$  observations, expectile also takes the distance into the account.  $\tau$ -expectile is defined such that  $\tau \cdot 100\%$  of the distance of observations to it corresponds to the observations below it. Thus, expectiles are more sensitive to the extreme observations and outliers.

However, in reality  $Y$  is usually associated with a vector of covariates  $X$ , e.g. the variable  $X$  can express the development over time, i.e.:

$$e^\tau(x) = \arg \min_{\theta} E \{ \rho_\tau(Y - \theta) \mid X = x \}.$$

One is then interested in studying the conditional expectile as a function of  $x$ . For that purpose we define the generalized regression  $\tau$ -expectile as:

$$e^\tau(x) = \arg \min_{f \in \mathcal{F}} E \{ \rho_\tau(Y - f(X)) \}.$$

where  $f(\cdot)$  is a nonparametric function of covariates  $X$  from a set of functions  $\mathcal{F}$ , such that the expectation is well defined.

There are more possibilities on how to estimate such an expectile curve from an observed dataset. For example, expectile curve  $e^\tau(x)$  can be approx-

imated by any basis and estimated iteratively. Schnabel and Eilers (2009) proposed to approximate the curves with P-splines and combine it with the LAWS (least average weighted squares) algorithm.

The aim of our work is to model a collection of  $N$  generalized expectiles curves  $e_n^\tau(x), n = 1, \dots, N$  with semi-parametric factor model.

### 3 Dynamic semi-parametric factor model for expectile curves

Let us fix the level  $\tau$  and assume the functional time series  $e_n, n = 1, \dots, N$ . We represent such a random process via factor model :

$$e_n(t) = \sum_{k=1}^K Z_{nk} m_k(t) = Z_n^\top m(t),$$

with time-varying factor loadings  $Z_{nk}$  and functional factors  $m_k(t)$ . Index  $t$  captures the spatial dependency while the index  $n$  express the evolution over time.

Suppose both sequences factorize over space and time with respect to some fixed bases. Thus, for some  $J$ -dimensional time basis  $U^\top = (U_1, \dots, U_J)$  with  $U_i = (U_i(1), \dots, U_i(N))$ ,  $i = 1, \dots, J$  and  $L$ -dimensional space basis  $\Psi = (\Psi_1, \dots, \Psi_L)$  with  $\Psi_i = (\Psi_i(1), \dots, \Psi_i(T))$ , we have the decomposition:

$$Z_{nk} = \sum_{j=1}^J \alpha_{kj} u_j(n) \quad \text{i.e.} \quad Z_n = A_{K \times J} \cdot U(n)$$

and

$$m_k(t) = \sum_{l=1}^L \beta_{kl} \Psi_l(t) \quad \text{i.e.} \quad m(t) = B_{K \times L} \Psi(t),$$

which lead to the final dynamic semi-parametric factor model:



$$e_n = (AU)^\top (B\Psi) = U^\top C\Psi, \quad (1)$$

where  $C = A^\top B$  is a  $J \times L$  matrix of coefficients needed to be estimated.

For the choice of both basis, one may employ various basis functions. To capture the periodic variation in time one can use the fourier basis, for the global trend over time, any orthogonal polynomial basis may be suitable. For the purpose of space spaces either B-splines, any polynomial basis or even principal components or their alternatives such as principal expectile components defined by Tran et al. (2016) may be used.

In order to estimate this model we propose the iterative algorithm for minimising the penalized loss function. We define the weights in a similar manner as in Schnabel and Eilers (2009). Once the space and time basis are pre-specified, the choice of significantly loaded space and time basis functions is done via LASSO-penalization of the coefficient matrix  $C$ . As before, assume the fixed expectile level  $\tau \in (0, 1)$  and for the observed discrete dat-points  $Y_{n,t}$ ,  $n = 1, \dots, N$  and  $t = 1, \dots, T$ :

1. Start with a set up for the weights  $w_{n,t} = 0.5$ . That corresponds to the mean curves.
2. Estimate the matrix  $\hat{C}$  by minimising

$$\underset{C}{\operatorname{argmin}} \underbrace{\sum_{n=1}^N \sum_{t=1}^T w_{n,t} \{Y_{n,t} - U(n)^\top C\Psi(t)\}}_{l(C)} + \lambda \sum_{j=1}^J \|c_{\mathcal{G}_j}\|_2,$$

where the penalization term  $\lambda \sum_{j=1}^J \|c_{\mathcal{G}_j}\|_2$  is a group-Lasso penalization.

3. Update the weights

$$w_{n,t} = \begin{cases} \tau & \text{if } Y_{n,t} > U(s)^\top \widehat{C}\Psi(t), \\ 1 - \tau & \text{otherwise.} \end{cases}$$

4. Iterate via Steps 2. and 3. Recompute the weights until convergence, i.e. until there is no change in weights  $w_{n,t}$ .

Even-though we can set the separate elements of  $C$ -matrix as the groups in LASSO-penalization, the group-LASSO would also allow us to give some specific structure or importance into the pre-defined basis if needed. Note that function  $l(C)$  in Step 2 is continuously differentiable and obtains a global minimum. Yang and Zou (2015) proposed the algorithm to solve such optimization problem and proved its convergence for different types of "empirical loss + group lasso penalty" optimisation problems satisfying a quadratic majorization condition. It is easy to show that our specific definition of weighted least squares in combination to group-lasso penalization in Step 2 for the fixed weights fulfill these requirements.

## 4 Simulation study

In order to evaluate the performance of the proposed model and the algorithm above we did a simulation study. We follow the set up of Guo et al. (2013) or Tran et al. (2016), since they both proposed the alternatives for the modelling of tail event expectile curves. The data  $Y_{n,i}$ ,  $n = 1, \dots, N$  and  $i = 1, \dots, T$ , are simulated as:

$$Y_{n,i} = \mu(t_i) + \alpha_{1,n}f_1(t_i) + \alpha_{2,n}f_2(t_i) + \varepsilon_{n,i},$$

where  $t_i$ 's are the equidistant points on  $[0, 1]$ . We set the mean function  $\mu(t)$  as  $\mu(t) = 1 + t + \exp\{-(t - 0.6)^2/0.05\}$  and the principal component curves as  $f_1(t) = \sqrt{2}\sin(2\pi t)$  and  $f_2(t) = \sqrt{2}\cos(2\pi t)$ .

Further, we consider the two following different scenarios for the scores of principal components  $\alpha_{1,n}$  and  $\alpha_{2,n}$  and 4 different error scenarios:

1. The scores set as  $\alpha_{1i} \sim N(0, 36)$  and  $\alpha_{2i} \sim N(0, 9)$  are both iid. The error term  $\varepsilon_{n,i}$ 's is: (1) iid  $N(0, \sigma_1^2)$ , (2) iid  $t(5)$ , (3) independent  $N\{0, \mu(t_j)\sigma_1^2\}$  and (4) iid  $\log N(0, \sigma_1^2)$ . With  $\sigma_1^2 = 1$ .
2. The scores set as  $\alpha_{1i} \sim N(0, 16)$  and  $\alpha_{2i} \sim N(0, 9)$  are both iid. The error term  $\varepsilon_{ni}$ 's is: (1) iid  $N(0, \sigma_2^2)$ , (2) iid  $t(5)$ , (3) independent  $N\{0, \mu(t_j)\sigma_2^2\}$  and (4) iid  $\log N(0, \sigma_2^2)$ . With  $\sigma_2^2 = 0.5$ .

For each of the parameter settings we run the simulations 200 times. These scenarios allow us to analyse the different coefficient-to-coefficient-to-noise variations as well as the scenarios for fat tail errors (scenario of  $\varepsilon_2$ ), heteroscedastic (scenario of  $\varepsilon_3$ ) and skewed errors (scenario of  $\varepsilon_4$ ). We analyse the performance for  $\tau = 0.5, 0.6, \dots, 0.9$  based on the mean squared error (MSE) and its standard deviation (SD). Summary of the recorded MSE for the simulations is given in Table 4, the standard deviations are given in the brackets.

Regarding the choice of basis in all scenarios we choose to use  $T/2$  B-spline curves for the space basis. The time basis was create as  $N/2$  curves of the fourier basis and 3 trend curves: linear, quadratic and logarithmic. One has to be aware that the choice of basis, i.e. number of basis functions can also have an impact on the results.

From the observed MSEs we conclude that on whenever the error distribution is skewed ,the model is likely to produce big MSEs. The model performs, in general, very well for different  $\tau$ -levels and comparable to the already mentioned alternatives proposed by Guo et al. (2013) or Tran et al. (2016). However, since there is a lack of the extreme observations, the MSEs increase with higher  $\tau$ -s.

$\tau$	$\varepsilon$	N=20, T=100		N=50, T=150	
		$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$
$\tau = 0.5$	$\varepsilon_1 \sim N(0, \sigma_\varepsilon^2)$	0.238 (0.011)	0.247 (0.012)	0.467 (0.023)	0.462 (0.013)
	$\varepsilon_2 \sim t(5)$	0.779 (0.053)	0.773 (0.028)	0.777 (0.054)	0.761 (0.028)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.439 (0.022)	0.443 (0.013)	0.868 (0.054)	0.853 (0.023)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	1.444 (0.058)	1.444 (0.028)	1.444 (0.592)	1.444 (0.271)
	$\varepsilon_1$	0.251 (0.013)	0.267 (0.018)	0.488 (0.024)	0.485 (0.016)
	$\varepsilon_2 \sim t(5)$	0.824 (0.059)	0.822 (0.034)	0.820 (0.059)	0.806 (0.033)
$\tau = 0.6$	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.461 (0.023)	0.469 (0.019)	0.906 (0.046)	0.888 (0.026)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	0.721 (0.097)	0.713 (0.051)	3.523 (1.055)	3.257 (0.501)
	$\varepsilon_1$	0.288 (0.011)	0.355 (0.012)	0.540 (0.023)	0.562 (0.013)
	$\varepsilon_2 \sim t(5)$	0.937 (0.079)	0.971 (0.066)	0.927 (0.079)	0.936 (0.055)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.517 (0.028)	0.568 (0.061)	0.995 (0.050)	0.989 (0.043)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	0.985 (0.148)	1.014 (0.096)	5.087 (1.655)	4.733 (0.800)
$\tau = 0.7$	$\varepsilon_1$	0.418 (0.073)	0.786 (0.365)	0.629 (0.053)	0.816 (0.204)
	$\varepsilon_2 \sim t(5)$	1.094 (0.108)	1.392 (0.320)	1.056 (0.103)	1.217 (0.187)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.641 (0.069)	0.972 (0.347)	1.093 (0.068)	1.228 (0.182)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	1.289 (0.205)	1.598 (0.367)	6.514 (2.204)	6.209 (1.089)
	$\varepsilon_1$	2.316 (1.247)	3.819 (1.648)	1.660 (0.689)	2.221 (0.779)
	$\varepsilon_2 \sim t(5)$	2.669 (1.069)	3.821 (1.527)	2.020 (0.675)	2.484 (0.751)
$\tau = 0.8$	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	2.349 (1.087)	3.719 (1.577)	1.989 (0.652)	2.428 (0.741)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	3.078 (1.154)	4.551 (1.542)	8.486 (2.627)	8.295 (1.519)
$\tau = 0.9$	$\varepsilon_1$	2.316 (1.247)	3.819 (1.648)	1.660 (0.689)	2.221 (0.779)
	$\varepsilon_2 \sim t(5)$	2.669 (1.069)	3.821 (1.527)	2.020 (0.675)	2.484 (0.751)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	2.349 (1.087)	3.719 (1.577)	1.989 (0.652)	2.428 (0.741)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	3.078 (1.154)	4.551 (1.542)	8.486 (2.627)	8.295 (1.519)
	$\varepsilon_1$	2.316 (1.247)	3.819 (1.648)	1.660 (0.689)	2.221 (0.779)
	$\varepsilon_2 \sim t(5)$	2.669 (1.069)	3.821 (1.527)	2.020 (0.675)	2.484 (0.751)

Table 1: Average MSE and its standard deviation in brackets by 200 simulation runs and different scenarios.

 DYTEC\_simulation\_all

## 5 Application for temperature curves

We apply our model to Chinese temperature data, which consists of daily average temperatures of 159 weather stations for the years 1957 to 2009. In this case  $n = 1, \dots, 53$  corresponds to the year and  $t = 1, \dots, 365$  corresponds to the day during the year. Model is applied for each station separately. It is obvious that while the factor loadings  $Z_{nk}$  vary over the years, the dependence within the year is captured by factors  $m_k(t)$  themselves.

### 5.1 The choice of time basis $U$

The proper time basis allows us to capture any periodic variation as well as any trend. In case of temperature data, we do not assume only linear trend, but also add quadratic and logarithmic function to the basis:

$$u_1(t) = \frac{t}{T}, \quad u_2(t) = \frac{t^2}{T^2}, \quad u_3(t) = \frac{\log t}{\log T}.$$

For the periodicity we use  $N - 1$  fourier basis functions with period  $N = 53$ :

$$\begin{aligned} u_4 &= \text{constant}, \\ u_5 &= \frac{1}{\sqrt{\frac{N}{2}}} \sin\left(\frac{2 \cdot \pi \cdot t}{N}\right), \\ u_6 &= \frac{1}{\sqrt{\frac{N}{2}}} \cos\left(\frac{2 \cdot \pi \cdot t}{N}\right), \\ &\dots \\ u_{52} &= \frac{1}{\sqrt{\frac{N}{2}}} \sin\left(\frac{\frac{(N-1)}{2} \cdot \pi \cdot t}{N}\right), \\ u_{53} &= \frac{1}{\sqrt{\frac{N}{2}}} \cos\left(\frac{\frac{(N-1)}{2} \cdot \pi \cdot t}{N}\right). \end{aligned}$$

In general, one may operate with various types of basis functions, such as higher-power polynomials, local polynomials, trigonometric or periodic functions, splines, etc., with regard to follow various types of non-linearity concerning the specific design of a given data.

## 5.2 The choice of space basis $\Psi$

In order to model the specific structure and pattern within the year we set the space basis  $\Psi$  as simple B-splines, particularly  $\frac{T-1}{2}$  functions of the order 5. One can also choose to use first few principal components explaining 85% of variance or even more complex Principal expectile components introduced by Tran et al. (2016) not to loose the specific information in tails.

## 5.3 Forecasting

One of the traditional approaches for the forecasting of functional data can be done via Karhunen-Loève expansion. The functional time series is rewritten via principal components and their scores are consequently modeled separately with an appropriate model. The forecast obtained by the model of the scores together with the original principal components are used for the prediction of functional time series. Similar approach can be done for expectiles, see e.g. López-Cabrera and Schulz (2017) who did a two-step approach. In the first step, for a fixed level of  $\tau$ , the series of expectiles curves is computed. Second, the principal component decomposition of the curves and the forecast of their scores is done via a vector auto-regressive model. However, one of the main restriction of such an approach is the assumption of the weak-dependent data.

The proposed DSFM model for expectile curves provides us with the easy method of forecasting. Since the matrix of coefficients  $\hat{C}$  is estimated once and the space basis is already predefined as well, one only needs to forecast the time basis. The basis consists of a set of functions, of which each can be

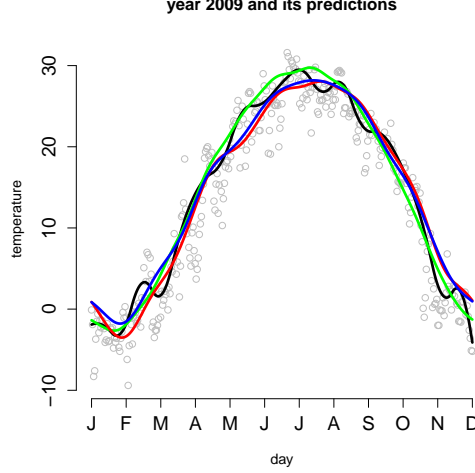



Figure 1: Chinese temperature data for year 2009 (grey) with its smoothed expectile curve (black). Predictions of expectile curve for level  $\tau = 0.8$  by DSFM model (green), VAR model (red) and ARIMA model (blue).

 DYTEC\_temperature

simply prolonged for the value in time  $N + 1$ . For each  $i = 1, \dots, J$  denote the values of prolonged basis function as  $\tilde{U}_i = (U_i(1), \dots, U_i(N), U_i(N + 1))$  and the new updated prolonged time basis as  $\tilde{U} = (\tilde{U}_1, \dots, \tilde{U}_J)$ . We obtain the forecast as:

$$\hat{e}_{N+1}(t) = (A\tilde{U})^\top (B\Psi) = \tilde{U}^\top \hat{C}\Psi.$$

With the aim to demonstrate the model we applied the proposed DSFM-model to Chinese temperature data sets. The daily observations for a specific station No.1. from 1957 to 2008 were used for the estimation of the model and matrix  $\hat{C}$ . Consequently the model with prolonged time basis was used for the prediction of an expectile curve for the upcoming year 2009. We used two other benchmark approaches to compare the quality of our forecast. First model, VAR-model, uses the principal components of the pre-computed expectiles curves from years 1957-2008 and forecasts their scores via VAR(4) model. The second model takes into consideration the possible

non-stationarity and thus uses the ARIMA model to forecast the score of each component separately. Figure 1 shows the data for year 2009, together with the expectile curve for fixed level  $\tau = 0.8$ . The DSFM-forecast (green) better predicts the expectile curve than VAR-model (red) or ARIMA-model (blue), which are constructed by forecasting the scores of principal components.

## 6 Application to Wind speed data

As a second application we use our DSFM-model for the modeling expectile curves of the wind speed of hurricanes in a hurricane season across the North Atlantic basin over the period 1965-2011. As earlier, the observed data has the form  $X_n(t_i)$ , where the times  $t_i$  are here separated by six hours, and the index  $n$  stands for year. The value  $X_n(t_i)$  is the wind speed in knots (1 kn = 0.5144 m/s). The data is accessible at the website of Unisys Weather Information, UNISYS (2015). We focus only on the hurricane-period from July till October, thus having  $T = 400$  observations for every year  $n = 1962, \dots, 2011$ , i.e.  $N = 50$ . We treat time  $0 \leq t \leq T$  within a year as continuous, and the observed curves as functional data.

Motivated by the work of Burdejova et al. (2017), who tested the hypothesis of linear trend for hurricanes we model the hurricane data with our DSFM-model and focus primarily on the estimation of coefficients for different trend curves in time basis.

### 6.1 The choice of time basis $U$

For the periodicity we use also 10 fourier basis functions with period  $N = 50$  and a constant function. Since we are mainly interested not only in linear trend, but also add quadratic and logarithmic function to the basis, so we set as before:

$$u_{12}(t) = \frac{\log t}{\log T}, \quad u_{13}(t) = \frac{t^2}{T^2}, \quad u_{14}(t) = \frac{t}{T}.$$



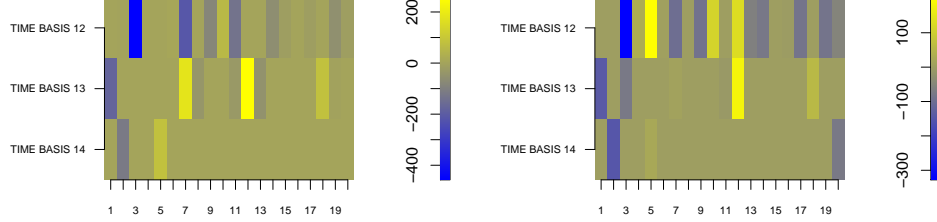


Figure 2: Graphical representation of  $\hat{C}$  for hurricane data and level  $\tau = 0.5$  (left) and  $\tau = 0.8$  (right). The rows correspond to the last three time basis functions, i.e. logarithmic, quadratic and linear. The columns correspond to the space basis, i.e. 20 principal components.

 data\_load\_hurricanes  
 DYTEC\_hurricanes

## 6.2 The choice of space basis $\Psi$

In order to model the specific structure within the yearly period we use the principal components. We set first 20 components as the space basis  $\Psi$ , since they explain 90% of the variance.

## 6.3 The estimates of matrix $\hat{C}$

The proposed algorithm is performed for two different  $\tau = 0.5$  and  $\tau = 0.8$ . The estimations of matrices are shown in Figure 2, for  $\tau = 0.5$  left and  $\tau = 0.8$  right. Two conclusions are obvious from the estimations:

1. The coefficient related to the linear trend for all principal component are not as much significant as the coefficients for quadratic and even logarithmic trend.
2. The linear trends has similar pattern for both  $\tau$ -levels. But this does not hold true for other two trends, whose coefficients related to all principal components differ with respect to  $\tau$ , especially for logarithmic trend.

One could conclude that in case of the historical observation of hurricanes, the question of testing models incorporating other than linear trend is lying in the interest of future research.

## 7 Conclusion

In this paper we propose the dynamic semi-parametric factor model for a joint estimation of expectile curves. For that purpose we utilized a non-parametric series expansion for both factors and their (time-developing) scores. We have provided the convergent algorithm for its estimation that is based on the idea of iterative least squares. The presented model is thus an utile extension of commonly known factor model for the mean, where  $L_2$  norm is used.

This novel approach provides us with several advantages. One can easily directly estimate the extreme curves from the data without any need of pre-computing the expectile curves separately. Moreover, the method may be applied for a non-stationary data as well. Any dynamics, hidden intra-dependencies, trend or patterns of such tail event curves can be easily captured with the proper choice of time basis.

We have demonstrated the good estimation properties in a simulation study for different set-ups of error term and different expectile  $\tau$ -levels as well. A method was applied to the Chinese temperature data set of average daily temperatures over years in order to show its easy usability not only for modelling but mainly for forecasting, where it performs as good as any traditional approaches used for the prediction of this type of functional data. The second application to the wind speed data of hurricanes shows not only the importance of considering various trends but also pointed out the fact of diverse factor structure for different  $\tau$ -levels.

## References

- BENKO, M., W. HÄRDLE, AND A. KNEIP (2009): “Common functional principal components,” *Ann. Statist.*, 37, 1–34.
- BORKE, L. AND W. K. HÄRDLE (2017a): “GitHub API based QuantNet Mining infrastructure in R,” *Discussion Paper SFB2017-08, Humboldt-Universität zu Berlin*.
- (2017b): “Q3-D3-LSA,” in *Handbook of Big data Analytics*, Springer Verlag.
- BRECKLING, J. AND R. CHAMBERS (1988): “M-quantiles,” *Biometrika*, 75, 761–771.
- BURDEJOVA, P., W. HÄRDLE, P. KOKOSZKA, AND Q. XIONG (2017): “Change point and trend analyses of annual expectile curves of tropical storms,” *Econometrics and Statistics*, 1, 101 – 117.
- CHOROS-TOMCZYK, B., W. K. HÄRDLE, AND O. OKHRIN (2016): “A semiparametric factor model for CDO surfaces dynamics,” *Journal of Multivariate Analysis*, 146, 151 – 163, special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.
- ERBAS, B., R. J. HYNDMAN, AND D. M. GERTIG (2007): “Forecasting age-specific breast cancer mortality using functional data models,” *Statistics in Medicine*, 26, 458–470.
- FERRATY, F. AND P. VIEU. (2006): *Nonparametric Functional Data analysis; Theory and Practice*, Springer.
- GUO, M., L. ZHOU, W. HÄRDLE, AND J. HUANG (2013): “Functional Data Analysis for Generalized Quantile Regression,” *Statistics and Computing*, 1–14.

- HÄRDLE, W. K. AND P. MAJER (2016): “Yield curve modeling and forecasting using semiparametric factor dynamics,” *The European Journal of Finance*, 22, 1109–1129.
- HÖRMANN, S. AND P. KOKOSZKA (2010): “Weakly dependent functional data,” *Ann. Statist.*, 38, 1845–1884.
- HORVÁTH, L. AND P. KOKOSZKA (2012): *Inference for Functional Data with Applications*, Springer.
- IGNACCOLO, R., S. GHIGO, AND E. GIOVENALI (2008): “Analysis of air quality monitoring networks by functional clustering,” *Environmetrics*, 19, 672–686.
- JONES, M. C. (1994): “Expectiles and M-quantiles are Quantiles,” *Statistics & Probability Letters*, 20, 149–153.
- KOKOSZKA, P., H. MIAO, AND X. ZHANG (2014): “Functional Dynamic Factor Model for Intraday Price Curves,” *Journal of Financial Econometrics*, 1–22.
- LEE, R. D. AND L. R. CARTER (1992): “Modeling and Forecasting U. S. Mortality,” *Journal of the American Statistical Association*, 87, 659–671.
- LÓPEZ-CABRERA, B. AND F. SCHULZ (2017): “Forecasting Generalized Quantiles of Electricity Demand: A Functional Data Approach,” *Journal of the American Statistical Association*, 112, 127–136.
- NELSON, C. R. AND A. F. SIEGEL (1987): “Parsimonious Modeling of Yield Curves,” *The Journal of Business*, 60, 473–489.
- NEWHEY, W. AND J. POWELL (1987): “Asymmetric least squares estimation and testing,” *Econometrica*, 819–847.

- PARK, B. U., E. MAMMEN, W. HÄRDLE, AND S. BORAK (2009): “Time series modelling with semiparametric factor dynamics,” *Journal of the American Statistical Association*, 104, 284–298.
- RAMSAY, J. AND B. SILVERMAN (2005): *Functional data analysis*, Springer, New York.
- SCHNABEL, S. K. AND P. H. EILERS (2009): “Optimal expectile smoothing,” *Computational Statistics & Data Analysis*, 53, 4168 – 4177.
- SONG, S., W. K. HÄRDLE, AND Y. RITOV (2014): “Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series,” *The Econometrics Journal*, 17, S101–S131.
- TRAN, N. M., P. BURDEJOVA, M. OSIPENKO, AND W. K. HÄRDLE (2016): “Principal Component Analysis in an Asymmetric Norm,” *Discussion Paper SFB2016-40, Humboldt-Universität zu Berlin*.
- UNISYS (2015): “Data in Atlantic and West Pacific,” Unisys Weather Information Systems, <http://weather.unisys.com/hurricane/index.php>, Accessed: February 20, 2015.
- YANG, Y. AND H. ZOU (2015): “A fast unified algorithm for solving group-lasso penalized learning problems,” *Statistics and Computing*, 25, 1129–1141.

# SFB 649 Discussion Paper Series 2017

For a complete list of Discussion Papers published by the SFB 649, please visit <http://sfb649.wiwi.hu-berlin.de>.

- 001 "Fake Alpha" by Marcel Müller, Tobias Rosenberger and Marliese Uhrig-Homburg, January 2017.
- 002 "Estimating location values of agricultural land" by Georg Helbing, Zhiwei Shen, Martin Odening and Matthias Ritter, January 2017.
- 003 "FRM: a Financial Risk Meter based on penalizing tail events occurrence" by Lining Yu, Wolfgang Karl Härdle, Lukas Borke and Thijs Benschop, January 2017.
- 004 "Tail event driven networks of SIFIs" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle and Yarema Okhrin, January 2017.
- 005 "Dynamic Valuation of Weather Derivatives under Default Risk" by Wolfgang Karl Härdle and Maria Osipenko, February 2017.
- 006 "RiskAnalytics: an R package for real time processing of Nasdaq and Yahoo finance data and parallelized quantile lasso regression methods" by Lukas Borke, February 2017.
- 007 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, February 2017.
- 008 "GitHub API based QuantNet Mining infrastructure in R" by Lukas Borke and Wolfgang K. Härdle, February 2017.
- 009 "The Economics of German Unification after Twenty-five Years: Lessons for Korea" by Michael C. Burda and Mark Weder, April 2017.
- 010 "Data Science & Digital Society" by Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, May 2017.
- 011 "The impact of news on US household inflation expectations" by Shih-Kang Chao, Wolfgang Karl Härdle, Jeffrey Sheen, Stefan Trück and Ben Zhe Wang, May 2017.
- 012 "Industry Interdependency Dynamics in a Network Context" by Ya Qian, Wolfgang Karl Härdle and Cathy Yi-Hsuan Chen, May 2017.
- 013 "Adaptive weights clustering of research papers" by Larisa Adamyan, Kirill Efimov, Cathy Yi-Hsuan Chen, Wolfgang K. Härdle, July 2017.
- 014 "Investing with cryptocurrencies - A liquidity constrained investment approach" by Simon Trimborn, Mingyang Li and Wolfgang Karl Härdle, July 2017.
- 015 "(Un)expected Monetary Policy Shocks and Term Premia" by Martin Kliem and Alexander Meyer-Gohde, July 2017.
- 016 "Conditional moment restrictions and the role of density information in estimated structural models" by Andreas Tryphonides, July 2017.
- 017 "Generalized Entropy and Model Uncertainty" by Alexander Meyer-Gohde, August 2017.
- 018 "Social Security Contributions and the Business Cycle" by Anna Almosova, Michael C. Burda and Simon Voigts, August 2017.
- 019 "Racial/Ethnic Differences In Non-Work At Work" by Daniel S. Hamermesh, Katie R. Genadek and Michael C. Burda, August 2017.
- 020 "Pricing Green Financial Products" by Awdesch Melzer, Wolfgang K. Härdle and Brenda López Cabrera, August 2017.
- 021 "The systemic risk of central SIFIs" by Cathy Yi-Hsuan Chen and Sergey Nasekin, August 2017.
- 022 "Das deutsche Arbeitsmarktwunder: Eine Bilanz" by Michael C. Burda and Stefanie Seele, August 2017.

**SFB 649, Spandauer Straße 1, D-10178 Berlin**  
**<http://sfb649.wiwi.hu-berlin.de>**

This research was supported by the Deutsche  
Forschungsgemeinschaft through the SFB 649 "Economic Risk".



# SFB 649 Discussion Paper Series 2017

For a complete list of Discussion Papers published by the SFB 649, please visit <http://sfb649.wiwi.hu-berlin.de>.

- 023 "Penalized Adaptive Method in Forecasting with Large Information Set and Structure Change" by Xinjue Li, Lenka Zbonakova and Wolfgang Karl Härdle, September 2017.
- 024 "Smooth Principal Component Analysis for High Dimensional Data" by Yingxing Li, Wolfgang K. Härdle and Chen Huang, September 2017.
- 025 "Realized volatility of CO2 futures" by Thijs Benschop and Brenda López Cabrera, September 2017.
- 026 "Dynamic Semiparametric Factor Model with a Common Break" by Likai Chen, Weining Wang and Wei Biao Wu, November 2017.
- 027 "Dynamic semi-parametric factor model for functional expectiles" by Petra Burdejová and Wolfgang K. Härdle, November 2017.

**SFB 649, Spandauer Straße 1, D-10178 Berlin**  
**<http://sfb649.wiwi.hu-berlin.de>**

This research was supported by the Deutsche  
Forschungsgemeinschaft through the SFB 649 "Economic Risk".

