Mathematisches Forschungsinstitut Oberwolfach

Report No. 48/2013

DOI: 10.4171/OWR/2013/48

Statistical Inference for Complex Time Series Data

Organised by
Rainer Dahlhaus, Heidelberg
Oliver Linton, Cambridge
Wei-Biao Wu, Chicago
Qiwei Yao, London

22 September – 28 September 2013

Abstract. During recent years the focus of scientific interest has turned from low dimensional stationary time series to nonstationary time series and high dimensional time series. In addition new methodological challenges are coming from high frequency finance where data are recorded and analyzed on a millisecond basis. The three topics "nonstationarity", "high dimensionality" and "high frequency" are on the forefront of present research in time series analysis. The topics also have some overlap in that there already exists work on the intersection of these three topics, e.g. on locally stationary diffusion models, on high dimensional covariance matrices for high frequency data, or on multivariate dynamic factor models for nonstationary processes. The aim of the workshop was to bring together researchers from time series analysis, nonparametric statistics, econometrics and empirical finance to work on these topics. This aim was successfully achieved and the workshops was very well attended.

Mathematics Subject Classification (2010): 62M10.

Introduction by the Organisers

The workshop *Statistical Inference for Complex Time Series Data*, organised by Rainer Dahlhaus (Heidelberg), Oliver Linton (Cambridge), Wei-Biao Wu (Chicago) and Qiwei Yao (London), was held in 22-28 September 2013. The workshop was well attended with 51 participants with broad geographic representation from Europe, Australia, Canada and USA. The participants formed a nice blend of researchers with various backgrounds including statistics, probability, machine

learning and econometrics. A considerably large proportion of the participants were early career academics, post-doctoral researchers and some PhD.

29 talks of varying lengths were delivered during the five days. The talks were given by both leading experts in the field as well as by up-coming young scientists. In addition, there were seven 10 minute sessions with title "People and Topics" which featured short presentations on ongoing research projects and brief introductions on themselves of young researchers. Participants found those short presentations informative and effective.

There were several major themes in the various sessions, including local stationary time series models, high-dimensional modeling, high-frequency data, volatility estimation in finance, change-point detection for dependent data, and GARCH models. Overall the meeting generated a great deal of discussion and often smaller groups of people met in the evenings for additional spontaneous lectures and detailed discussions. A number of important research contacts were made which we expect to stimulate new collaborative research projects.

In addition to the excellent scientific exchanges, the traditional Wednesday afternoon hike was blessed by excellent weather and delicious black-forest cake. It is important to note that this social event also has a high impact on scientific exchange and on stimulating new collaborative research. Those participants who had never visited Oberwolfach before, left with a clear impression on the MFO and its high valued contribution to the global mathematical community. There was also a strong consensus that the theme "Complex Time Series" should appear more regularly in the Oberwolfach workshop program to reflect the rapid development in mathematics and statistics driven by this information age.

Workshop: Statistical Inference for Complex Time Series Data

Table of Contents

Abstracts

The Marčenko–Pastur Law for Time Series

Alexander Aue
(joint work with Haoyang Liu, Debashis Paul)

Spectra of high-dimensional covariance matrices appear in portfolio choice problems in finance, determining the channel capacity in wireless communications and modeling the highly excited states in nuclear physics, among other applications. In this talk, we consider high-dimensional zero mean linear time series of the form $X_t = \sum_{\ell=0}^{\infty} \mathbf{A}_\ell Z_{t-\ell}$, where $(Z_t \colon t \in \mathbb{Z})$ is a sequence of p-dimensional real or complex-valued random vectors with independent, zero mean, unit variance entries, and $(\mathbf{A}_\ell \colon \ell \in \mathbb{N}_0)$ are symmetric and simultaneously diagonalizable $p \times p$ coefficient matrices. Assuming that $X_1, \ldots, X_n$ are observed, we consider the asymptotic framework where $p \to \infty$, $n \to \infty$ while $p/n \to c > 0$. The sample covariance matrix is defined as $S = \frac{1}{n} \sum_{t=1}^{n} X_t X_t^*$. We also study symmetrized autocovariance matrices which, for lags $\tau \in \mathbb{N}_0$, are defined as

$$\mathbf{C}_\tau = \frac{1}{2n} \sum_{t=1}^{n-\tau} (X_t X_{t+\tau}^* + X_{t+\tau} X_t^*).$$

When $\tau = 0$, C_τ reduces to S. The empirical spectral distribution (ESD) $F^{\mathbf{S}}$ of $\mathbf{S}$ is given by $F^{\mathbf{S}}(\lambda) = \frac{1}{p} \sum_{j=1}^{p} 1_{\{\lambda_j \le \lambda\}}$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ are the eigenvalues of $\mathbf{S}$. When p/n stays away from 0, as $n \to \infty$, the ESD of $\mathbf{S}$ is away from the spectrum of the population covariance matrix even for large p. Deriving the limiting ESD for $\mathbf{S}$ when $p \to \infty$, $n \to \infty$ while $p/n \to c > 0$ is fundamental to understanding the correlation structure of a large number of variables using finite samples. There have been works on the limiting ESD of the sample covariance matrix for different models of X_t. The classical Marčenko–Pastur law established in [2] is concerned with i.i.d. X_t without temporal or dimensional dependence. [3], among others, considered random vectors of the form $X_t = AZ_t$ allowing for dependence across rows but not time. Recently, [1], [4] and [5] considered p-dimensional time series with independent and identically distributed rows. Our work goes beyond the existing contributions and is the first one to derive the limiting behavior of both sample covariance and symmetrized autocovariance matrices. The large-sample results for the symmetrized autocovariance matrices can be particularly useful for constructing high-dimensional diagnostic tests for the presence of correlation, setting up high-dimensional prediction equations, determining the order of a high-dimensional time series and estimating the spectrum of the population autocovariance matrices. Indeed, a computational method is developed to estimate the spectra of the coefficients of low order moving average processes by using the results presented here.

The main result (Theorem 1 below) involves the derivation of a system of nonlinear equations for the Stieltjes transform of the probability distribution representing the limiting ESD of the sample covariance matrix and the lag-τ symmetrized sample autocovariance matrices. The Stieltjes transform of a probability distribution P on $\mathbb{R}$ is defined as

$$s_P(z) = \int \frac{dP(\lambda)}{\lambda - z}, \qquad \lambda \in \mathbb{C}^+ = \{u + iv \colon v > 0\}.$$

Under mild technical conditions, pointwise convergence of the Stieltjes transforms for a sequence of probability distributions ensures the weak convergence of the sequence. For simplicity of the exposition, Theorem 1 is stated for MA(1) processes. In order to state the result, define $h(\lambda, \nu) = 1 + 2\cos(\nu)\lambda + \lambda^2$, where $\nu \in [0, 2\pi]$, the spectrum (up to normalization) of the univariate MA(1) process $x_t = z_t + \lambda z_{t-1}$, $t \in \mathbb{Z}$, where $\lambda \in \mathbb{R}$ and $(z_t \colon t \in \mathbb{Z})$ a univariate white noise process.

Theorem 1. *Suppose that $(X_t \colon t \in \mathbb{Z})$ satisfies the following assumptions:*

(a) *n is assumed to be a function of p, that is, $c_p = p/n(p) \to c > 0$ as $p \to \infty$.*
(b) *$Z_t = [Z_{1t}, \ldots, Z_{pt}]' \in \mathbb{C}^p$ with i.i.d. Z_{jt}. If the observations are complex-valued, it is assumed that $\mathbb{E}[Z_{11}] = 0$, $\mathbb{E}[|\Re(Z_{11})|^2] = 1/2$, $\mathbb{E}[|\Im(Z_{11})|^2] = 1/2$ and $\mathbb{E}[|Z_{11}|^4] < \infty$, and real and imaginary parts are assumed to be independent. If the observations are real-valued, $\mathbb{E}[Z_{11}] = 0$, $\mathbb{E}[|Z_{11}|^2] = 1$ and $\mathbb{E}[|Z_{11}|^4] < \infty$.*
(c) *$\mathbf{A}_1$ is a $p \times p$ Hermitian (symmetric) matrix, independent of the Z_t, and with uniformly bounded eigenvalues.*
(d) *Almost surely, the ESD of $\mathbf{A}_1$ converges weakly to a nonrandom probability distribution function $F^{\mathbf{A}_1}$ as $p \to \infty$.*

Then, almost surely, the ESD of C_τ, $F^{\mathbf{C}_\tau}$, converges weakly to a probability distribution F_τ with Stieltjes transform $s_\tau(z)$, determined by the equation

$$s_\tau(z) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{\cos(\tau\nu) h(\lambda, \nu)}{1 + c\cos(\tau\nu) K_\tau(z, \nu)} d\nu - z \right] dF^{\mathbf{A}_1}(\lambda), \tag{1}$$

where $K_\tau(z, \nu)$, for $z \in \mathbb{C}^+$ and $\nu \in [0, 2\pi]$, is a Stieltjes kernel, i.e., for every $\nu \in [0, 2\pi]$, $K_\tau(z, \nu)$ is the Stieltjes transform of a measure with total mass $\int h(\lambda, \nu) dF^{\mathbf{A}_1}(\lambda)$. Moreover, $K_\tau(z, \nu)$ is the unique solution to

$$K_\tau(z, \nu) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{\cos(\tau\nu') h(\lambda, \nu')}{j1 + c\cos(\tau\nu') K_\tau(z, \nu')} d\nu' - z \right]^{-1} h(\lambda, \nu) dF^{\mathbf{A}_1}(\lambda), \tag{2}$$

subject to the restriction that $K_\tau(z, \nu)$ is a Stieltjes kernel.

Theorem 1 has been extended to MA(∞) processes for which the coefficient matrices $(\mathbf{A}_\ell \colon \ell \in \mathbb{N}_0)$ are simultaneously diagonalizable in a unitary or orthogonal basis, and $\sum_{\ell=0}^{\infty} \ell \|\mathbf{A}_\ell\| < \infty$. For MA($\infty$) processes, the systems of equations are still given by (1) and (2), but $h(\lambda, \nu)$ is now related to the spectrum of a univariate MA(∞) process, namely, $h(\lambda, \nu) = |\sum_{\ell=0}^{\infty} e^{i\ell\nu} f_\ell(\lambda)|$, where it is further assumed that the jth eigenvalue of A_ℓ equals $f_\ell(\lambda'_j)$, for continuous functions f_ℓ, where

$\lambda'_j \in \mathbb{R}^d$ for some $d \geq 1$, for each j. In addition, $F^{\mathbf{A}_1}$ is replaced by a distribution on $\mathbb{R}^d$ that describes the limiting joint ESD of the coefficient matrices. It is to be noted that this class contains the class of causal invertible ARMA(q, r) processes whose coefficient matrices are simultaneously diagonalizable and the corresponding linear process representations satisfy the requirement $\sum_{\ell=0}^{\infty} \ell \|\mathbf{A}_\ell\| < \infty$.

References

[1] B. Jin, C. Wang, B. Miao and M.-N. Lo Huang, *Limiting spectral distribution of large-dimensional sample covariance matrices generated by VARMA*, Journal of Multivariate Analysis **100** (2009), 2112–2125.

[2] V. Marčenko and L. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mathematics of the USSR-Sbornik **1** (1967), 457–483.

[3] J.W. Silverstein and Z.D. Bai, *On the empirical distribution of eigenvalues of a class of large dimensional random matrices*, Journal of Multivariate Analysis **54** (1995), 175–192.

[4] O. Pfaffel and E. Schlemm, *Eigenvalue distribution of large sample covariance matrices of linear processes*, arXiv:1201.3828 (2012).

[5] J.-F. Yao, *A note on a Marčenko–Pastur type theorem for time series*, Statistics & Probability Letters **82** (2012), 22–28.

Largest eigenvalues of the sample covariance matrix for p-variate time series with heavy-tails

Richard A. Davis

(joint work with Thomas Mikosch, Oliver Pfaffel)

In the classical multivariate statistics or time series setting, the data consist of n observations of p-dimensional random vectors, where p is relatively small compared to the sample size n. With the recent advent of large data sets, the dimension p can be large relative to the sample size and hence standard asymptotics, assuming p is fixed relative to n may provide misleading results. Structure in multivariate data is often summarized by the sample covariance matrix. For example, principal component analysis, extracts principal component vectors corresponding to the largest eigenvalues. Consequently, there is a need to study asymptotics of the largest eigenvalues of the sample covariance matrix. In the case of p fixed and the $p \times n$ data matrix consists of iid N(0,1) observations, Anderson [1] showed that the largest eigenvalue is asymptotically normal. In a now seminal paper, Johnstone [6] showed that if $p_n \to \infty$ at the rate $p_n/n \to \gamma \in (0, \infty)$, then the largest eigenvalues, suitable normalized, converges to the Tracey-Widom distribution with $\beta = 1$. Johnston's result has been generalized by Tao and Vu [8] where only 4 moments are needed to determine the limit. The theory for the largest eigenvalues of sample covariance and Wigner matrices based on heavy tails is not as well developed as in the light tailed case. The largest eigenvalues of sample covariance matrices with iid entries that are regularly varying with index $-\alpha$ were studied by Soshnikov [7] for the $\alpha \in (0, 2)$ case and subsequently extended in Auffinger et el. [2] to the $\alpha \in (2, 4)$ case. They showed that the point process of eigenvalues, normalized by the square of the $1 - (np)^{-1}$ quantile converges in distribution to

a Poisson point process with intensity $(\alpha/2)x^{-\alpha/2-1}$, provided $p/n \to \gamma$, where $\gamma \in (0,1)$. These results were extended in Davis et al. [4] to the case where the rows of the data matrix are iid linear heavy-tailed processes. They also had more general growth conditions on p_n in the case of iid entries and $\alpha \in (0,2)$.

In this paper, we study the asymptotic behavior of the largest eigenvalues of the sample covariance matrices of a multivariate time series. The time series is assumed to be heavy-tailed and linearly dependent in time and between the components. This generalizes and extends Davis et al. [4], who consider multivariate time series with heavy tails, but the component time series were assumed to be iid copies of a linear time series. We show that allowing dependence between the rows can appreciable impact the limit behavior of the largest eigenvalues. Instead of obtaining a Poisson point process as the limit of the extreme eigenvalues, we now get a "cluster" Poisson point process. That is, the limit can be described by a Poisson point process in which each point produces a "cluster" of points. Interestingly, the limit point process is identical to the limit point process derived by Davis and Resnick [3] for the extremes of a linear process.

To make the model precise, consider a double array of iid random variables $(Z_{it})_{i,t\in\mathbb{Z}}$, a double array of real numbers $(h_{kl})_{k,l\in\mathbb{Z}}$ and construct an infinite-dimensional time series,

$$X_{it} = \sum_{l=0}^{\infty}\sum_{k=0}^{\infty} h_{kl} Z_{i-k,t-l}\,, \quad i,t \in \mathbb{Z}\,. \tag{1}$$

We also assume that a generic element Z of the Z-field satisfies the regularly varying and tail balance condition

$$P(Z > x) \sim p_+ \frac{L(x)}{x^\alpha} \quad \text{and} \quad P(Z \le -x) \sim p_- \frac{L(x)}{x^\alpha}\,, \quad x \to \infty\,,$$

for some tail index $\alpha > 0$, where $p_+, p_- \ge 0$ with $p_+ + p_- = 1$ and L is a slowly varying function . To ensure the a.s. absolute convergence of the series (1) we will need further conditions on (h_{kl}) to be discussed later.

Consider the $p \times n$ data matrix

$$\mathbf{X}_n = (X_{it})_{i=1,\dots,p,t=1,\dots,n}\,, \quad n \ge 1\,,$$

where $p = p_n$ is an integer sequence such that $p_n \to \infty$.

The main focus of study in this paper is the asymptotic behavior of the eigenvalues n times the sample covariance matrix $\mathbf{X}_n\mathbf{X}_n'$ in the case $\alpha \in (0,2)$ and its centered version $\mathbf{X}_n\mathbf{X}_n' - E\mathbf{X}_n\mathbf{X}_n'$ in the case $\alpha \in (2,4)$. Our main result, yields an approximation for the sequence of the order statistics of the sample covariance matrices, showing that the largest eigenvalues of these matrices are to a large extent determined by the order statistics of the vector $D_1, \dots, D_p$, where, for $n \ge 1$, we define the iid sequence

$$D_s = D_s^{(n)} = \sum_{t=1}^{n} Z_{st}^2\,, \quad s \in \mathbb{Z},.$$

A consequence of this approximation is the point process convergence of the normalized eigenvalues of the sample covariance matrices. Based on the point process convergence , the continuous mapping theorem yields a variety of asymptotic results for the largest eigenvalues of the sample covariance matrix as well as joint limit theory for the trace and the largest eigenvalue. In particular, we show that the ratio of the largest eigenvalue to their sum converges in distribution to the ratio of a max-stable to a sum-stable random variable. In the special case when the filter (h_{kl}) is separable, $h_{kl} = \theta_k c_l$, the limit ratio does not depend on the filter weights (θ_k), (c_l). As a further special case, if the time series consists of iid vectors with linear dependence between the components the limit behavior of the eigenvalues is the same as that for iid components.

References

[1] Anderson, T.W., (1963) *Asymptotic theory for principal component analysis. Ann. Math. Statist.* **34**, 122–148.

[2] Auffinger, A., Ben Arous, G. and Péché, S. (2009) Poisson convergence for the largest eigenvalies of heavy tailed random matrices. *Ann. Inst. H. Poincaré, Probab. Statist.* **45**, 589–610.

[3] Davis, R.A. and Resnick, S.I. (1985) Limit theory for moving averages of random variables with regularly varying tail probabilities. *Ann. Probab.* **13**, 179–195.

[4] Davis, R.A., Pfaffel, O. and Stelzer, R. (2014) Limit theory for the largest eigenvalues of sample covariance matrices with heavy tails. *Stoch. Proc. Appl.* **124**, 18–50.

[5]

[6] Johnstone, I.M. (2001) *on the distribution of the largest eigenvalue in principal component analysis. Ann. Statist.* **29** (2), 295–327.

[7] Soshnikov, A. (2006) Poisson statistics for the largest eigenvalues in random matrix ensembles. In: *Mathematical Physics of Quantum Mechanics.* Lecture Notes in Physics **690**. Springer, Berlin, pp. 351–364.

[8] Tao, T. and Vu, V. (2012) Random covariance matrices; universality of local statistics of eigenvalues. *Ann. Probab.* **40** (3), 1285–1315.

Asymptotic distribution of some robust change-point tests for time series

Herold Dehling

(joint work with Roland Fried, Aeneas Rooch, Murad Taqqu, Martin Wendler)

1. Introduction. We study robust change-point tests for time series, and derive their asymptotic distribution, both in the short range as well as in the long-range dependent case. We study the model where the data are generated by $X_i = \mu_i + \epsilon_i$, where μ_i is an unknown signal, and where ϵ_i is a stationary ergodic noise process with $E(\epsilon_i) = 0$. Given the data $X_1, \ldots, X_n$, we wish to test the hypothesis H : $\mu_1 = \ldots = \mu_n$ against the alternative

$$A : \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n, \text{ for some } 1 \leq k \leq n-1.$$

In what follows, we will treat both the case of short range dependent (SRD) as well as long range dependent (LRD) noise.

Many change point tests can be derived from the two-sample problem that one obtains when the change point k is known. In this case, we have two samples $X_1, \ldots, X_k$ and $X_{k+1}, \ldots, X_n$, where we want to test for a difference in location. We consider three tests, namely the Gauss test, the Wilcoxon test and the Hodges-Lehmann test, which are associated with the (non-normalized) test statistics

$$\frac{1}{n-k}\sum_{i=k+1}^{n} X_i - \frac{1}{k}\sum_{i=1}^{k} X_i$$

$$\sum_{i=1}^{k}\sum_{j=k+1}^{n} 1_{\{X_i \le X_j\}}$$

$$\mathrm{median}\{(\mathrm{X_j} - \mathrm{X_i}) : 1 \le \mathrm{i} \le \mathrm{k}, \mathrm{k}+1 \le \mathrm{j} \le \mathrm{n}\}$$

After a proper normalization, the first two statistics are special cases of a two-sample U-statistic

$$U_{k,n-k} = \frac{1}{k(n-k)}\sum_{i=1}^{k}\sum_{j=k+1}^{n} h(X_i, X_j).$$

The Hodges-Lehmann statistic is the median of the empirical distribution of the pairwise differences $X_j - X_i, 1 \le i \le k < j \le n$. More generally, we can study the empirical distribution and the quantiles of $g(X_i, X_j), 1 \le i \le k < j \le n$, defining

$$U_{k,n-k}(t) = \frac{1}{k(n-k)}\sum_{i=1}^{k}\sum_{j=k+1}^{n} 1_{\{g(X_i,X_j)\le t\}},$$

and the quantile function $Q_{k,n-k}(p) = U^{-1}_{k,n-k}(p)$, where U^{-1} denotes the generalized inverse.

In the case of an unknown change point, one takes some summary statistics, e.g. the maximum over all $1 \le k \le n$. In order to derive the corresponding asymptotic distribution, we study convergence of the processes obtained by replacing k by $[n\lambda]$, $0 \le \lambda \le 1$. In this way, e.g., we obtain the two-sample U-statistic process $(U_{[n\lambda],n-[n\lambda]})_{0\le\lambda\le1}$, and the two-sample quantile process $(Q_{[n\lambda],n-[n\lambda]}(p))_{0\le\lambda\le1}$.

2. Short Range and Long Range Dependence. We have obtained results under the assumption of Short Range Dependent (SRD) as well as of Long Range Dependent (LRD) noise. In the SRD case, we assume that the noise has a representation as a functional of a β-mixing process $(Z_i)_{i\in\mathbb{Z}}$, i.e. that $\epsilon_i = f(Z_i, Z_{i-1}, \ldots)$, where f is a Lipschitz continuous function. Specific results require further technical assumptions on the rate of decay of the β-mixing coefficient and on the continuity of the function f. In the LRD case, we consider Gaussian subordinated processes, i.e. we assume that $\epsilon_i = H(\xi_i)$, where $(\xi_i)_{i\ge1}$ is a stationary Gaussian process with standard normal marginals and autocorrelation function $\rho_k = k^{-D}L(k)$, $0 < D < 1$, and where H is a measurable function.

3. Two-sample U-processes for SRD data. Consider the Hoeffding decomposition of the kernel $h(x,y) = \theta + h_1(x) + h_2(y) + \psi(x,y)$, where $\theta = Eh(X,Y)$, $h_1(x) = Eh(x,Y) - \theta$, $h_2(y) = Eh(X,y) - \theta$ and $\psi(x,y) = h(x,y) - \theta - h_1(x) - h_2(y)$, and where X, Y are independent with the same distribution as X_1.

Theorem 1 (Dehling, Fried, Garcia, Wendler 2013) *Under some technical conditions, concerning the β-mixing coefficients and the continuity of f and h, and under the null hypothesis of no change,*

$$\sqrt{n}\lambda(1-\lambda)(U_{[n\lambda],n-[n\lambda]}-\theta)_{0\le\lambda\le 1} \xrightarrow{\mathcal{D}} ((1-\lambda)W_1(\lambda)+\lambda(W_2(1)-W_2(\lambda)))_{0\le\lambda\le 1},$$

where (W_1, W_2) is 2-dimensional Brownian motion with covariance structure

$$E(W_i(\lambda)W_j(\lambda)) = (\lambda\wedge\mu)\sum_{k\in\mathbb{Z}}\mathrm{Cov}(\mathrm{h_i}(\mathrm{X_0}),\mathrm{h_j}(\mathrm{X_k})),$$

for $i, j \in \{1, 2\}$ and $0 \le \lambda, \mu \le 1$.

The proof of this theorem uses the Hoeffding decomposition. The crucial part of the proof is to show that the remainder term $\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\psi(X_i, X_j)$ is small, uniformly in λ. This is achieved by generalized correlation inequalities.

4. Two-sample U-processes for LRD data. Let $\epsilon_i = H(\xi_i)$, $i \ge 1$, be a Gaussian subordinated process, and define $J_k(x) = E(1_{\{H(\xi)\le x\}}H_k(\xi))$, where H_k is the k-th order Hermite polynomial. The smallest integer m such that $J_m(x) \not\equiv 0$ is called the Hermite rank. Define the normalizing constants $d_n = \mathrm{Var}(\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{H_m}(\xi_\mathrm{i}))$, and recall that $d_n \sim cn^{2H}L^m(n)$, where $H := 1 - mD/2$ is the Hurst coefficient.

Theorem 2 (Dehling, Rooch, Taqqu 2013a) *Let $m < 1/D$. Then, under the null hypothesis of no change,*

$$\frac{1}{d_n}\sum_{i=1}^{[n\lambda]}\sum_{j=[n\lambda]+1}^{n}\left(1_{\{X_i\le X_j\}}-\frac{1}{2}\right) \to \frac{\int J_m(x)dF(x)}{m!}(Z_m(\lambda)-\lambda Z_m(1)),$$

where $(Z_m(\lambda))_{0\le\lambda\le 1}$ denotes an m-th order Hermite process.

Dehling, Rooch and Taqqu (2013b) have investigated the asymptotic distribution of the Wilcoxon and the CUSUM change point test under local alternatives and calculated their asymptotic relative efficiencies. In the case of Gaussian errors, the ARE equals 1, while for heavy-tailed data, the Wilcoxon test has superior power. For finite samples, these results are confirmed by simulations.

Rooch (2012) has studied the asymptotic behavior of the U-statistics process of LRD data for arbitrary kernels $h(x, y)$, using two different techniques, namely an empirical process representation of the two-sample U-statistic, and a bivariate Hermite expansion of the kernel.

5. Hodges-Lehmann change-point test for SRD data. We study the two-sample U-statistic process with kernel $h(x, y; t) = 1_{\{g(x,y)\le t\}}$, indexed by $t \in \mathbb{R}$, and the associated quantile process, and define $U(t) = P(g(X, Y) \le t)$, for X and Y independent with the same distribution as X_0. Moreover, we define the quantile function $Q(p) = U^{-1}(p)$. We denote the terms of the Hoeffding decomposition of $h(x, y; t)$ by $h_1(x; t)$ and $h_2(y; t)$, and let $u(t) = U'(t)$.

Theorem 3 (Dehling, Fried, Wendler 2013) *Under some technical conditions, concerning the β-mixing coefficients and the continuity of f and g, and under the null hypothesis of no change, the two-sample quantile process*

$$\sqrt{n}\lambda(1-\lambda)\left(Q_{[n\lambda],n-[n\lambda]}(p)-Q(p)\right)_{0\le\lambda\le 1}$$

converges in distribution to the process $((1-\lambda)W_1(x)+\lambda(W_2(1)-W_2(\lambda))_{0\le\lambda\le 1}$, where $(W_1(\lambda),W_2(\lambda))$ is 2-dimensional Brownian motion with covariance function

$$\mathrm{Cov}(\mathrm{W_i}(\mu),\mathrm{W_j}(\lambda))=\frac{\mu\wedge\lambda}{\mathrm{u^2(Q(p))}}\sum_{\mathrm{k}\in\mathbb{Z}}\mathrm{E}(\mathrm{h_i}(\mathrm{X_0},\mathrm{Q(p)})\mathrm{h_j}(\mathrm{X_k},\mathrm{Q(p)})).$$

The proof uses a Bahadur-Kiefer representation of the quantile process, together with the above mentioned convergence results for the two-sample U-process.

Acknowledgement. Financial support by the Collaborative Research Center SFB 823 *Statistical Modeling of Nonlinear Dynamic Processes* of the German Research Foundation DFG is gratefully acknowledged.

References

[1] H.Dehling and R.Fried (2012): Asymptotic Distribution of Two-Sample Empirical U-Quantiles with Applications to Robust Tests for Shifts in Location. *Journal of Multuvariate Analysis* **105**, 124–140.

[2] H.Dehling, R.Fried, I.Garcia and M.Wendler (2013): Change-point detection under dependence based on two-sample U-statistics. To appear in: *Asymptotic Methods in Statistics - Festschrift in Honor of Miklos Csörgő's 80th Birthday.*

[3] H.Dehling, R.Fried and M.Wendler (2013): A Robust Method for Shift Detection in Time Series. *Work in Progress.*

[4] H.Dehling, A.Rooch and M.S.Taqqu (2013a): Nonparametric Change-Point Tests for Long-Range Dependent Data. *Scandinavian Journal of Statistics* **40**, 153–173.

[5] H.Dehling, A.Rooch and M.S.Taqqu (2013b): Power of Change-Point Tests for Long-Range Dependent Data. *arXiv*:1303.4917v1 [math.ST].

[6] A.Rooch: *Change-point tests for long-range dependent data.* Dissertation, Ruhr-Universität Bochum 2012.

Weak dependence, models limit theory and an application to DNA modeling

Paul Doukhan

1. Dependence

1.1. **Independence.** The question is: *how to weaken stochastic independence ?* $\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B)$ relating the events $A\in\sigma(P)$ of the past history with those $B\in\sigma(F)$ in a (not so close) future. also written $\mathrm{Cov}(f(P),g(F))=0$, $\forall f,g,\ \|f\|_\infty,\|g\|_\infty\le 1$. If $X=(X_t)_{t\in\mathbb{Z}}$ this relation should be weakened with $P=(X_{i_1},\dots,X_{i_u}),F=(X_{j_1},\dots,X_{j_v})$, $i_1\le\cdots\le i_u, j_1\le\cdots\le j_v$ and large $r=j_1-i_u$. For more general index sets, past and future are simply understood as distant index subsets.

1.2. **Strong mixing (Rosenblatt, 1956).** i.e. $\lim_{r\to\infty}\alpha(r)=0$ with $\alpha(r)=\sup_{P,F}\alpha(\sigma(P),\sigma(F))$ and $\alpha(\sigma(P),\sigma(F))=\sup_{A,B}|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|$. Bradley (2007) provides a complete and elementary presentation, Rio (2000) proposes sharp technical results and Doukhan (1994) gives examples. It is usually hard to prove mixing and it some cases one even may prove non-mixing:
A non mixing AR(1)-model, Andrews (1984) and Rosenblatt (1984)

$$X_t=\frac{1}{2}\left(X_{t-1}+\xi_t\right),\qquad \left(\xi_t\sim b\left(\frac{1}{2}\right)\ \text{iid}\right),\quad X_{t-1}=\text{frac}(2X_t)$$

Exhibit sets I_r, with dyadic extremities and $A=(X_0\in[0,\frac{1}{2}])\equiv B=(X_r\in I_r)$, hence: $\alpha(r)\geq|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|=\mathbb{P}(A)-\mathbb{P}^2(A)=\frac{1}{4}$.
A bilinear type non-mixing model, Doukhan, Mayo, Truquet (2009)

$$X_t=\xi_t(1+aX_{t-1}),P(\xi_0=\pm1)=1/2\ \text{iid},a\in\big(\frac{3-\sqrt{5}}{2},\frac{1}{2}\big],\ X_t=\sum_{j\geq0}a^j\xi_t\cdots\xi_{t-j}$$

here, a direct approach following that of Andrews leads to the result. Those simple examples are by themselves a sufficient reason to exhibit other tools!

1.3. **Covariance.** Independence sometimes coincides with orthogonality as for the cases of Gaussian an associated vectors. Recall that a vector $V\in R^p$ is associated if $\text{cov}(f(V),g(V))\geq0$ for coordinatewise non-decreasing functions $f,g:R^p\to R$ such that this expression is well defined.

1.4. **General formulation.** $(X_t)_{t\in Z}$ $(\in E)$, $f:E^u\to R$ from a class of functions $\mathbf{F}$, $g:E^v\to R$ from $\mathbf{G}$:

$$|\text{Cov}\left(f(X_{i_1},\ldots,X_{i_u}),g(X_{j_1},\ldots,X_{j_v})\right)|\leq\Psi(f,g)\epsilon(r),\quad \epsilon(r)\downarrow0$$

and if $\Psi(f,g)=v\text{Lip}g$, then we set $\epsilon(r)=\theta(r)$;
if $\Psi(f,g)=u\text{Lip}f+v\text{Lip}g$, then $\epsilon(r)=\eta(r)$;
if $\Psi(f,g)=uv\text{Lip}f\cdot\text{Lip}g$ then $\epsilon(r)=\kappa(r)$;
if $\Psi(f,g)=u\text{Lip}f+v\text{Lip}g+\ uv\text{Lip}f\cdot\text{Lip}g$ then $\epsilon(r)=\lambda(r)$, with
$\text{Lip}f=\sup_{y\neq x}\dfrac{|f(y)-f(x)|}{|y_1-x_1|+\cdots+|y_u-x_u|}$.
The limit theory and many applications may be found in [3].

2. Models

2.1. **Chaotic models.** Vector valued models $X_t=\xi_t\big(a+\sum_{j=1}^{\infty}a_jX_{t-j}\big)$ if $\phi=\|\xi_0\|_m\sum_j\|a_j\|<1$ then a strictly stationary $\mathbb{L}^m$-solution writes

$$X_t=\xi_t\Big(a+\sum_{k=1}^{\infty}\sum_{j_1,\ldots,j_k\geq1}a_{j_1}\xi_{t-j_1}\cdots a_{j_k}\xi_{t-j_1-\cdots-j_k}a\Big)$$

- GARCH(p,q) (Engle, Granger) $r_t=\sigma_t\epsilon_t$, $\sigma_t^2=\sum_{j=1}^{p}\beta_j\sigma_{t-j}^2+\gamma_0+\sum_{j=1}^{q}\gamma_jr_{t-j}^2$.
- ARCH(∞) (Surgailis *et al.* 2001) $r_t=\sigma_t\epsilon_t$, $\sigma_t^2=\beta_0+\sum_{j=1}^{\infty}\beta_jr_{t-j}^2$.

- Bilinear (Giraitis, Surgailis, 2003) $X_t = \zeta_t\Big(a + \sum_{j=1}^{\infty} a_j X_{t-j}\Big) + b + \sum_{j=1}^{\infty} b_j X_{t-j}$.

2.2. **Memory models.** $X_t = F(X_{t-1}, X_{t-2}, X_{t-3}, \ldots; \xi_t)$ (Doukhan and Wintenberger, 2008) if $(\xi_t)_{t\in\mathbb{Z}}$ iid, $F : (\mathbb{R}^d)^{\mathbb{N}} \times \mathbb{R}^D \to \mathbb{R}^d$, $m \geq 1$, $\|F(0,0,0,\ldots;\xi_t)\|_m < \infty$ and $\|F(x_1,x_2,x_3,\ldots;\xi_t) - F(y_1,y_2,y_3,\ldots;\xi_t)\|_m \leq \sum_{j=1}^{\infty} a_j \|x_j - y_j\|$ with $a = \sum_{j=1}^{\infty} a_j < 1$, then there exists a weakly dependent strictly stationary solution $X_t = H(\xi_t, \xi_{t-1}, \ldots) \in \mathbb{L}^m$.

2.3. **Integer valued models.** Thining (or Steutel & van Harn) operator is defined as $a \circ X = sign(X) \sum_{i=1}^{|X|} Y_i$ for $a > 0$, $X \in \mathbb{Z}$, $(Y_i)_i$ is iid, context-independent, $\mathbb{E}Y_0 = a$ (e.g. Poisson or Bernoulli).

- Galton-Watson process with immigration, INAR $X_t = a \circ X_{t-1} + \xi_t$. More generally Random INAR models $X_t = a_t \circ X_{t-1} + \xi_t$, are defined through stationary (a_t) such $E(a_t | lF_{t-1}) < 1$ (working paper).
- Integral bilinear models $X_t = a \circ X_{t-1} + b \circ (\varepsilon_{t-1} X_{t-1}) + \varepsilon_t$. Estimation from moments (Doukhan, Latour and Oraichi, 2006).
- GLM integer models $X_t | F_{t-1} \sim P(\lambda_t)$ with $\lambda_t = g(\lambda_{t-1}, X_{t-1}, \ldots)$ with Fokianos & Tjostheim (2012) under Lipschitz, with Douc & Moulines for loglinear and threshold models (2013).

3. Promotors in DNA analysis

The end of the presentation is a fast discussion on joint working papers with Jean-Paul Feugeas, Xiaoyin Li, and Wei Biao Wu. A zone of DNA chains is located before protein zones and nucleotids are distributed on $\{A, C, G, T\}$ with distributions depending smoothly on there location: $\mathbb{P}(X_k = A) = p_A(k/n)$ for $k = 1, \ldots, n$ and analogously for other values of the nucleotid.
Let $(U_k)_k$ be a stationary sequence with uniformly distributed margins the a simple model is $X_k = \mathbb{1}_{\{U_k \leq p_A(k/n)\}}$. One may rewrite this as a regression model with $\mathbb{L}^2-$stationary innovations:

$$X_k = p_A(k/n) + \sqrt{p_A(k/n)(1 - p_A(k/n))}\xi_k$$

A first way to infer on this function p_A is to work out a kernel regressor

$$\widehat{p}_A(k/n) = \frac{1}{nh} \sum_{k=1}^{n} X_k K\left(\frac{t - k/n}{h}\right)$$

Under standard smoothness arguments a CLT provides us with confidence bounds

$$Z_n(t) = \sqrt{nh}(\widehat{p}_A(t) - p_A(t) - b_n(t)) \to N(m(t), \sigma^2(t))$$

Now a non-stationary invariance principle by Wu and Shao (2007) allows to derive uniform confidence bounds through asymptotics for $\|Z_n\|_\infty$ and allows qualitative tests of hypothesis. Anyway rates as in Wu and Zhou (2011) need to be improved in order to avoid bandwidth restrictions.

References

[1] Andrews, D. (1984) *Non strong mixing autoregressive processes.* J. Appl. Probab. 21, 930-934.

[2] Bradley, R. (2007) *Introduction to strong mixing conditions,* Kendricks Press.

[3] Dedecker, J., Doukhan, P., León, J. R., Lang, G., Louhichi, S., Prieur, C. (2007) *Weak dependence: models, theory and applications.* Lecture Notes in Statistics 190, Springer-Verlag.

[4] Doukhan, P. (1994) *Mixing: Properties and Examples.* Lecture Notes in Statistics 85. Springer Verlag.

[5] Doukhan, P. Mayo, N. Truquet, L. (2009) *Weak dependence, models and some applications.* Metrika 69:2-3, 199-225.

[6] Rio, E. (2000) *Théorie asymptotique pour des processus aléatoires faiblement dépendants,* SMAI, Mathématiques et Applications 31, Springer.

[7] Wu,W., Shao, Z. (2007) . Inference of trends in time series. *J. R. Statist. Soc. B 69*, 391–410.

[8] Wu, W., Zhou, Z. (2011) . Gaussian approximations for non-stationary multiple time series. *Statistica Sinica 21*, 1397- 1413.

Graphical time series models

Michael Eichler

Graphical models have become an important tool for the statistical analysis of complex multivariate data sets. The key feature of these models is to merge the probabilistic concept of conditional independence with graph theory by representing possible dependences among the variables of a multivariate distribution in a graph. This leads to simple graphical criteria for identifying conditional independences that hold, for instance, in a submodel. Further important advantages of the graphical modelling approach are statistical efficiency due to parsimonious parameterizations of the joint distribution of the variables and the visualization of complex dependence structures, which allows an intuitive understanding of the interrelations among the variables and, thus, facilitates the communication of statistical results.

While graphical models originally have been developed for variables that are sampled with independent replications, they have been applied more recently also to the analysis of time dependent data (e.g. [1, 2, 3, 4, 5]). However, due to the added serial dependence, time series exhibit even for a relatively small number of variables quite complex dependence structures. This leads to theoretical as well as practical problems with the traditional graphical modelling approaches, where the variables at different time points are represented by separate nodes.

In this talk, we review new types of graphical models that have been developed for modelling and analysing multivariate time series. These models are based on graphs in which each node represents a full stochastic process. This leads to much simpler graphs with as many nodes as there are variables. With time series as basic stochastic entities, there are more possibilities to choose the type of conditional independences encoded by the graph.

The first example for this approach are the partial correlation graphs introduced by Dahlhaus [2]. These are undirected graphs with an edge $i \longrightarrow j$ omitted

whenever the corresponding two processes are uncorrelated after removing the linear effects of all other time series. Thus, they are straightforward extension of Markov random fields to the time series case. Eichler [6] describes graphical vector autoregressive (VAR) models constrained by such undirected graphs. As the constraints imposed by the graph on the standard autoregressive parameters are non-linear, a new parameterization is proposed which yields simple zero constraints and thus is better suited for model fitting.

One disadvantage of partial correlation graphs is that they do not provide information about the dynamic dependences among the variables. As an alternative, Eichler [7] introduced graphical representations that utilize the concept of Granger causality [8]. In these Granger causality graphs, directed edges ($\longrightarrow$) indicate possible Granger–causal relationships between variables while undirected edges (---) are used to map the contemporaneous dependence structure. For Gaussian processes, this leads to graphical VAR models with zero–constraints on the autoregressive parameters. More general graphical time series models with non-linear dynamics have been discussed in Eichler [9].

In contrast to undirected graphical time series models, the class of Granger–causal time series models is not closed under marginalization. This means that it is possible that the Granger–causal relations of a subprocess in a graphical VAR model cannot be completely endcoded by a Granger causality graph. The problem is of importance for the identification of causal structures when part of the variables are unobserved. The problem can be solved by allowing an additional type of dashed directed edges (--→) representing Granger–causal relationships that are induced by latent variables.

For modelling, we consider multivariate stationary Gaussian processes $\boldsymbol{X}_V$ that are given by

$$\boldsymbol{X}_V(t) = \sum_{u=1}^{p} \boldsymbol{\Phi}(u)\, \boldsymbol{X}_V(t-u) + \boldsymbol{\varepsilon}_V(t),$$

where $\boldsymbol{\varepsilon}_V$ is a stationary Gaussian process with mean zero and covariances

$$\operatorname{cov}\big(\boldsymbol{\varepsilon}_V(t), \boldsymbol{\varepsilon}_V(t-u)\big) = \begin{cases} \boldsymbol{\Omega}(u) & \text{if } |u| \le q \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for some $q \in \mathbb{N}$. For a given mixed graph G, the parameters are constrained by

a) $\Phi_{ba}(u) = 0$ for all $u > 0$ if $a \longrightarrow b$ is not in G;
b) $\Omega_{ba}(u) = 0$ for all $u > 0$ if a --→ b is not in G;
c) $\Omega_{ba}(0) = \Omega_{ab}(0) = 0$ if a --- b is not in G.

Since the process $\boldsymbol{\varepsilon}_V$ is a multivariate moving average process of order q, the processes $\boldsymbol{X}_V(t)$ form a *graphical multivariate ARMA(p,q) model*. The following results hold (see [10]):

a) the model satisfies the global Granger–causal Markov property with respect to G;
b) if G satisfies an ancestrality condition (i --→ $j \notin G$ whenever there exist a directed path $i \longrightarrow \cdots \longrightarrow j$ in G) the model is identifiably.

As in the case of graphical VAR(p) models with respect to undirected graphs, the constraints on the parameters are best expressed in an alternative parameterization. In contrast to the other graphical time series models discussed above, the graph is not defined in terms of a pairwise Markov property. As a consequence model identification in this more general framework requires new model search strategies that will be topic of future research.

References

[1] E. Stanghellini and J. Whittaker, *Analysis of multivariate time series via a hidden graphical model*, in: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics (1999), 250–254.

[2] R. Dahlhaus, *Graphical interaction models for multivariate time series*, Metrika **51** (2000), 157–172.

[3] M. Reale and G. Tunnicliffe Wilson, *Identification of vector AR models with recursive structural errors using conditional independence graphs*, Statistical Methods and Applications **10** (2001), 49–65.

[4] R. Dahlhaus and M. Eichler, *Causality and graphical models in time series analysis*, in: Highly Structured Stochastic Systems (eds P. Green, N. Hjort and S. Richardson). Oxford: University Press (2003), pp. 115–137.

[5] A. Moneta and P. Spirtes, *Graphical models for the identification of causal structures in multivariate time series models*, in: Joint Conference on Information Sciences Proceedings, Atlantis Press (2006).

[6] M. Eichler, *Fitting graphical interaction models to multivariate time series*, in: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (2006), 147–154.

[7] M. Eichler, *Granger causality and path diagrams for multivariate time series*, Journal of Econometrics **137** (2007), 334–353.

[8] C.W.J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica **37** (1969), 424–438.

[9] M. Eichler, *Graphical modelling of multivariate time series*, Probability Theory and Related Fields **153** (2012), 233–268.

[10] M. Eichler, *Causal inference in time series analysis*, in: Causality: Statistical Perspectives and Applications (eds C. Berzuini, A.P. Dawid, L. Bernardinelli) Wiley, Chichester (2012).

Incidental Endorgeneity in high-dimensional statisticse

Jianqing Fan

(joint work with Yuan Liao)

Consider the sparse linear model

$$Y = X^T\beta + \varepsilon$$

where β is a high-dimensional sparse vector. Most papers on high-dimensional statistics are based on the assumption that none of the regressors are correlated with the regression error, namely, they are exogenous. In other words, the fundamental assumption

$$EX_j\varepsilon = 0, \qquad j = 1, \cdots, p \tag{1}$$

has been made. Yet, incidental endogeneity arises easily in a large pool of regressors in a high-dimensional regression, namely some of the equations in (1) do

not hold when p is large. This is demonstrated by various examples in scientific studies and validated via statistical tests. Incidental endorgeneity causes the inconsistency of the penalized least-squares method [1, 3], and possible false scientific discoveries. A necessary condition for model selection consistency of a very general class of penalized regression methods is given, which allows us to prove formally the inconsistency claim [2]. To cope with the possible incidental endogeneity, we [2] construct a novel penalized focused generalized method of moments (FGMM) criterion function and offer a new optimization algorithm. The FGMM is an extra filter that excludes all incidental endogenous predictors and rely on the over identification conditions. To establish its asymptotic properties, we first study the variable selection consistency for a general class of penalized regression methods. These results are then used to show that the FGMM possesses the oracle property even in the presence of incidental endogenous predictors, and that the solution is also near global minimum under the over-identification assumption. Finally, we also show how the semi-parametric efficiency of estimation can be achieved via a two-step approach.

References

[1] J. Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348-1360.

[2] item Fan, J. and Liao, Y. (2012). Endogeneity in ultrahigh dimension. *http://arxiv.org/abs/1204.5536*

[3] R. Tibshirani (1996). Regression shrinkage and selection via lasso. *Jour. Roy. Statist. Soc. B.*, **58**, 267-288. M. Muster, *Computing certain invariants of topological spaces of dimension three*, Topology **32** (1990), 100–120.

Stable estimates for high-dimensional hidden Markov models

Jürgen Franke

(joint work with Mark Fiecas, Rainer von Sachs, Joseph Tadjuidje-Kamgaing)

Hidden Markov models (HMM) or, more generally, Markov switching models have found considerable interest as models for time series with sudden changes between various regimes - compare [2] for a review. In particular, [4] showed that even the simple HMMs are able to generate many of the so-called stylized facts of financial time series. Here, we consider a sample $X_1, \ldots, X_N$ from a d-dimensional time series which is controlled by a hidden Markov chain Q_t assuming only finitely many states. For this exposition, we assume that there are only two states 0 and 1, but our approach can be generalized to $K \geq 2$ different states easily. Let $\varepsilon_t \in \mathbb{R}^d$ be i.i.d. with mean vector 0 and unit covariance matrix I_d. Then,

$$X_t = \mu_k + \Sigma_k^{1/2} \varepsilon_t \quad \text{iff} \quad Q_t = k, \quad k = 0, 1, \;\; t = 1, \ldots, N.$$

Together with the transition probabilities $q_{k\ell} = \mathbb{P}(Q_t = \ell | Q_{t-1} = k), k = 0, 1$, the free model parameter is given by $\vartheta = (q_{00}, q_{11}, \mu_0, \mu_1, \Sigma_0, \Sigma_1)$. Given the data, we would like to estimate ϑ as well as to reconstruct the hidden variables $Q_1, \ldots, Q_N$. For low dimension d, there are various approaches giving satisfactory solutions

of this estimation and filtering problem. We consider here the EM algorithm, providing numerical approximations to the maximum Gaussian pseudo likelihood estimate of ϑ and the Viterbi algorithm for reconstructing the most likely path of the hidden process given the data. For large dimension d compared to sample size N, however, this procedure breaks down in various respects. E.g., the lower plot in Figure 1 shows the true path Q_t (lower half) and the reconstructed path (upper half), shifted by 1 for better visibility. The latter is quite erratic. The sample of size $N = 256$ has been simulated from a $d = 20$-dimensional HMM with $q_{00} = q_{11} = 0.95, \mu_0 = \mu_1 = 0$, a block diagonal Σ_0 with four 5×5 blocks and a tridiagonal Σ_1 with Gaussian innovations ε_t.

We propose a remedy for those problems which can be traced back to the instability of sample covariance matrices, and even more of their inverses, in higher dimensions. The resulting stabilized algorithm is able to reproduce the hidden Markov variables quite well - compare Figure 1 (upper plot). Let us first assume that we know the Q_t from an oracle. Then, we get immediately estimates for the unknown parameters from their sample versions:

$$
\begin{aligned}
&N_1 = \sum_{t=1}^{N} Q_t, \quad N_0 = N - N_1, \quad q^{\mathbf{o}}_{kk} = \frac{1}{N_k - 1} \sum_{t=2}^{N} 1_{\{Q_t = k, Q_{t-1} = k\}}, k = 0, 1, \\
&\mu^{\mathbf{o}}_1 = \frac{1}{N_1} \sum_{t=1}^{N} Q_t X_t, \quad \mu^{\mathbf{o}}_0 = \frac{1}{N_0} \sum_{t=1}^{N} (1 - Q_t) X_t, \\
&\Sigma^{\mathbf{o}}_1 = \frac{1}{N} \sum_{t=1}^{N} Q_t (X_t - \mu^{\mathbf{o}}_1)(X_t - \mu^{\mathbf{o}}_1)', \\
&\Sigma^{\mathbf{o}}_0 = \frac{1}{N} \sum_{t=1}^{N} (1 - Q_t)(X_t - \mu^{\mathbf{o}}_0)(X_t - \mu^{\mathbf{o}}_0)', \qquad (1)
\end{aligned}
$$

where we standardize the covariance estimates by the sample size N instead of N_0, N_1 which enhances the numerical stability in cases where a state is rarely visited, i.e. where N_0 resp. N_1 is small. Hence, $\Sigma^{\mathbf{o}}_k$ does not estimate Σ_k, but $\pi_k \Sigma_k$ where $\pi_k = \mathbb{P}(Q_t = k)$ in the stationary state. In high dimension, $\Sigma^{\mathbf{o}}_k, k = 0, 1$, frequently have large condition numbers which is the main reason for the observed problems. We, therefore, shrink the sample covariances towards a highly stable matrix. In particular, we choose a multiple of the unit matrix I_d and set for some weights $0 \le W_k \le 1$:

$$
\Sigma^{\mathbf{s}}_k = (1 - W_k)\Sigma^{\mathbf{o}}_k + W_k \alpha_k I_d \ \text{ with } \operatorname{tr}(\alpha_k I_d) = \alpha_k d = \mathrm{E}\, \operatorname{tr}(\Sigma^{\mathbf{o}}_k), \ k = 0, 1.
$$

Afterwards, the unknown α_k is replaced by its estimate $\alpha^{\mathbf{o}}_k = \frac{1}{d}\operatorname{tr}(\Sigma^{\mathbf{o}}_k)$. Based on the work of [3] and [5], we can derive explicit formulas for the optimal weights $W^{\mathbf{o}}_k$, minimizing the mean squared error $\mathrm{E}\ \|\Sigma^{\mathbf{s}}_k - \pi_k \Sigma_k\|^2$, as well as feasible estimates $\widehat{W^{\mathbf{o}}_k}$ for them. We can prove under appropriate conditions that the difference between the optimally shrinked covariance estimates with weights $W^{\mathbf{o}}_k$ and the

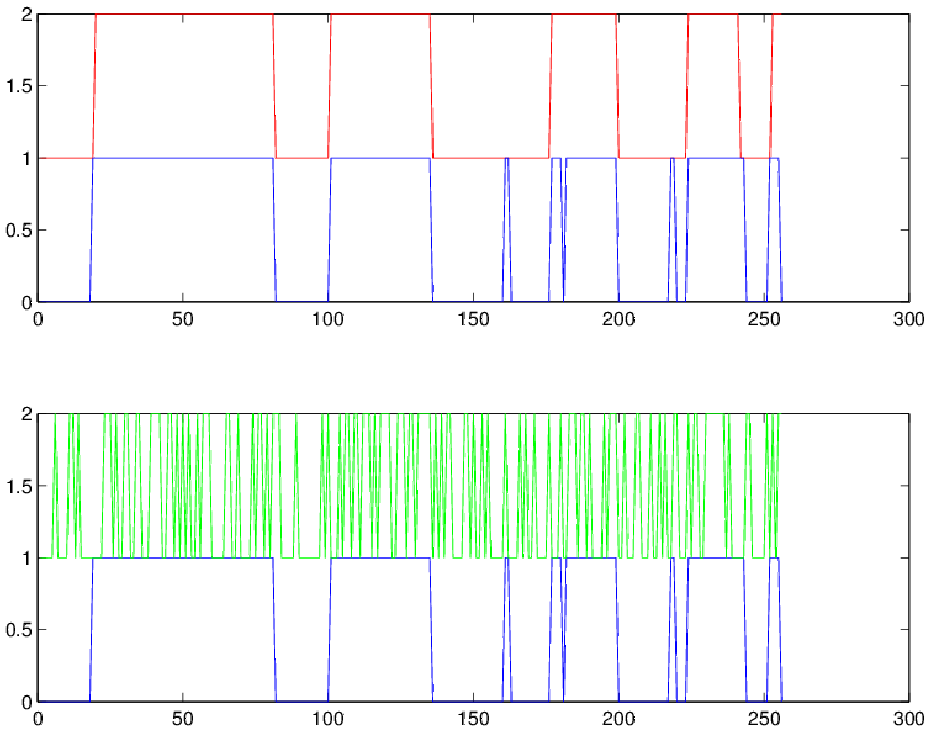


FIGURE 1. State sequence Q_t and reconstruction using standard EM (lower half) and EM with shrinkage (upper half).

realizable shrinkage estimates with weights $\widehat{W}_k^{\mathbf{o}}$ is asymptotically negligible for $N \to \infty$ even for increasing dimension $d \to \infty$, $\frac{d}{N} \to 0$.

Without an oracle, we approximate the unknown Q_t by their conditional expectations $\widehat{Q}_t^{(i)}$ given the data $X_1, \ldots, X_N$. This is done in the E step of the EM algorithm pretending that the parameter ϑ coincides with an estimate $\widehat{\vartheta}^{(i-1)}$ from the previous iteration. In the M step, we get a new parameter estimate $\widehat{\vartheta}^{(i)}$ by replacing the unobservable Q_t in (1) by their approximations $\widehat{Q}_t^{(i)}$. E.g., we have

$$\widehat{\Sigma}_1^{(i)} = \frac{1}{N} \sum_{t=1}^{N} \widehat{Q}_t^{(i)} (X_t - \widehat{\mu}_1^{(i)})(X_t - \widehat{\mu}_1^{(i)})'.$$

To get covariance matrix estimates with small condition numbers, we then apply shrinkage to the sample covariance matrices $\widehat{\Sigma}_1^{(i)}, \widehat{\Sigma}_0^{(i)}$ as the final part of the M step before proceeding to the $(i+1)^{th}$ iteration. This modification of the EM algorithm works well for simulated and real data, compare [1] for the details.

Finally, let us remark that the shrinkage procedure used for stabilizing the EM algorithm is related to another approach for improving the condition number of covariance matrix estimates which adds a penalty term to the pseudo log likelihood. To keep the notation simple, we depart for the moment from the HMM setting and consider only an i.i.d. sample $X_1, \ldots, X_N$ of random vectors with mean μ and covariance matrix Σ. We maximize the Gaussian pseudo log likelihood after adding a penalty term $\lambda pen(\Sigma)$ with tuning parameter λ. Following [6], we use

$pen(\Sigma) = \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}\Omega)$ as a penalty term for some fixed matrix Ω. Let $\widehat{\Sigma}$ denote the sample covariance matrix. [6] have shown that, for the special choice $\Omega = \alpha I_d, \alpha = \frac{1}{d}\mathrm{E}\ \mathrm{tr}(\widehat{\Sigma})$, the penalized pseudo maximum likelihood estimate of Σ coincides with the shrinkage estimate of [3] where the shrinkage weight W and the penalty weight λ are related in a simple manner. The approach of [3], however, goes beyond plain penalized maximum likelihood as it includes a data adaptive choice of the weight W as well as an estimate of α. These considerations can be extended straightforwardly to HMM, where in the M step the shrinkage estimates $\Sigma_k^{\mathbf{o}}, k = 0, 1$, can be interpreted as the results of adding corresponding penalty terms to the approximations of the complete Gaussian pseudo log likelihood before maximization.

References

[1] M. Fiecas, J. Franke, R. von Sachs, J. Tadjuidje-Kamgaing, *Shrinkage estimation for multivariate hidden Markov models*, Discussion papers, DP2012/16, Universite Catholique de Louvain, submitted for publication (2012).

[2] J. Franke, *Markov switching time series models*, In: Handbook of Statistics **30**, Time Series Analysis and Applications, C.R. Rao, S. Subba Rao and T. Subba Rao eds., Elsevier (2012), 99–122.

[3] O. Ledoit, M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivariate Analysis **88** (2004), 365–411.

[4] T. Rydén, T. Terasvirta, S. Asbrink, *Stylized facts of daily return series and the hidden Markov model of absolute returns*, J. Appl. Econometrics **13** (1998), 217–244.

[5] A. Sancetta, *Sample covariance shrinkage for high-dimensional dependent data*, J. Multivariate Analysis **99** (2008), 949–967.

[6] M. Yuan, J. Huang, *Regularized parameter estimation of high dimensional t distribution*, J. Statist. Planning and Inference **139** (2009), 2284–2292.

On multi-zoom autoregressive time series models

Piotr Fryzlewicz

In classic linear time series autoregression (AR), the univariate time series X_t under consideration is modelled as a linear but otherwise unconstrained function of its own past values $X_{t-1}, X_{t-2}, \ldots$, plus white-noise-like innovation ε_t. That is,

$$X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p} + \varepsilon_t. \tag{1}$$

In some situations, it appears to be a good idea to model X_t as depending explicitly on some other features of its own past, rather than on the individual variables $X_{t-1}, \ldots, X_{t-p}$.

As an example, consider the problem of modelling mid- and high-frequency financial returns, where X_t represents a fine-scale, e.g. one-minute, return on a financial instrument. In the hope of improving the predictive power, the analyst may wish to model X_t as depending not only on the past few one-minute returns, but also perhaps on past returns on lower frequencies, such as one hour or one day. Representing this in an unconstrained way as in (1) with a large value of p would lead to obvious over-parameterisation.

Our proposed way to resolve this issue is to adopt what we call a "multi-zoom" approach to time series analysis. The main idea of the approach is to include as regressors for X_t features of the path $X_1, \ldots, X_{t-1}$ which "live" on multiple time-scales, and hence correspond to considering the time series at different zoom levels.

For example, in the financial time series context described above, we could entertain a multi-zoom AR model of the form

$$X_t = \alpha_1 \frac{1}{\tau_1}(X_{t-1} + \ldots + X_{t-\tau_1}) + \ldots + \alpha_p \frac{1}{\tau_p}(X_{t-1} + \ldots + X_{t-\tau_p}) + \varepsilon_t, \tag{2}$$

where the time scales τ_k are such that $1 = \tau_1 < \tau_2 < \ldots < \tau_p$. Note that $X_{t-1} + \ldots + X_{t-\tau_k}$ represents the most recent τ_k-minute return. There is nothing to stop τ_k, $k > 1$, from being large, e.g. of the order of tens or hundreds. The number of scales p would typically be much smaller than the longest time scale τ_p (note that the standard AR model (1) can always be rewritten in the form (2) if we take $\tau_p = p$). Including the regressors $X_{t-1} + \ldots + X_{t-\tau_k}$, rather than the individual variables X_{t-s}, corresponds to "zooming out" of the original time scale on which the data were collected, and explicitly incorporating information from coarser time scales. In this instance, the returns $X_{t-1} + \ldots + X_{t-\tau_k}$ represent the multi-zoom "features" that we believe have some predictive power with respect to X_t.

The following questions are of immediate methodological interest:

- *Model identification and stationarity.* We note that the multi-zoom AR model in equation (2) is a particular, sparsely parameterised, instance of the AR(τ_p) model. Therefore, stationarity (or otherwise) of multi-zoom AR can be established via the usual route for AR processes.
- *Estimation of p, τ_k and α_k.* In the simplest case, the values of p and $\{\tau_k\}_{k=1}^p$ are chosen by the analyst, and only the coefficients $\{\alpha_k\}_{k=1}^p$ need to be estimated. This can be done e.g. via OLS, or by performing an unconstrained estimation for AR(τ_p) and then grouping the estimated coefficients into sections of piecewise constancy. If $\{\tau_k\}_{k=1}^p$ are unknown, the grouping can be achieved via change-point detection techniques. If p is also unknown, change-point detection needs to be coupled with devices for model choice based e.g. on thresholding or on the use of information criteria.
- *Use of other multi-zoom features.* It is of interest to generalise model (2) to other multi-zoom features, for example the wavelet coefficients of the original price process at different scales, or nonlinear breakout-type statistics (the latter being of interest in e.g. algorithmic trading). The introduction of non-linearity introduces particularly challenging methodological questions of model identifiability and estimation. Note that the linear dependence on non-linear features that this induces goes in the opposite direction to the non-linear dependence on linear features seen, for example, in Generalised Linear Models.

- *Applicability in financial statistics.* Preliminary results suggest that multi-zoom AR processes are good at explaining the apparent lack of serial dependence in time series of financial returns, when measured via the sample autocorrelation, which can be blind to multi-scale dependencies such as those in (2) due to its single-scale nature. Moreover, empirically, multi-zoom AR processes appear to have relatively good predictive power for forecasting high- and mid-frequency financial returns.

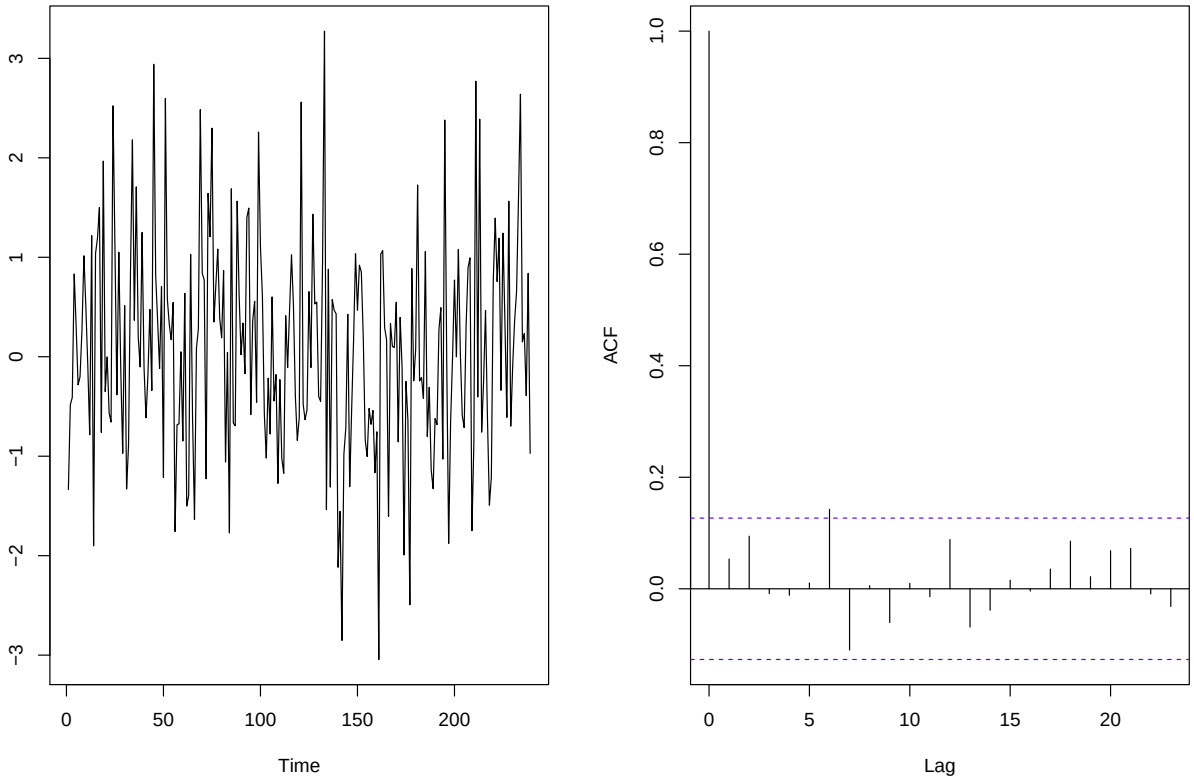


FIGURE 1. Left: sample path simulated from model (2) with length $n = 250$, $p = 2$, $\tau_1 = 1$, $\tau_2 = 10$, $\alpha_1 = 0.1$, $\alpha_2 = 0.5$, ε_t iid standard normal. Right: the sample autocorrelation of the simulated sample path.

The fact that multi-zoom AR processes can "mask" as white noise from the point of view of the sample autocorrelation (and hence be potentially be attractive from the point of view of modelling financial returns, which tend to exhibit this empirical feature) is illustrated in Figure 1. Despite the model being far from white noise, the sample autocorrelation fails to detect the serial dependence in the process, which is in part due to the fact that this measure takes no account of the multi-zoom structure of the model.

We are grateful to the workshop participants for pointing us to some other related literature, and in particular to the models described in [3], [1], [2]. We emphasize again that in contrast to these, our approach enables, in particular, automatic selection of the relevant time-scales τ_k. This also sets it apart from the autoregressive index models in [4].

References

[1] F. Corsi, *A simple approximate long-memory model of realized volatility*, Journal of Financial Econometrics **7** (2009), 174–196.

[2] S. Heston, R. A. Korajczyk and R. Sadka, *Intraday patterns in the cross-section of stock returns*, Journal of Finance **65** (2010), 1369–1407.

[3] U. Müller, M. Dacorogna, R. Dav, R. Olsen, O. Pictet, and J. von Weizsacker, *Volatilities of different time resolutions – Analysing the dynamics of market components*, Journal of Empirical Finance **4** (1997), 213–239.

[4] G. Reinsel, *Some results on multivariate autoregressive index models*, Biometrika **70** (1983), 145–156.

Functional Ridge Regularization

Yulia R. Gel

Our main interest is in real-time estimation and prediction of the same realization of a linear time series that does not degenerate to a finite dimensional autoregressive (AR) form. To reduce bias, the order of AR approximation p_t is to increase with the sample size t (Gerencser, 1992; Ing and Wei, 2003). However, the classical model selection methods such as the Akaike Information Criterion (AIC) are too conservative and typically choose too low p_t to achieve strong asymptotic efficiency (Ing and Wei, 2005; Ing, 2007). Thus, the desired estimation algorithm should preferably account for as much previous history as possible, yield a non-degenerative sample covariance matrix and be relatively computationally inexpensive. The possible approaches in this direction is to use either banding or thresholding algorithms for the sample autocovariance matrix (Wu and Pourahmadi, 2009; McMurry and Politis, 2010; Bickel and Gel, 2011). However, neither banding nor thresholding allow for direct recursive implementation. As an alternative, we can use functional ridge regularization of the sample covariance matrix in such a way that the inverse of a functional ridge is a nuclear operator in $\ell_2(\mathbb{N})$. Such estimation procedure allow for an immediate implementation via the stationary recursive Kalman filter, which computational simplicity makes it particularly attractive for real-time, or online estimation and forecasting. Here, we aim to extend the results of Barabanov and Gel (2005) and Gel and Barabanov (2007) on exponential and polynomial regularizers derived for a case of a short memory linear process (i.e. with exponentially decaying autocorrelation functions) to a more general case of a linear process with absolutely summable coefficients.

Suppose that $y_1, y_2, \ldots, y_t$ come from

$$y_t + \sum_{i=1}^{\infty} a_i y_{t-i} = v_t, \quad t = \ldots, -1, 0, 1, \ldots \tag{1}$$

where

$$a(\lambda) = 1 + a_1\lambda + \ldots \neq 0, \ \ \forall|\lambda| \leq 1, \tag{2}$$

and

$$\sum_{i=1}^{\infty} |i^{\alpha} a_i| < \infty, \quad \alpha \geq 0. \tag{3}$$

Here v_t is the martingale difference ($\mathbf{E}(v_t|F_{t-1}) \equiv 0$) and $\mathbf{E}(v_t^2|F_{t-1}) = 1$ a.s., F_{t-1} is the σ-algebra generated by r. v. $(v_1, v_2, \ldots, v_{t-1})$, and $\sup_t \mathbf{E} v_t^4 < \infty$.

Hence, $\{y_t\}$ can be also represented as MA(∞)

$$y_t = \sum_{j=0}^{\infty} b_j v_{t-j}, \quad b_0 = 1, \tag{4}$$

where

$$b(\lambda) = \frac{1}{a(\lambda)} = 1 + b_1\lambda + \ldots \neq 0, \ \forall |\lambda| \leq 1, \quad \text{and} \quad \sum_{\mathrm{i}=1}^{\infty} |\mathrm{b_i}| < \infty.$$

Hence, the spectral density of $f(\lambda)$ of $\{y_t\}$ satisfies:

$$0 < F_1 < f(\lambda) < F_2, \quad F_1, F_2 > 0 \tag{5}$$

The condition (3) also implies that $\sum_{j=1}^{\infty} |j^{\alpha} b_j| < \infty$, $\alpha \geq 0$.

To estimate unknown AR(∞), we re-write it in a state-space form:

$$y_t = \Phi_{t-1}'\tau + v_t, \tag{6}$$

where $\Phi_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_1, 0, \ldots)$ and $\tau = -(a_1, a_2, \ldots)$ are in $\ell_2(\mathbb{N})$. Here we assume w.l.g. that $y_t = 0$, $t < 0$.

We can form a Yule-Walker (YW) system of infinite order

$$\Sigma\tau = r, \quad r = (\rho_1, \rho_2, \ldots), \tag{7}$$

where $\Sigma : \ell_2(\mathbb{N}) \to \ell_2(\mathbb{N})$ and $\Sigma > 0$. Hence, $\tau = \Sigma^{-1} r$ is a unique solution.

Let P_p be orthogonal projector in $\ell_2(\mathbb{N})$. We can then consider a truncated Yule-Walker system

$$P_p \Sigma P_p \tau = P_p r, \tag{8}$$

which leads to

$$\hat{\Sigma}_{p,t} \hat{\tau}_{p,t} = \hat{r}_{p,t}, \tag{9}$$

where $\hat{\Sigma}_{p,t}$ and $\hat{r}_{p,t}$ are the sample estimates of Σ_p and $r_{p,t}$, respectively.

Two main interrelated problems are selection of p and consistent estimation of Σ, and the possible approaches are:

(1) select p from AIC, i.e.

$$\mathrm{AIC(p)} = \ln \hat{\sigma_{\mathrm{p}}}^2 + \frac{2\mathrm{p}}{\mathrm{n}},$$

where $\hat{\sigma}_p^2$ is the sample variance of 1-step ahead forecast, and then estimate the truncated model. However, the order of truncation is typically very conservative, e.g. for $\{y_t\}_{t=1}^T$ with short memory, typically $p^{\mathrm{AIC}} = O(\log(T))$.

(2) if $p > p_{AIC}$, $\hat{\Sigma}_{p,t}$ might be a deficient estimator. Hence, we can use a **thresholding** operator $T_s(\Sigma) = (\rho_{i-j}\mathbf{1}_{|\rho_{i-j}|\geq s})$,

$$T_s(\hat{\Sigma}_t)\hat{\tau}_t^{th} = \hat{r}_t,$$

(3) **banding** operator $B_k(\Sigma) = (\rho_{i-j}\mathbf{1}_{|i-j|\leq k})$,

$$B_k(\hat{\Sigma}_t)\hat{\tau}_t^{b} = \hat{r}_t,$$

(4) **functional ridge** regularization.

Define a Functional Ridge (FR) estimator as

$$\hat{\Sigma}_t^f = \frac{1}{t}\left[\sum_{k=1}^{t}\Phi_k\Phi_k' + \varepsilon R\right], \quad \varepsilon > 0, \tag{10}$$

where R is a nuclear operator:

$$R = \mathrm{diag}\{\mathrm{f_k}\}_{\mathrm{k}=1}^{\infty}, \quad \mathrm{f_k} \to \infty, \mathrm{k} \to \infty.$$

For example, $f_k = e^{\mu k}$ for $\mu > 0$ or $f_k = k^p$ for $p > 0$.

The immediate benefit is the direct implementation via the recursive Kalman filter:

$$\begin{aligned} \hat{\tau}_{t+1}^f &= \hat{\tau}_t^f + \gamma_t^{\varepsilon}\Phi_t(y_{t+1} - \Phi_t'\hat{\tau}_t^f) \\ \gamma_{t+1}^f &= \gamma_t^f - \gamma_t^f\Phi_{t-1}(1 + \Phi_t'\gamma_t^f\Phi_t)^{-1}\Phi_t'\gamma_t^f, \end{aligned} \tag{11}$$

where $\gamma_0^f = (\varepsilon R)^{-1}$ and $\hat{\tau}_0^f = 0$.

Statement. Let $\sum_{i=1}^{\infty}|ia_i| < \infty$. Then, for $\delta > 0$ with probability 1

$$\lim_{T\to\infty} T^{1-\delta}|\hat{\tau}_T^f - \tau|^2 = 0.$$

References

[1] A. Barabanov and Y. Gel , *Strong consistency of the Least-Squares method with a polynomial regularizer for infinite AR models*, Automation and Remote Control **1**(66) 2005, 92–108.

[2] P. Bickel P. and Y. Gel, *Banded regularization of covariance matrices in application to parameter estimation and forecasting of time series*, Journal of the Royal Statistical Society, Series B **73**(5) 2011, 711–728.

[3] Y. Gel and A. Barabanov, *Strong Consistency of the Regularized Least-Squares Estimates of Infinite Autoregressive Models*, Journal of Statistical Planning and Inference, **137** 2007, 1260–1277.

[4] L. Gerencser, *AR(∞) estimation and nonparametric stochastic complexity*, IEEE Transactions on Information Theory **38** 1992, 1768–1778.

[5] C.-K. Ing and C.-Z. Wei, *On same-realization prediction in an infinite-order autoregressive process*, Journal of Multivariate Analysis **85** 2003, 130–155.

[6] C.-K. Ing and C.-Z. Wei, *Order selection for same-realization predictions in autoregressive processes* Annals of Statistics **33**(5) 2005, 2423–2474.

[7] T.L. McMurry and D.N. Politis, *Banded and tapered estimates for autocovariance matrices and the linear process bootstrap*, Journal of Time Series Analysis **31**(6) 2010, 471–482.

[8] C.-K. Ing, *Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series*, Annals of Statistics **35**(3) 2007, 1238–1277.

[9] W.B. Wu and M. Pourahmadi, *Banding sample covariance matrices of stationary processes*, Statistica Sinica **19** 2009, 1755–1768.

Composite Quantile Regression for the Single-Index Model

Wolfgang Karl Härdle

(joint work with Yan Fan, Weining Wang, and Lixing Zhu)

Regression between response Y and covariates X is a standard element of statistical data analysis. When the regression function is supposed to be estimated in a nonparametric context, the dimensionality of X plays a crucial role. Among the many dimension reduction techniques the single index approach has a unique feature: the index that yields interpretability and low dimension simultaneously. In the case of ultra high dimensional regressors X though it suffers, as any regression method, from singularity issues. Efficient variable selection is here the strategy to employ. Specifically we consider a composite regression with general weighted loss and possibly ultra high dimensional variables. Our setup is general, and includes quantile, expectile (and therefore mean) regression. We offer theoretical properties and demonstrate our method with applications to firm risk analysis in a CoVaR context.

Quantile regression(QR) is one of the major statistical tools and is "gradually developing into a comprehensive strategy for completing the regression prediction" [13]. In many fields of applications like quantitative finance, econometrics, marketing and also in medical and biological sciences, QR is a fundamental element for data analysis, modeling and inference. An application in finance is the analysis of conditional Value-at-Risk (VaR). [5] proposed the CaViaR framework to model VaR dynamically. [12] used their QR techniques to test heteroscedasticity in the field of labor market discrimination. Like expectile analysis it models the conditional tail behavior.

The QR estimation implicitly assumes an asymmetric ALD (asymmetric Laplace distribution) likelihood, and may not be efficient in the QMLE case. Therefore, different types of flexible loss functions are considered in the literature to improve the estimation efficiency, such as, composite quantile regression, [29], [9] and [10]. Moreover, [3] proposed a general loss function framework for linear models, with a weighted sum of different kinds of loss functions, and the weights are selected to be data driven. Another special type of loss considered in [17] corresponds to expectile regression (ER) that is in spirit similar to QR but contains mean regression as its special case. Nonparametric expectile smoothing work with application to demography could be found in [19]. The ER curves are alternatives to the QR curves and give us an alternative picture of regression of Y on X.

The difficulty of characterizing an entire distribution partly arises from the high dimensionality of covariates, which asks for striking a balance between model flexibility and statistical precision. To crack this tough nut, dimension reduction techniques of semiparametric type such as the single index model came into the focus of statistical modeling. [23] considered quantile regression via a single index

model. However, to our knowledge there are no further literatures on generalized QR for the single-index model.

In addition to the dimension reduction, there is however the problem of choosing the right variables for projection. This motivates our second goal of this research: variable selection. [14], [22] and [27] focused on variable selection in mean regression for the single index model. Considering the uncertainty on the multi-index model structure, we restrict ourselves to the single-index model at the moment. An application of our research is presented in the relevant financial risk area: to investigate how the revenue distribution of companies depends on financial ratios describing risk factors for possible failure. Such kind of research has important consequences for rating and credit scoring.

When the dimension of X is high, severe nonlinear dependencies between X and the expectile (quantile) curves are expected. This triggers the nonparametric approach, but in its full gear, it runs into the "curse of dimensionality" trap, meaning that the convergence rate of the smoothing techniques is so slow that it is actually impractical to use in such situations. A balanced dimension reduction space for quantile regression is therefore needed. The MAVE technique, [24] provides us 1) with a dimension reduction and 2) good numerical properties for semiparametric function estimation. The set of ideas presented there, however, have never been applied to composite quantile framework or an even more general composite quasi-likelihood framework. The semiparametric multi-index approach that we consider herein will provide practitioners with a tool that combines flexibility in modeling with applicability for even very high dimensional data. Consequently the curse of dimensionality is circumvented. The Lasso idea in combination with the minimum average contrast estimate (MACE) technique will provide a set of relevant practical techniques for a wide range of disciplines. The algorithms used in this project are published on the quantlet database www.quantlet.org.

References

[1] Adrian, T. and Brunnermeier, M. K. (2011). CoVaR. *Staff Reports 348, Federal Reserve Bank of New York.*

[2] Berkowitz, J., Christoffersen, P. and Pelletier, D. (2009). Evaluating value-at-risk models with desk-level data. *Working Paper 010, North Carolina State University, Department of Economics.*

[3] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B.* **73** (3) 325–349.

[4] Chao, S. K., Härdle, W. K. and Wang, W. (2012). *Quantile regression in Risk Calibration. In Handbook for Financial Econometrics and Statistics (Cheng-Few Lee, ed.).* Springer Verlag, forthcoming, SFB 649 DP 2012-006.

[5] Engle, R. F. and Manganelli, S. (2004). CaViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econ. Stat.*. **22** 367–381.

[6] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*. **96** 1348–1360.

[7] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*. **84** 986–995.

[8] Huber, P. J. (1985). Projection pursuit. *Ann. Math. Statist.*. **13** 435–475.

[9] Kai, B., Li, R. and Zou, H. (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *J. R. Statist. Soc. B.* **72** 49–69.
[10] Kai, B., Li, R. and Zou, H. (2011) New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *Ann. Statist.*. **39** (1) 305–332.
[11] Koenker, R. and Bassett, G. W. (1978). Regression quantiles. *Econometrica*. **46** 33–50.
[12] Koenker, R. and Bassett, G. W. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*. **50** 43–61.
[13] Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Econometric Perspectives*. **15** (4) 143–156.
[14] Kong, E. and Xia, Y. (1994). Variable selection for the single-index model. *Biometrika*. **94** 217–229.
[15] Leng, C., Xia, Y. and Xu, J. (2008). An adaptive estimation method for semiparametric models and dimension reduction. *WSPC-Proceedings*.
[16] Li, Y. and Zhu, J. (2008). L1- norm quantile regression. *J. Comput. Graph. Stat.*. **17** 163–185.
[17] Newey, W. and Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*. **55** 819–847.
[18] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*. **90** 1257–1270.
[19] Schnabel, S. and Eilers, P. (2009). Optimal expectile smoothing. *Comput. Stat. Data. An.*. **53** (12) 4168–4177.
[20] Serfling, R. J. (2001). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
[21] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.* **58** (1) 267–288.
[22] Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection methods for high-dimensional data: sparse mave. *Comput. Stat. Data. An.*. **52** 4512–4520.
[23] Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.*. **101** 1607–1621.
[24] Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Statist. Soc. B.* **64** 363–410.
[25] Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.*. **93** 228–237.
[26] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B.* **68** (1) 49–67.
[27] Zeng, P., He, T. H. and Zhu, Y. (2012). A lasso-type approach for estimation and variable selection in single-index models. *J. Comput. Graph. Stat.*. **21** 92–109.
[28] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*. **101** 476.
[29] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.*. **36** (3) 1108–1126.

Efficient estimation of integrated volatility in presence of jumps with activity bigger than 1

Jean Jacod
(joint work with Viktor Todorov)

We consider the problem of estimating the continuous part of the quadratic variation (henceforth referred to as integrated volatility) of a discretely-observed one-dimensional Itô semimartingale X over a finite time interval $[0,T]$: the observations are regularly spaced, with mesh Δ_n going to 0.

When the jump activity index (or Blumenthal-Getoor index) of X is smaller than 1, there are so far two kinds of estimators for the integrated volatility C_T, which both converge with the rate $1/\sqrt{\Delta_n}$: the truncated realized volatility (see [5]), and the multipower variations (see [1], [2], and [6] when the degree of activity is 1). On the other hand, when the jump activity is not bigger than some $r \in [1,2]$, a general rate-minimax result (on suitably bounded classes of Itô semimartingales) is $\big(\frac{\log(1/\Delta_n)}{\Delta_n}\big)^{(2-r)/r}$, see [3].

Despite this rate-minimax result, we show that under some specific structural assumptions on the jump part of X, it is possible to find estimators with rate $1/\sqrt{\Delta_n}$ even when the activity index is bigger than 1. Namely we suppose that X has the form

$$X_t = X_0 + \int_0^t b_s\,ds + \int_0^t \sigma_s\,dW_s + U_t + V_t, \tag{1}$$

where b_t is locally bounded adapted, and σ_t is itself an Itô semimartingale with "locally bounded" coefficients (it can jump), and U_t is a pure jump Itô semimartingale with degree of activity $r < 1$ (and some local boundedness of its Lévy measure). So when $V_t = 0$ we are in the setting where truncated realized volatility and multipower variations achieve the rate $1/\sqrt{\Delta_n}$.

The novelty here lies in the additional process V_t, which has the form

$$V_t = \sum_{m=1}^{M} \int_0^t \gamma^m_{s-}\,dY^m_t. \tag{2}$$

Here the Y^m are independent pure jump symmetric Lévy processes, with Lévy measures F^m satisfying $\big|F^m((x,\infty)) - 1/x^{\beta}_m\big| \le K/x^r$ for all $x \in (0,1]$ and r as before and K a constant, and $1 \le \beta_M < \cdots < \beta_1$. Note that the β_m's, as well as the γ^m_t-s are unknown. We need basically the same assumption on the γ^n's than on σ, namely that they are Itô semimartingales with locally bounded coefficients.

The estimators for C_t are based on estimators for the (real part of the) empirical characteristic function of the returns $\Delta^n_i = X_{i\Delta_n} - X_{(i-1)\delta_n}$, evaluated on successive time windows of size $k_n\Delta_n$ for a suitable sequence k_n of integer going

to infinity, but such that $k_n\Delta_n \to 0$. For any $u > 0$ we set

$$(3)\qquad \begin{aligned} L(u)^n_j &= \tfrac{1}{k_n}\sum_{l=0}^{k_n-1}\cos(u\Delta^n_{1+jk_n+l}X/\sqrt{\Delta_n}\,)\\ \widehat{c}(u)^n_j &= -\tfrac{2}{u^2}\log\Big(L(u)^n_j \bigvee \tfrac{1}{\sqrt{k_n}}\Big),\end{aligned}$$

so $\widehat{c}(u)^n_j$ serves as a local estimator of the average of $c_t = \sigma_t^2$ over the interval $(jk_n\Delta_n, (j+1)k_n\Delta_n]$, and our first estimator for C_t will be

$$(4)\qquad \widehat{C}(u)^n_t = k_n\Delta_n\sum_{j=0}^{[t/k_n\Delta_n]-1}\Big(\widehat{c}(u)^n_j - \frac{1}{u^2k_n}\big(\sinh(u^2\widehat{c}(u)^n_j)/\big)^2\Big).$$

For stating the asymptotic behavior we need some notation. We set

$$\chi(\beta) = \int_0^\infty \frac{\sin y}{y^\beta}\,dy, \qquad A^m_t = 2\chi(\beta_m)\int_0^t |2\gamma^{m+}_s|^{\beta_m}\,ds$$

$$(5)\qquad Z(u)^n_t = \widehat{C}(u)^n_t - C_t - \sum_{m=1}^M u^{\beta_m-2}\Delta_n^{1-\beta_m/2}A^m_t.$$

The key Central Limit Theorem, proved in [4], is then as follows:

Theorem 1. *Choose any $\theta_i > 0$ and any $Q \times L$ matrix α_{ij} with $\sum_{l=1}^L \alpha_{i,l} = 0$., and also two sequences k_n (of integers) and u_n (of positive reals) as follows:*
• *without prior knowledge on β_1 and r, then*

$$k_n\sqrt{\Delta_n} \to 0, \qquad k_n\Delta_n^{1/2-\varepsilon} \to \infty\ \forall\varepsilon > 0, \quad u_n \to 0, \quad \frac{k_n\sqrt{\Delta_n}}{u_n^2} \to 0;$$

• *if we know that $\beta_1 \le \beta_0$ and $r \le r_0$ for some given $\beta_0 \in [1,2)$ and $r_0 < 1$, then*

$$k_n \asymp \tfrac{1}{\Delta_n^\varpi}, \quad u_n \asymp \Delta_n^{\varpi'}, \quad \text{where } \tfrac{1}{3}\bigvee\tfrac{2-\beta_0}{2} < \varpi < \tfrac{1}{2}$$
$$\text{and } 0 < \varpi' < \tfrac{1-2\varpi}{8}\bigwedge\tfrac{1-r_0}{8(1-r_0+\beta_0)}\bigwedge\tfrac{(1-\beta_0+2\varpi)\wedge(1-\beta_0\varpi)}{8-2\beta_0}\bigwedge\tfrac{2-\beta_0}{(8\beta_0)\vee 12}$$

Then the $(Q+1)$-dimensional processes with components $\frac{1}{\sqrt{\Delta_n}}Z(u_n)^n$ and $\frac{1}{u_n^2\sqrt{\Delta_n}}\sum_{l=1}^L\alpha_{q,l}\,Z(\theta_l u_n)^n$ converge (functionally) stably in law to a limit which is defined on an extension $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ of the space $(\Omega,\mathcal{F},\mathbb{P})$ and is, conditionally on $\mathcal{F}$, a continuous Gaussian centered martingale, and the conditional variance of the first component Z of the limit is

$$(6)\qquad \widetilde{\mathbb{E}}((Z_t)^2 \mid \mathcal{F}) = 2\int_0^t c_s^2\,ds.$$

When $\gamma^m \equiv 0$ for all m, that is (1) holds with $V_t = 0$, the de-biasing terms in (5) disappear, and the estimators $\widetilde{C}(u_n)_T$ converge with rate $1/\sqrt{\Delta_n}$ and are asymptotically efficient (as is the truncated realized volatility) , since (6) gives the minimal estimating (conditional) variance for estimating C_T. Moreover in this

case and if we know that $r \leq r_0$, the conditions on ϖ, ϖ' become $\frac{1}{3} < \varpi < \frac{1}{2}$ and $0 < \varpi' < \frac{1-2\varpi}{4} \bigwedge \frac{1-r_0}{8}$.

Remark: If instead of taking $u_n \to 0$ we take $u_n = u > 0$ a constant, the same result holds when $r \leq 1$ instead of $r < 1$, except that the variance in (6) becomes $8 \int_0^t \left(\frac{\sinh(u^2 c_s/2)}{u^2}\right)^2 ds$, so we keep the rate $1/\sqrt{\Delta_n}$ but loose asymptotic variance-efficiency.

Note that the bias in (5), once divided by $\sqrt{\Delta_n}$, goes to infinity. However, it is possible to de-bias, and for this we choose any number $\zeta > 1$ and set (with $T > 0$ fixed)

$$\widehat{C}(u,\zeta)^n_T = \widehat{C}u)^n_T - \frac{(\widehat{C}(\zeta u)^n_T - \widehat{C}^n_T)^2}{\widehat{C}(\zeta^2 u)^n_T - 2\widehat{C}(\zeta u)^n_T + \widehat{C}(u)^n_T}. \tag{7}$$

This one-step procedure is enough when $M = 1$ in (2):

Theorem 2. *If $M = 1$ and $C_T > 0$ almost surely, and with k_n, u_n as in Theorem 1, the variables $\frac{1}{\sqrt{\Delta_n}}(\widehat{C}(u,\zeta)^n_T - C_T)$ converge stably in law to the variable Z_T of that theorem.*

The additional assumption $C_T > 0$ a.s. here, which is necessary to ensure that the denominator in (7) is asymptotically "non-degenerate" in an appropriate sense.

When $M \geq 2$ we need to "iterate" the previous de-biasing method, and this can be done only under an additional assumption, which basically requires the β_m's to be on a lattice. This is of course a very strong assumption, from a theoretical viewpoint, but practically speaking it seems rather innocuous:

> The numbers $2 - \beta_m$ all belong to the set $\{j\rho : j = 1, 2, \cdots\}$
> for some (unknown) constant $\rho \in (0, 1)$, so necessarily $M \leq [1/\rho]$

Then the de-biasing procedure goes as follows, for some integer N:

(1) - initialization: Choose a real $\zeta > 1$ and put $\widehat{C}(u,\zeta,0)^n_T = \widehat{C}(u)^n_t$.

(2) - iteration: Assuming $\widehat{C}_n(u,\zeta,j-1)$ known for some integer j between 1 and N, define (similar with (7)): $\widehat{C}_n(u,\zeta,j)$ as

$$\widehat{C}(u,\zeta,j)^n_T = \widehat{C}_n(u,\zeta,j-1)^n_T + \frac{(\widehat{C}_n(\zeta u,\zeta,j-1)^n_T - \widehat{C}(u,\zeta,j-1)^n_T)^2}{\widehat{C}(\zeta^2 u,\zeta,j-1)^n_T - 2\widehat{C}(\zeta u,\zeta,j-1)^n_T + \widehat{C}(u,\zeta,j-1)^n_T}.$$

(3) - end: The final estimator is set to be $\widehat{C}(u_n,\zeta,N)^n_T$.

The following is then a (relatively) simple consequence of Theorem 1:

Theorem 3. *Assume (8) with $\rho \geq \rho_0$ for some $\rho_0 \in (0, 1)$, and let N be the biggest integer such that $N\rho_0 \leq 1$. Assume also that $C_T > 0$ almost surely, and choose k_n, u_n as in Theorem 1, with $\beta_0 = 2 - \rho_0$ in the second case. Then the variables $\frac{1}{\sqrt{\Delta_n}}(\widehat{C}(u_n,\zeta,N)^n_T - C_T)$ converge stably in law to the variable Z_T described above.*

Finally, one can relax the symmetry assumption on the Y^m's: by this, we mean that the Lévy measures F^m satisfy $\left|F^m((x,\infty)) - a^m_+/x^\beta_m\right| \leq K/x^r$ and

$\left|F^m((-\infty,-x)) - a_-^m/x_m^\beta\right| \leq K/x^r$ for all $x \in (0,1]$, with a_+^m and a_-^m possibly different (one of them can even vanish). Setting

$$L(u)_j^n = \frac{1}{k_n}\sum_{l=0}^{k_n-1}\cos(u\Delta^n_{1+2jk_n+2l}X/\sqrt{\Delta_n} - u\Delta^n_{2+2jk_n+2l}X/\sqrt{\Delta_n}) \tag{8}$$

(a symmetrized version of (3)), we sum up from 1 to $[t/2k_n\Delta_n]$ in (4). Then all previous results hold, except that in (6) we need the multiplicative factor 2, so we loose variance-efficiency.

If we still use the original version (3)-(4), the bias term in Theorem 1 is more complicated. As it turns out, when $M = 1$ the result stated in Theorem 2 fails, but one can iterate the procedure as above and get a result similar to Theorem 3, see [4]. When $M > 1$, though, the iteration never ends: at each step there is still (in general) a bias which, once divided by $\sqrt{\Delta_n}$, goes to infinity.

From a practical viewpoint, the iteration procedure is probably quite unstable. So if we know that $M = 1$, and in the non-symmetric case, it is advisable to use the one step bias-correction with the version (8), despite the efficiency loss.

References

[1] Barndorff-Nielsen, O.E. and N. Shephard (2004). Power and bipower variation with stochastic volatility and jumps. *J. of Financial Econometrics* **2**, 1–48.

[2] Barndorff-Nielsen, O.E., N. Shephard and M. Winkel (2006). Limit theorems for multipower variation in the presence of jumps. *Stochastic Processes and their Applications* **116**, 796-806.

[3] Jacod, J. and M. Reiss (2013). A remark on the rates of convergence for integrated volatility estimation in the presence of jumps. Preprint.

[4] Jacod, J. and V. Todorov (2013). Efficient estimation of integrated volatility in presence of infinite variation jumps. Preprint.

[5] Mancini, C. (2001). Disentangling the jumps of the diffusion in a geometric jumping Brownian motion. *Giornale dell'Istituto Italiano degli Attuari* **LXIV**, 19–47.

[6] Vetter, M. (2010). Limit theorems for bipower variation of semimartingales. *Stoch. Proc. Appl.* **120**, 22–38.

Change-Points in High-Dimensional Settings

Claudia Kirch

(joint work with John A D Aston)

While there is considerable work on change-point analysis in univariate time series, more and more data being collected comes from high dimensional multivariate settings, where the number of components is of the same order or even larger than the number of time points. An appropriate asymptotic framework to investigate statistical procedures for such data assumes that the number of components increases to infinity with the number of time points. In this setup we would like to investigate the properties of univariate tests after the data has been projected onto a vector $\boldsymbol{p}_d$. To this end, we consider the following model:

$$X_{i,t} = \mu_i + \delta_{i,T}\,1_{\{t>\lfloor\vartheta T\rfloor\}} + e_{i,t}, \quad 1 \leq i \leq d = d_T, 1 \leq t \leq T,$$

where (for simplicity) $\{(e_{1,t},\dots,e_{d,T})^T, t=1,\dots,T\}$ is i.i.d. and $0<\vartheta<1$ is the rescaled change-point. We call the vector $\boldsymbol{\Delta}_d=(\delta_{1,T},\dots,\delta_{d,T})^T$ the change and test

$$H_0:\boldsymbol{\Delta}_d=0, \qquad H_1:\boldsymbol{\Delta}_d\neq 0.$$

In this setting, it is apparent that the change $\boldsymbol{\Delta}_d$ is always a one-dimensional object no matter the number of components d. This observation suggests that knowledge about where the change-point is located in addition to the underlying covariance structure can significantly increase the signal-to-noise ratio. In applications, certain changes are either expected or of particular interest e.g. an economist looking at the performance of several companies expecting changes caused by a recession will have a good idea which companies will profit or lose. This knowledge can then be used to increase the power in directions close to the search direction $\boldsymbol{p}_d$ while decreasing it for changes that are close to orthogonal to it.

In order to understand this informal statement better and to compare the power behavior of different statistics, we consider contiguous changes, where $\|\boldsymbol{\Delta}_d\|\to 0$ but with such a rate that the power of the corresponding test is strictly between the size and one. We can then compare these **contiguous rates** to understand the power of the test. Concerning a fixed projection $\boldsymbol{p}_d$ it turns out that the contiguous rate is given by

$$T\,\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2\cos^2(\alpha_{\Sigma^{-1/2}\boldsymbol{\Delta}_d,\Sigma^{1/2}\boldsymbol{p}_d}),$$

where Σ is the covariance of the vector $(e_{1,t},\dots,e_{d,T})^T$ and $\alpha_{\boldsymbol{a},\boldsymbol{b}}$ is the smallest angle between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. From this it is obvious that the **oracle** projection $\boldsymbol{o}=\Sigma^{-1}\boldsymbol{\Delta}_d$ maximizes the contiguous rate. This can be compared to a random projection on the unit sphere after standardizing the data, which is equivalent to projecting with the vector $\boldsymbol{r}_{d,\Sigma}=\Sigma^{-1/2}\boldsymbol{r}_d$, where $\boldsymbol{r}_d$ is a random projection on the unit sphere. Furthermore, we can compare the procedure with a generalization of multivariate change-point procedures for independent components in the above asymptotic framework proposed by Horváth and Hušková [2].

The following table compares the contiguous rates in all three cases:

	Contiguous Rate
Oracle projections	$T\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2$
HH statistic ($\Sigma=\mathrm{Id}$)	$T\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2/\sqrt{d}$
Scaled random projection	$T\|\Sigma^{-1/2}\boldsymbol{\Delta}_d\|^2/d$ (stochastic order)

It becomes apparent that we lose an order $\sqrt{d}$ between the oracle and the HH statistic as well as another order $\sqrt{d}$ between the HH statistic and the scaled random projection. Figure 1 confirms these theoretical findings and gives an impression on how wide the angle between $\Sigma^{-1/2}\boldsymbol{\Delta}_d$ and $\Sigma^{1/2}\boldsymbol{p}_d$ can be before the HH procedure is better than the projection. Please note, however, that the space covering these angles increases for increasing dimensions.

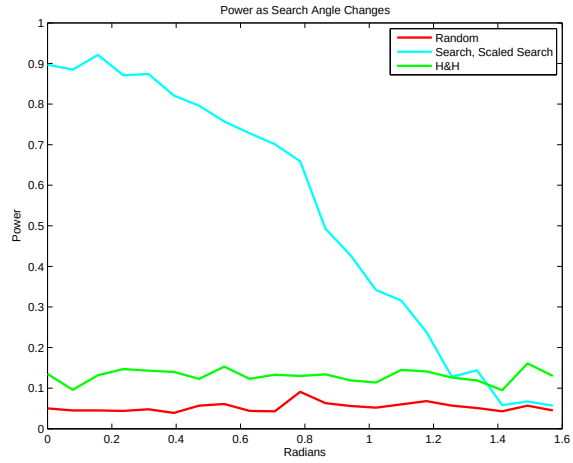


FIGURE 1. Empirical size-corrected power for increasing angles, $\Sigma = \mathrm{Id}$

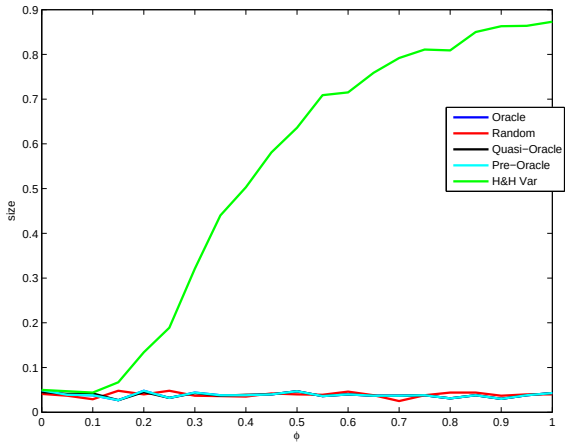


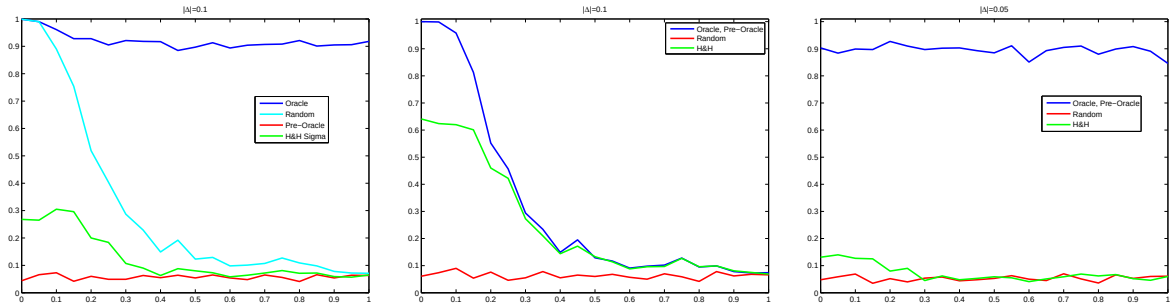
FIGURE 2. Empirical size for increasing contamination by a common factor

Usually, in applications Σ is not known and needs to be estimated, which is rather problematic particularly in high-dimensional settings without additional parametric or sparsity assumptions. For change-point tests the inverse is needed which results in additional numerical problems for large d. Consequently, it is of importance to check the robustness of the procedures with respect to not knowing Σ.

To this end, we first consider the size of the different procedures. For the projection procedures and a large class of dependency across components only the variance of the projected sequence is needed, which is not difficult to estimate. The HH procedure on the other hand strongly depends on the independence between components or after some possible extensions on the knowledge of Σ^{-1}. Consequently, it suffers sincere size problems if Σ is misspecified. In order to show this effect we consider the situation where $e_{i,t} = s_i\eta_{i,t} + \Phi\xi_i$, where $\eta_{i,t}$ are independent and standardized and ξ_i is a common standardized disturbance factor across all channels (independent of η). Figure 2 clearly shows that the projection is much more robust with respect to size.

Considering contiguous rates again we can also investigate the robustness in terms of the power of the different procedures. To this end, we consider the **pre-oracle** ${}_p\mathbf{o} = \mathbf{\Delta}_d$ as well as the **quasi-oracle** ${}_q\mathbf{o} = (\delta_1/\operatorname{var}(e_{1,1}), \ldots, \delta_d/\operatorname{var}(e_{d,1}))^T$. If the Variances are all of the same order, i.e. $0 < c \leq \operatorname{var}(e_{i,1}) \leq C < \infty$, then in the uncorrelated case quasi- and pre-oracle are of the same order, in the general case both of them are always at least as good as the unscaled random projection $\boldsymbol{r}_d$ but can be better, while the HH procedure is always of the same order as the random projection. This fact is confirmed by the simulations in Figure 3.

In summary, projections can greatly increase the power of corresponding change-point tests in high-dimensional settings particularly if the covariance structure is accessible and some information about the location of the change of interest is



$\boldsymbol{\Delta}_d \sim (1,\ldots,1,0,\ldots,0)^T$, ${}_p\mathbf{o} \sim {}_q\mathbf{o}$ | $\boldsymbol{\Delta}_d \sim (1,\ldots,1,1,\ldots,1)^T$, $\boldsymbol{o} \sim {}_p\mathbf{o} \sim {}_q\mathbf{o}$ | $\boldsymbol{\Delta}_d \sim (1,\ldots,1,-1,\ldots,-1)^T$, ${}_p\mathbf{o} \sim {}_q\mathbf{o}$

FIGURE 3. Empirical size-corrected power for increasing contamination by a common factor, $s_j = 1$

known. Additionally, such projections are much more robust with respect to both size and power than competing fully multivariate procedures if the covariance structure is misspecified.

REFERENCES

[1] J.A.D. Aston, C. Kirch, *Change-poins in high-dimensional settings*, in preparation, 2013.
[2] L. Horváth, M. Hušková, *Change-point detection in panel data*, J. Time Series Anal. **33** (2012), 631-648.

Estimation of spatial weight matrices in large spatial lag/error panels

CLIFFORD LAM
(joint work with Pedro CL Souza)

Spatial lag/error models are used for incorporating spatial dependence explicitly among the components of a panel. They are commonly used in fields such as economics, social science, or biology, where observations are dependent on each other on top of individual characteristics. An important element of these models is the *spatial weight matrix* that describes how the components depend on each other. Unfortunately, in most applications, it needs to be explicitly specified either by expert knowledge, or by a proxy, which can give rise to estimation error if the specification is not good enough. Majority of papers assume these as known matrices, and go on to estimate other parameters of the model. In [1], we argue that wrong specification of the spatial weight matrix can lead to inaccurate estimation of the regression parameters in the model, and quantify how serious this can be. Many practitioners just use rough proxies for the spatial weight matrix, which can be too rough for accurate estimation of the model overall. Since in many applications the spatial weight matrix is sparse overall, we are motivated to estimate the sparse spatial weight matrix using LASSO penalization.

In more details, the model we consider for $\mathbf{y}_t$ of length N is

$$\mathbf{y}_t = \mathbf{W}_1^* \mathbf{y}_t + \mathbf{W}_2^* \mathbf{X}_t \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T,$$

where $\mathbf{W}_1^*, \mathbf{W}_2^*$ are spatial weight matrices with diagonal entries 0 and 1 respectively. This model entails the spatial error model. Put $\mathbf{W}_2^* = \mathbf{I} - \mathbf{W}_1^*$, and the model becomes $\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta}^* + (\mathbf{I} - \mathbf{W}_1^*)^{-1} \boldsymbol{\epsilon}_t$, which is of the form of a spatial error model. Using the compact "regression" form of the model

$$\mathbf{y} = \mathbf{M}_{\boldsymbol{\beta}^*} \boldsymbol{\xi}^* + \boldsymbol{\epsilon},$$

where $\mathbf{y} = \mathrm{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}\}$, $\boldsymbol{\epsilon} = \mathrm{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\mathrm{T}}\}$, $\boldsymbol{\xi}^* = (\mathrm{vec}(\mathbf{W}_1^{*\mathrm{T}})^{\mathrm{T}}, \mathrm{vec}(\mathbf{W}_2^{*\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{M}_{\boldsymbol{\beta}^*} = (\mathbf{Z}, \mathbf{X}_{\boldsymbol{\beta}^*})$ with $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}$, $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{I}_N \otimes \{(\mathbf{I}_T \otimes \boldsymbol{\beta}^{*\mathrm{T}})(\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\mathrm{T}}\}$, we formulate the problem as a LASSO penalization,

$$\begin{aligned} (\widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\beta}}) &= \arg\min_{\boldsymbol{\xi}, \boldsymbol{\beta}} \frac{1}{2T} ||\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}} \boldsymbol{\xi}||^2 + \gamma_T ||\boldsymbol{\xi}||_1, \\ &\text{subj. to } \sum_{j \neq i} |w_{1,ij}|, \sum_{j \neq i} |w_{2,ij}| < 1. \end{aligned} \tag{1}$$

The adaptive LASSO is also explored theoretically and it gives better sensitivity and specificity results for estimating the spatial weight matrices in practice. A block coordinate descent algorithm is developed for carrying out all computations. We proved asymptotic sign consistency of the elements in the spatial weight matrices, as well as giving error bounds for both the estimators of the spatial weight matrices and the regression parameter.

We have also considered the problem of identifying block structure in a model without covariates in [2], that is

$$\mathbf{y}_t = \mathbf{W} \mathbf{y}_t + \boldsymbol{\epsilon}_t.$$

This problem is closely related to graphical model estimation, but our aim is to find blocks in the spatial weight matrix. This is motivated by a set of US voting data where there are no obvious covariates, and block structure in the spatial weight matrix is anticipated because of political affiliations. We have proved that even when blocks are slightly overlapping in the spatial weight matrix, we can identify these blocks with probability approaching 1 in the absence of covariates. Indeed, our results show clearly that over the year 2012, the Republicans forms one block and the Democrats forms another in the spatial weight matrix, even though they overlap slightly.

Future directions include using instrumental variables with LASSO to relax the assumptions in [1] and improve the estimators. We also explore the adaptation of expert knowledge in specifying a spatial weight matrix, and make sparse adjustment to this using the data available. This way the practitioners can see if their knowledge on the spatial dependence structure and specification of the spatial weight matrix align with the data.

References

[1] Lam, C. and P. C. L. Souza (2013). Regularization for spatial panel time series using the adaptive lasso. Manuscript, submitted to the Journal of Econometrics.

[2] Lam, C. and P. C. L. Souza (2013). Detection And Estimation Of Block Structure In Spatial Weight Matrix. Manuscript, submitted to the Econometrics Review.

Testing for GARCH(1,1) model specification

Anne Leucht

(joint work with Michael H. Neumann and Jens-Peter Kreiß)

Conditionally heteroscedastic time series are frequently used in the finance literature to model the evaluation of stock prizes, exchange rates and interest rates. The question of parameter estimation in these models has been studied intensively. In contrast, we focus on a test for the GARCH(1,1) model.

There is already an overwhelming amount of model specification tests in the econometric literature. However, these methods typically rely on the assumption that the information variables as well as the response variables are observable. This condition is violated in the case of GARCH models, where unobserved quantities enter the information variable. Hence, standard tests cannot be applied and certain additional approximation procedures have to be invoked. It turns out that the literature on specification tests for conditionally heteroscedastic time series is comparatively rare. [1] proposed a Portmanteau goodness-of-fit test for GARCH(1,1) models. Their test statistic is a quadratic form of weighted autocorrelations of the squared residuals of a GARCH(1,1) process fitted to the data, whose dimension increases with the sample size. They showed that its limit distribution is an (infinite) weighted sum of independent χ^2_1-distributed random variables under the null hypothesis but did not consider the behavior under alternatives.

We propose a specification test of Cramér-von Mises type for a GARCH(1,1) hypothesis against general alternatives. Here, we face the particular problem that some of the explanatory variables are not observed and have to be approximated. It turns out that our test statistic can be approximated by a V-statistic and it follows from results of [4] that the latter converges to a weighted sum of independent χ^2_1 variables. In contrast to [1], where the weights in the limit correspond to the weights in the test statistic itself, here these quantities depend on the properties of the underlying process in a complicated way. Therefore, the asymptotic result cannot be used for determining an appropriate critical value.

We propose to apply a model-based bootstrap method to approximate the null distribution of the test statistic which eventually yields an appropriate critical value for the test. [4] prove consistency of model-based bootstrap for statistics that can be approximated by a V-statistic. In contrast to the method of proof used in [4], we present a different approach of proving bootstrap consistency: Rather than imitating the derivation of the limit distribution of the test statistic also on the bootstrap side, we use coupling arguments to show consistency. This approach was successfully applied to U- and V-statistics of independent random variables

by [2] and [3], however, it seems to be new in the context of dependent data. Simulations indicate a desirable finite sample behavior of our test under the null as well as under certain asymmetric alternatives.

Finally, we conjecture that our theory can be generalized to GARCH models of higher order and to augmented GARCH processes. To present the main ideas in a transparent manner, we restrict ourselves to the simple GARCH(1,1) case.

References

[1] Berkes, I., Horváth, L., and Kokoszka, P., *A weighted goodness-of-fit test for GARCH(1,1) specification.* Lithuanian Mathematical Journal **44** (2004), 3–22.

[2] Dehling, H. and Mikosch, T., *Random quadratic forms and the bootstrap for U-statistics.* Journal of Multivariate Analysis **51** (1994), 392–413.

[3] Leucht, A. and Neumann, M. H., *Consistency of general bootstrap methods for degenerate U- and V-type statistics.* Journal Multivariate Analysis **100** (2009), 1622–1633.

[4] Leucht, A. and Neumann, M. H., *Degenerate U- and V-statistics under ergodicity: asymptotics, bootstrap and applications in statistics.* Annals of the Institute of Statistical Mathematics **65** (2013), 349–386.

Inference for Lévy driven continuous time moving average processes

Alexander Lindner

(joint work with Serge Cohen)

Let $L = (L_t)_{t\in\mathbb{R}}$ be a two-sided Lévy process, i.e. a process with independent and stationary increments, and such that $L_0 = 0$ and that L has càdlàg paths almost surely. Assume further that L has zero mean, finite variance $\sigma^2 \in (0,\infty)$ and that $f : \mathbb{R} \to \mathbb{R}$ is a kernel function which is in $L^2(\mathbb{R})$. Let $\mu \in \mathbb{R}$. Then the process $(X_t)_{t\in\mathbb{R}}$, given by

$$X_t = \mu + \int_{\mathbb{R}} f(t-s)\, dL_s, \quad t \in \mathbb{R},$$

can be defined in an L^2-sense and is called a *continuous time moving average process with mean μ and kernel function f, driven by L.* The aim of this talk, which was based on the paper [3], was to derive a central limit theorem for the sample mean of X and the sample autocovariance functions of X when sampled at a discrete grid size Δ, which for convenience we take as $\Delta = 1$. Denote the autocovariance function of the process X by

$$\gamma(h) = \mathrm{Cov}(X_h, X_0) = \sigma^2 \int_{-\infty}^{\infty} f(-s) f(h-s)\, ds, \quad h \in \mathbb{R}.$$

Regarding the sample mean, we have the following result:

Theorem 1 [3, Thm. 2.1]
Let L have zero mean and variance $\sigma^2 \in (0,\infty)$ and let $\mu \in \mathbb{R}$. Suppose that

$$\left(F_\Delta : [0,1] \to [0,\infty], \quad u \mapsto F(u) := \sum_{j=-\infty}^{\infty} |f(u+j)| \right) \in L^2([0,1]).$$

Then $\sum_{j=-\infty}^{\infty} |\gamma(j)| < \infty$,

$$v := \sum_{j=-\infty}^{\infty} \gamma(j) = \sigma^2 \int_0^1 \left(\sum_{j=-\infty}^{\infty} f(u+j) \right)^2 du,$$

and the sample mean $\overline{X}_n := n^{-1}(X_1 + \ldots + X_n)$ is asymptotically normal with mean μ and variance v/n as $n \to \infty$, i.e.

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, v), \quad n \to \infty.$$

This is very similar to the discrete time setting of infinite moving average processes driven by i.i.d. noise, cf. Brockwell and Davis [2, Thm. 7.1.2]. The situation is however different for the asymptotic analyis of the sample autocovariance and sample autocorrelation. Suppose that $\mu = 0$, so that $EX_t = 0$ for all t. Denote

$$\gamma_n^*(h) := n^{-1} \sum_{t=1}^{n} X_t X_{t+h}, \quad h \in \mathbb{N}_0,$$

which is the (modified) sample autocovariance function of X based on observations $X_1, \ldots, X_n$ (and $X_{n+1}, \ldots, X_{n+h}$). Denote the sample autocorrelation by

$$\rho_n^*(h) := \frac{\gamma_n^*(h)}{\gamma_n^*(0)},$$

which is an estimator for the autocorrelation $\rho(h) := \frac{\gamma(h)}{\gamma(0)}$. Then we have the following theorem:

Theorem 2 [3, Thm. 3.5]
Suppose the Lévy process L satisfies $EL_1 = 0$, $\sigma^2 = EL_1^2 \in (0,\infty)$, and that $\eta := \sigma^{-4} EL_1^4 < \infty$. Suppose the kernel function f is in $L^2(\mathbb{R}) \cap L^4(\mathbb{R})$ and that

$$\left(G : [0,1] \to [0,\infty], \quad u \mapsto \sum_{j=-\infty}^{\infty} f(u+j)^2 \right) \in L^2([0,1]).$$

Suppose further that

$$\sum_{j=-\infty}^{\infty} \left(\int_{\mathbb{R}} |f(s) f(s+j)| \, ds \right)^2 < \infty.$$

Then the function

$$g_q : [0,1] \to \mathbb{R}, \quad u \mapsto \sum_{j=-\infty}^{\infty} f(u+j)f(u+j+q)$$

is in $L^2([0,1])$ *for each* $q \in \mathbb{Z}$, $\sum_{j=-\infty}^{\infty} \gamma(j)^2 < \infty$, *and the sample autocorrelation* ρ_n^* *of the process* $X_t = \int_{\mathbb{R}} f(t-s)\, dL_s$ *based on the observations* $X_1, \ldots, X_n$ *satisfies*

$$\sqrt{n}(\rho_n^*(1) - \rho(1), \ldots, \rho_n^*(h) - \rho(h)) \xrightarrow{d} N(0, W), \quad n \to \infty,$$

where $W = (w_{ij})_{i,j=1,\ldots,h}$ *is given by*

$$w_{ij} = \widetilde{w}_{ij} + \frac{(\eta - 3)}{\gamma(0)^2} \int_0^1 \big(g_i(u) - \rho(i)g_0(u)\big)\big(g_j(u) - \rho(j)g_0(u)\big)\, du,$$

and

$$\begin{aligned}\widetilde{w}_{ij} \quad = \quad & \sum_{k=-\infty}^{\infty} \big[\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) \\ & +2\rho(i)\rho(j)\rho(k)^2 - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\big]\end{aligned}$$

is given by Bartlett's formula.

This theorem is quite in contrast to the corresponding discrete time result (cf. Brockwell and Davis [2, Thm. 7.2.1]). While there the asymptotic variance of the sample autocorrelation is given by Bartlett's formula, here we need the correction term

$$\frac{(\eta - 3)}{\gamma(0)^2} \int_0^1 \big(g_i(u) - \rho(i)g_0(u)\big)\big(g_j(u) - \rho(j)g_0(u)\big)\, du.$$

It is easy to see that, unless $\eta = 3$ which corresponds to a Brownian motion, the correction term may be non-zero. For example, consider the function $f = \mathbf{1}_{(0,1/2]} + \mathbf{1}_{(1,2]}$. Then $g_1 = \mathbf{1}_{(0,1/2]}$ and $g_0 = 2 \cdot \mathbf{1}_{(0,1/2]} + \mathbf{1}_{(1/2,1]}$ and it is easy to see that $g_1 - \rho(1)g_0$ is not almost everywhere zero, so that $w_{11} \neq \widetilde{w}_{11}$ if $\eta \neq 3$. This is an interesting example, since it corresponds to a discrete time moving average process on the grid $\frac{1}{2}\mathbb{Z}$ with i.i.d. noise when sampled only at integer times. A more detailed analysis of this phenomenon in discrete time can be found in Niebuhr and Kreiß [5].

The case when $f = \sum_{j=-\infty}^{\infty} \psi_j \mathbf{1}_{(j,j+1]}$ deserves special attention. This corresponds to a discrete time moving average process on $\mathbb{Z}$ when sampled at integer times, and indeed it is easy to show that $w_{ij} = \widetilde{w}_{ij}$ so that the correction term is equal to zero in that case.

An application to fractional Lévy noise. Theorem 2 can be applied to derive an estimator of the Hurst index of fractional Lévy noise. Let L be a Lévy process with mean zero and finite variance $\sigma^2 \in (0, \infty)$, and define the *fractional Lévy process*

$$M_t^d := \frac{1}{\Gamma(d+1)} \int_{-\infty}^{\infty} \big[(t-s)_+^d - (-s)_+^d\big]\, dL_s, \quad t \in \mathbb{R},$$

where $d \in (0, 1/2)$ and $H := d + 1/2$ is the *fractional Hurst parameter* (cf. Marquardt [4]; different types of fractional Lévy processes can also be defined, cf. Benassi et al. [1]). Then the *fractional Lévy noise* is given by

$$X_t := M_t^d - M_{t-1}^d, \quad t \in \mathbb{R}.$$

Then X_t can be written as a continuous time moving average process

$$X_t = \int_{\mathbb{R}} \left[(t-s)_+^d - (t-s-1)_+^d \right] \, dL_s, \quad t \in \mathbb{R},$$

$EX_t = 0$ and

$$E(X_h X_0) = \gamma(h) = \frac{C\sigma^2}{2} \left(|h+1|^{2d+1} - 2|h|^{2d+1} + |h-1|^{2d+1} \right),$$

for some constant C depending on d. In particular,

$$\gamma(0) = C\sigma^2, \gamma(1) = \frac{C\sigma^2}{2}(2^{2d+1} - 2),$$

hence $\rho(1) = 2^{2d} - 1$ so that

$$d = \frac{1}{2} \left(\frac{\log(\rho(1)+1)}{\log 2} \right).$$

Hence the sample autocorrelation $\rho_n^*(1)$ based on $X_1, \ldots, X_n$ leads to a moment estimator

$$\widehat{d} := \frac{1}{2} \left(\frac{\log(\rho_n^*(1)+1)}{\log 2} \right).$$

It can be shown that this estimator is strongly consistent for $d \in (0, 1/2)$, and an application of Theorem 2 shows that it is asymptotically normal if L has finite fourth moment and $d \in (0, 1/4)$. Observe that fractional Lévy noise is in general not strongly mixing.

If $d \in [1/4, 1/2)$, then Theorem 2 is not applicable to X_t, but it can be applied to the differenced fractional noise

$$Y_t = X_t - X_{t-1} = M_t^d - 2M_{t-1}^d + M_{t-2}^d,$$

and an asymptotically normal estimator for Y can be derived. See [3] for details.

References

[1] A. Benassi, S. Cohen and J. Istas, *On roughness indices for fractional fields*, Bernoulli **10** (2004), 357–373.

[2] P.J. Brockwell and R.A. Davis, *Time Series. Theory and Methods*, (1987), Springer, New York.

[3] S. Cohen and A. Lindner, *A central limit theorem for the sample autocorrelations of a Lévy driven continuous time moving average process*, J. Statist. Planning Inference **143** (2013), 1295–1306.

[4] T. Marquardt, *Fractional Lévy processes with an application to long memory moving average processes*, Bernoulli **12** (2006), 1099–1126.

[5] T. Niebuhr and J.-P. Kreiß, *Asymptotics for autocovariances and integrated periodograms for linear processes observed at lower frequencies*, International Statistical Review, to appear.

Asymptotics for the Parametric GARCH-in-Mean Model

Enno Mammen
(joint work with Christian Conrad)

The aim of this talk is to develop asymptotic theory for the Quasi-Maximum Likelihood Estimator in GARCH-in-Mean (GARCH-M) models for the special case of GARCH(1,1)-innovations. We will explain why the proof of asymptotic normality is so difficult in this simple classical parametric model. There is a mathematical motivation to look at this model because difficulties in the study of the model come from nonstationarities of derivatives of the likelihood function which creates some nonstandard mathematical difficulties. There is also some applied statistical motivation because the model is related to some recent proposals in empirical finance.

The GARCH-in-Mean (GARCH-M) model was proposed in [2]. The parametric GARCH(1,1)-M model is given by

$$\begin{aligned} Y_t &= m_\gamma(h_t(\theta)) + \varepsilon_t \\ \varepsilon_t &= \sqrt{h_t(\theta)} Z_t \\ h_t(\theta) &= \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}(\theta), \end{aligned}$$

with $Z_t \overset{iid}{\sim} (0,1)$. Standard specifications for the mean functions are: $m_\gamma(h_t) = \mu + \lambda g(h_t)$ with $g(h_t) = h_t$, $= \sqrt{h_t}$ or $= \ln(h_t)$. The quasi-maximum likelihood estimator $\hat{\theta}$ is defined as $\hat{\theta} = \text{argmax}_{\theta \in \Theta} \hat{L}_T(\theta)$, where $\hat{L}_T(\theta)$ is the quasi-likelihood function:

$$\hat{L}_T(\theta) = -\frac{1}{2} \sum_{t=1}^{T} log(h_t(\theta)) + h_t(\theta)^{-1} \ (Y_t - m_\gamma(h_t(\theta)))^2$$

with $h_t(\theta) = \omega + \alpha(Y_{t-1} - m_\gamma(h_{t-1}(\theta)))^2 + \beta h_{t-1}(\theta)$. For proving a theorem on asymptotic normality of the quasi-maximum likelihood estimator $\hat{\theta}$ our proof has the standard structure: in a first step we show a rate of convergence for $\hat{\theta}$. And in a second step we use the localization of the problem to prove asymptotic normality of $\hat{\theta}$.

The basic idea of the first step is to use an approach based on stochastic recurrence equations, compare [1, 3, 4]. The essential assumptions, needed in the first step are:

$$\begin{aligned} &(A) \quad E[\ln(\alpha_0 \ Z_t^2 + \beta_0)] < 0, \\ &(B) \quad E[ln(U_t)] < 0, \quad D_1 < +\infty, \quad D_2 < +\infty, \end{aligned}$$

where $U_t = \sup_{\alpha,\beta} 2\alpha[D_1 Z_t \sqrt{h_t} + D_1 m_0(h_t) + D_2] + \beta$, $D_1 = \sup_{\gamma,u} |m'_\gamma(u)|$, $D_2 = \sup_{\gamma,u} |m_\gamma(u) m'_\gamma(u)|$. Assumption (A) is rather standard. It implies that there exists a stationary ergodic solution h_t of the GARCH equation. We will now explain why assumption (B) will be useful by similar reasons. Afterwards we will discuss how restrictive the assumption is. We start with a short discussion of stochastic recurrence equations.

Consider first the iteration equation for $h_t = h_t(\theta_0)$:

$$\begin{aligned} h_t &= \omega_0 + \alpha_0 \varepsilon_{t-1}^2 + \beta_0 h_{t-1} \\ &= \omega_0 + h_{t-1}(\alpha_0 Z_{t-1}^2 + \beta_0). \end{aligned}$$

This equation has a stationary ergodic solution if $E[\ln(\alpha_0\ Z_t^2 + \beta_0)] < 0$. Why does this hold? Consider two sequences h_t and h_t^* that fulfill the recurrence equation:

$$\begin{aligned} h_t &= \omega_0 + h_{t-1}(\alpha_0 Z_{t-1}^2 + \beta_0), \\ h_t^* &= \omega_0 + h_{t-1}^*(\alpha_0 Z_{t-1}^2 + \beta_0). \end{aligned}$$

Then $h_t - h_t^* = (h_{t-1} - h_{t-1}^*)(\alpha_0 Z_{t-1}^2 + \beta_0)$ and our condition $E[\ln(\alpha_0\ Z_t^2 + \beta_0)] < 0$ implies that

$$h_t - h_t^* \to 0 \text{ a.s.}$$

The approach of stochastic recurrence equations has been generalized w.r.t. two aspects:

- One can consider nonlinear recurrence equations. Then one needs conditions of the type $E[ln(\Lambda)] < 0$ where Λ is the (random) Lipschitz constant of the recurrence equation.
- Instead of real valued random variables one can consider random elements of function spaces.

We use this approach with the random functions

$$h_t(\theta) = \omega + \alpha(Y_{t-1} - m_\gamma(h_{t-1}(\theta)))^2 + \beta h_{t-1}(\theta).$$

Consider two sequences $h_t(\theta)$ and $h_t^*(\theta)$

$$\begin{aligned} h_t(\theta) &= \omega + \alpha(Y_{t-1} - m_\gamma(h_{t-1}(\theta)))^2 + \beta h_{t-1}(\theta), \\ h_t^*(\theta) &= \omega + \alpha(Y_{t-1} - m_\gamma(h_{t-1}^*(\theta)))^2 + \beta h_{t-1}^*(\theta). \end{aligned}$$

One can show the following Lipschitz inequality:

$$|h_t(\theta) - h_t^*(\theta)| \le U_t \|h_{t-1} - h_{t-1}^*\|_\infty$$

with U_t defined above and $\|...\|_\infty$ equal to the sup-norm. In our assumptions we had assumed that $E\ln(U_t) < 0$. This assumption implies that the recurrence equation

$$h_t(\theta) = \omega + \alpha(Y_{t-1} - m_\gamma(h_{t-1}(\theta)))^2 + \beta h_{t-1}(\theta)$$

has a stationary ergodic solution $h_t(\theta)$. In particular, these considerations show that the quasi-likelihood function converges to its expectation. This can be used to show consistency of the quasi-maximum likelihood estimator $\hat{\theta}$.

Unfortunately, the theory of recurrence equations cannot be used in the next steps of the proof. In this respect, the model differs form other GARCH-models. Our approach is based on first showing that for $\|\theta - \theta_0\| \le \delta$ with $\delta > 0$ small enough

$$\left| \frac{h_t(\theta) - h_t(\theta_0)}{h_t(\theta_0)} \right| \le \|\theta - \theta_0\| W_t$$

for some not too ill-behaved process W_t. In a further step this result, consistency of $\hat{\theta}$, the likelihood equation and empirical process theory is used to show that

$$\|\hat{\theta} - \theta_0\| = O_P(\log(T)T^{-1/2})$$

for some $\rho > 0$ large enough. Then one uses that $h_t'(\theta)$ and $h_t''(\theta)$ are well defined and well behaved for $\|\theta - \theta_0\| \leq C \log(T)T^{-1/2}$.

Finally, we arrive at the following asymptotic result for the quasi-maximum likelihood estimator:

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}), \text{ in distribution,}$$

where

$$\begin{aligned}
\Sigma_1 &= E\left[\frac{1}{2}\frac{h_t'(h_t')^\top}{h_t^2} + \frac{1}{h_t}(\dot{m}_{\gamma_0}(h_t) + m_{\gamma_0}'(h_t)h_t')(\dot{m}_{\gamma_0}(h_t) + m_{\gamma_0}'(h_t)h_t')^\top\right], \\
\Sigma_2 &= E\left[\left\{\frac{1}{2}\frac{h_t'}{h_t}(Z_t^2 - 1) + h_t^{-1/2}(\dot{m}_{\gamma_0}(h_t) + m_{\gamma_0}'(h_t)h_t')Z_t\right\}\right. \\
&\qquad \left.\left\{\frac{1}{2}\frac{h_t'}{h_t}(Z_t^2 - 1) + h_t^{-1/2}(\dot{m}_{\gamma_0}(h_t) + m_{\gamma_0}'(h_t)h_t')Z_t\right\}^\top\right].
\end{aligned}$$

For Gaussian Z_t we have $\Sigma_1 = \Sigma_2$ and we get that the asymptotic covariance is equal to Σ_1^{-1}.

Assumption (B) is rather restrictive. It is always fulfilled if $\beta < 1$ and D_1 and D_2 are small enough. The assumption $D_2 < +\infty$ states that our function m does not grow faster than $x \to a\sqrt{x}$. The treatment of functions with faster growth would require another approach. Consider e.g. the recurrence equation for the linear function $m_\gamma(x) = \gamma_1 + \gamma_2 x$. Here, we get that

$$h_t(\theta) - h_t(\theta_0) = \omega - \omega_0 + ... + \alpha\gamma_2^2[h_{t-1}(\theta) - h_{t-1}(\theta_0)]^2 + ...$$

It needs a very careful check why the quadratic term in the recurrence equation does not lead to an explosive behaviour during $0 \leq t \leq T$. The process is not stationary and explodes for $t \to \infty$!

Follow-up work includes discussions of what happens for m_γ that do not fulfill Assumption (B). Furthermore, we apply the results in empirical work using nonparametric testing and GARCH-M models with nonparametric mean specifications.

References

[1] Bougerol, P., *Kalman filtering with random coefficients and contractions*, SIAM J. Control Optim. **31** (1993), 942–959.

[2] Engle, R. F., D. M. Lilien, and R. P. Robins, *Estimating time varying risk premia in the term structure*, Journal of Business and Economic Statistics **9** (1987), 345–359.

[3] Straumann, D., *Estimation in conditionally heteroskedastic time series models*, Lecture Notes in Statistics **181** (2005), Springer, Berlin.

[4] Straumann, D. and Mikosch, T., *Quasi-maximum-likelihood estimation in conditionally heteroskedastic time series: a stochastic recurrence equations approach*, The Annals of Statistics **34** (2006), 2449–2495.

Boundary Spot Volatility Estimation using the Laplace Transform

Maria Elvira Mancino
(joint work with Imma Curato, Maria Cristina Recchioni)

We show that the Laplace transform is an appropriate tool for building estimators of the instantaneous volatility based on a long time series of prices by smoothing past data and retaining recent price observations. Our estimation procedure is non-parametric and model-free, given that we assume the asset price model to be a continuous Brownian semi-martingale.

The Laplace transform has the same advantages as the Fourier estimation procedure (see [6], [7]) with respect to the quadratic variation methods. In particular, by its definition, it uses all the available observations and avoids any manipulation of the original data, because it is based on the integration of the time series of returns rather than on its differentiation. Moreover, it has other good features which can be highlighted. In [6] it is shown that the Fourier estimator performs better in the center of the time window. The use of the Laplace transform allows us to obtain an estimator of instantaneous volatility which becomes less sensitive to the boundary effects as it approaches to the present time.

From a conceptual point of view the introduction of the Laplace transform has two advantages: firstly, it avoids the artificial "periodization" subjacent to Fourier series methodology, which is responsible for the low precision in the boundary estimation; secondly, it leads to an estimator which constitutes a bridge between the two different methods of computation of the volatility: the method based on quadratic variation and our approach using Fourier analysis. This link is confirmed also by the fact that the Laplace and Fourier estimators of the spot volatility include some kernel-based estimators when the lead-and-lag terms are neglected. In particular, due to the presence of a cross-product term with zero mean the Laplace and Fourier estimators of the spot volatility generalize respectively the triangular and the Fejer kernel based spot volatility estimators. The kernel-based estimators have been recently studied, e.g. in [2] and [3], the asymptotic results for the triangular estimator has been proven in [2] and the ability of the triangular kernels on the boundary of the time horizon has been shown in [3]. On the other hand, the relevance of the cross-product terms in the Fourier estimators for the integrated variance and the quarticity (i.e. the fourth power of the diffusion coefficient) is highlighted in [8], [9], where the robustness of the Fourier estimator in the presence of microstructure noise is shown.

We prove the consistency of the Laplace estimator of spot volatility and study its performance at the boundary of the observation interval, showing that it outperforms most estimators even in the presence of microstructure noise. Extensive numerical simulations carried out using high frequency data and stochastic volatility models (see [4], [5]) support this thesis. In fact, our estimator outperforms the Fourier estimator at the boundary of the time horizon while showing the same accuracy in terms of mean squared error and bias of its kernel-based component. Furthermore, in the presence of microstructure noise it performs as well as the

Fourier estimator for the spot volatility. Actually, the Laplace estimator is slightly more robust than the Fourier one with respect to the choice of bandwidths. We exploit the optimal bandwidths, in term of the smallest mean squared error, for these four estimators. To this end we use 1-second sampled data and two stochastic volatility models [4], [5] and we distort the data by introducing ever higher intensities of noise. We observe that Laplace estimator optimal bandwidths remain unchanged while the Fourier optimal bandwidths must be modified for high intensity noise; on the contrary, the triangular and the Fejer estimators deteriorate when the noise-to-signal ratio increases. Finally, the values of the optimal bandwidths are coherent with the theoretical result on the consistency of the estimator and with the results shown in [1], [2].

References

[1] Y. Ait_Sahalia, and J. Yu, *High frequency market microstructure noise estimates and liquidity markets*, The Annals of Applied Statistics **3**(1) (2009), 422-457.

[2] J. Fan, and Y. Wang, *Spot volatility estimation for high-frequency data*, Statistics and Its Interface **1** (2008), 279-288

[3] D. Kristensen, D. (2010). *Nonparametric filtering of the realized spot volatility: a kernel-based approach*, Econometric Theory, **26** (2010), 60-93.

[4] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Review of Financial Studies **6** (1993), 327–343.

[5] J. Hull, and A. White, *The pricing of options on assets with stcohastic volatilities*, Journal of Finance, **42** (1987), 281–300.

[6] P. Malliavin, M.E. Mancino, *Fourier series method for measurement of multivariate volatilities*, Finance and Stochastics *4* (2002), 49–61.

[7] P. Malliavin, and M.E. Mancino, *A Fourier transform method for nonparametric estimation of multivariate volatility*, The Annals of Statistics **37**(4) (2009), 1983–2010.

[8] M.E. Mancino, and S. Sanfelici, *Robustness of the Fourier estimator of integrated volatility in presence of microstructure noise*, Computational Statistics and Data Analysis **52**(6) (2008), 2966–2989.

[9] M.E. Mancino, and S. Sanfelici, *Estimation of quarticity with high-frequency data*, Quantitative Finance **12**(4) (2012), 607-622.

The extremogram and the ex-periodogram

Thomas Mikosch

(joint work with Richard A. Davis, Yuwei Zhao)

This is joint work with Richard A. Davis (Columbia University) and Yuwei Zhao (Copenhagen) (in Bernoulli Journal 2009 and 2013).

In this talk we consider the periodogram calculated from the indicators of extreme events based on an underlying stationary sequence (ex-periodogram). The ex-periodogram shares many properties with the classical periodogram of a stationary sequence such as asymptotic independence at distinct frequencies and consistent estmation of the (ex-)spectral density from weighted averages of ex-periodogram ordinates. We consider integrated versions of the ex-periodogram and investigate whether these statistics can be used for testing the goodness of fit of time series models based on their extremal behavior.

Forecasting Volatility using Leverage Effect

Per A. Mykland
(joint work with Dan Christina Wang and Lan Zhang)

The research provides a theoretical foundation for our previous empirical finding [3] that leverage effect has a role in forecasting volatility. This empirics is related to earlier econometric studies of news impact curves [1, 2]. Our new theoretical development is based on the concept of projection on stable subspaces of semi-martingales. We show that this projection provides a framework for forecasting (across time periods) that is internally consistent with the semi-martingale model which is used for the intra-day high frequency asymptotics.

References

[1] Xilong Chen and Eric Ghysels, *News - good or bad - and the impact on volatility predictions over multiple horizons*, Review of Financial Studies **24** (2011), 46-81.
[2] Robert F. Engle and Victor K. Ng, *Measuring and testing the impact of news on volatility*, Journal of Finance **48** (1993), 1749–1778.
[3] Dan Christina Wang and Per Mykland, *The Estimation of Leverage effect with High Frequency Data*, in revision for the Journal of the American Statistical Association.

Limit theorems for Lévy moving average processes

Mark Podolskij
(joint work with Andreas Basse-O'Connor, Raphael Lechieze-Rey)

We consider a Lévy moving average process of the form

$$X_t = \int_{-\infty}^{t} g(t-s)dL_s,$$

where L is a pure jump Lévy motion. The kernel function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is assumed to be of the type

$$g(x) = x^{\alpha} f(x), \qquad \alpha > 0,$$

where the function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is smooth, exponentially decaying at infinity and $f(0) \neq 0$. Since $g(0) = 0$ the process X turns out to be continuous and stationary.

Our main class of statistics are *power variations* that are defined via

$$V(X,p)_n := \sum_{i=1}^{n} |X_{\frac{i}{n}} - X_{\frac{i-1}{n}}|^p, \qquad p > 0.$$

In order to determine the first order asymptotics for $V(X,p)_n$, we need to introduce the *Blumenthal-Getoor index* of the Lévy process L. Let ν denote the Lévy measure of L. Then the Blumenthal-Getoor index β of L is defined as

$$\beta := \inf_{r \geq 0} \left\{ r : \int_{-1}^{1} |x|^r \nu(dx) < \infty \right\} = \inf_{r \geq 0} \left\{ r : \sum_{s \in [0,1]} |\Delta L_s|^r < \infty \right\} \in [0,2],$$

where $\Delta L_s = L_s - L_{s-}$ denotes the jumps size of L at time s. We remark that a β-stable Lévy process with $\beta \in (0,2)$ has Blumenthal-Getoor index β. Our main result is the following theorem.

Theorem: For the power variation $V(X,p)_n$ of a Lévy moving average process X the following results hold.

(i) When $\alpha \in (0, 1-1/p)$ and $p > \beta$, we obtain the stable convergence

$$n^{\alpha p} V(X,p)_n \xrightarrow{st} |f(0)|^p \sum_{m:T_m \in [0,1]} |\Delta L_{T_m}|^p \left(\sum_{k=0}^{\infty} |(k+U_m)_+^\alpha - (k-1+U_m)_+^\alpha|^p \right),$$

where $(T_m)_{m\geq 1}$ denote the jump times of L, $(U_m)_{m\geq 1}$ is a sequence of i.i.d. $U([0,1])$-distributed random variables and $x_+ := \max(0,x)$.

(ii) Assume that L is a symmetric β-stable Lévy process without drift. When $\alpha \in (0, 1-1/\beta)$ and $p < \beta$, we deduce the convergence in probability

$$n^{p(\alpha+1/\beta)-1} V(X,p)_n \xrightarrow{\mathbb{P}} \mathbb{E}[|\widetilde{L}_1|^p],$$

where $\widetilde{L}_1$ is a certain β-stable random variable.

(iii) When $\alpha > 1-1/p,\ p > \beta$ or $\alpha > 1-1/\beta,\ p < \beta$, we obtain the convergence in probability

$$n^{p-1} V(X,p)_n \xrightarrow{\mathbb{P}} \int_0^1 |F_s|^p ds \qquad \text{with} \qquad F_s := \int_{-\infty}^{s} g'(s-u) dL_u.$$

We conjecture that one can prove weak limit theorems associated with the ergodic case (ii). As in the Gaussian case, central and non-central limit theorems can be expected. These probabilistic results might be applied to estimate the parameters α and β.

The described limit theory can be extended in various directions. One of the most important class of models are the so called *ambit process*, which are defined as

$$Y_t = \int_{-\infty}^{t} g(t-s)\sigma_s dL_s,$$

where σ is a stochastic process that describes the intermittency of a turbulent flow. We think that our theoretical results are easily extended to this class of processes by the standard blocking technique. Another direction of our research lies in proving similar asymptotic results for a general class of statistics given via

$$\sum_{i=1}^{n} h\Big(X_{\frac{i}{n}} - X_{\frac{i-1}{n}} \Big),$$

where $h : \mathbb{R} \to \mathbb{R}$ is a measurable function. This setting corresponds to the so called Breuer-Major theorems in the Gaussian framework.

Detection of multiple structural breaks in multivariate time series

Philip Preuss
(joint work with Ruprecht Puchstein, Holger Dette)

We propose a new nonparametric procedure for the detection and estimation of multiple structural breaks in the autocovariance function of a multivariate (second-order) piecewise stationary process, which also identifies the components of the series where the breaks occur. The new method is based on a comparison of the estimated spectral distribution on different segments of the observed time series and consists of three steps: it starts with a consistent test, which allows to prove the existence of structural breaks at a controlled type I error. Secondly, it estimates sets containing possible break points and finally these sets are reduced to identify the relevant structural breaks and corresponding components which are responsible for the changes in the autocovariance structure. In contrast to all other methods which have been proposed in the literature, our approach does not make any parametric assumptions, is not especially designed for detecting one single change point and addresses the problem of multiple structural breaks in the autocovariance function directly with no use of the binary segmentation algorithm.

1. Testing for structural breaks

We assume to observe realizations of a centered $\mathbb{R}^d$ valued stochastic process $(\boldsymbol{X}_{t,T})_{t=1,...,T}$, where $\boldsymbol{X}_{t,T} = (X_{t,T,1}, ..., X_{t,T,d})^T$ has a piecewise stationary representation. This means that there exists an unknown number $K \in \mathbb{N}_0$ and points $0 = b_0 < b_1 < \cdots < b_k < b_{K+1} = 1$ such that

$$\boldsymbol{X}_{t,T} = \sum_{l=0}^{\infty} \boldsymbol{\Psi}_l(t/T)\boldsymbol{Z}_{t-l} \quad t = 1, ..., T, \tag{1}$$

where the functions $\boldsymbol{\Psi}_l : [0,1] \to \mathbb{R}^{d\times d}$, $l \in \mathbb{Z}$, are defined as $\boldsymbol{\Psi}_l(u) = \sum_{j=0}^{K} \boldsymbol{\Psi}_l^{(j)} 1_{S_j}(u)$ and 1_{S_j} denotes the indicator function of the set $S_j = \{u : b_j < u \leq b_{j+1}\}$, $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ denotes a centered Gaussian White Noise process with covariance matrix $\boldsymbol{I}_d$ and the matrices $\boldsymbol{\Psi}_l^{(j)} \in \mathbb{R}^{d\times d}$ correspond to the piecewise constant coefficents of the linear representations on the segment $(\lfloor b_j T \rfloor, \lfloor b_{j+1} T \rfloor]$. We assume that K is 'minimal' in the sense that for every pair $(i, i+1)$ with $i \in \{0, ..., K-1\}$ there exists an integer $l \in \mathbb{N}$ such that $\boldsymbol{\Psi}_l^{(i)} \neq \boldsymbol{\Psi}_l^{(i+1)}$. This ensures that, if K equals zero, there is no change point in the dependency structure, while structural breaks exist for $K \geq 1$. We introduce $\boldsymbol{f}_j(\lambda) = \frac{1}{2\pi} \sum_{l,m=0}^{\infty} \boldsymbol{\Psi}_l^{(j)} \big(\boldsymbol{\Psi}_m^{(j)}\big)^T \exp(-i\lambda(l-m))$ and obtain for the $\mathbb{C}^{d\times d}$ valued time-varying (piecewise constant) spectral density matrix

$$\boldsymbol{f}(u,\lambda) = \frac{1}{2\pi} \sum_{l,m=0}^{\infty} \boldsymbol{\Psi}_l(u)\big(\boldsymbol{\Psi}_m(u)\big)^T \exp(-i\lambda(l-m)) = \sum_{j=0}^{K} \boldsymbol{f}_j(\lambda) 1_{S_j}(u). \tag{2}$$

From this representation it follows that the spectral density has points of discontinuity in u direction at the break points b_i $(i = 1, ..., K)$ whenever $K \geq 1$. Therefore we propose to compare the spectral density $\lambda \mapsto \frac{1}{e} \int_{v-e}^{v} \boldsymbol{f}(u,\lambda) du$ with

$\lambda \mapsto \frac{1}{e}\int_v^{v+e} \boldsymbol{f}(u,\lambda)du$ for some 'small' constant e. If there exist structural breaks, the difference $\sup_{\omega\in[0,1]} \frac{1}{e}|\int_0^{\omega\pi}\int_v^{v+e} \boldsymbol{f}(u,\lambda)dud\lambda - \int_0^{\omega\pi}\int_{v-e}^{v} \boldsymbol{f}(u,\lambda)dud\lambda|$ will be positive for $v \in \{b_1,...,b_K\}$ while it vanishes for $v \in [0,1]\backslash\{b_1,...,b_K\}$ as $e \to 0$. In order to obtain a global measure for the presence of structural breaks we consider

$$D := \sup_{v,\omega\in[0,1]} ||\boldsymbol{D}(v,\omega)||_\infty \tag{3}$$

[$||\cdot||_\infty$ denotes the maximum norm of a matrix], where for $v \in [e, 1-e]$ and $\omega \in [0,1]$ the matrix $\boldsymbol{D}(v,\omega)$ is defined by

$$\boldsymbol{D}(v,\omega) := \frac{1}{e}\left(\int_0^{\omega\pi}\int_v^{v+e} \boldsymbol{f}(u,\lambda)dud\lambda - \int_0^{\omega\pi}\int_{v-e}^{v} \boldsymbol{f}(u,\lambda)dud\lambda\right) \in \mathbb{R}^{d\times d} \tag{4}$$

and we set $\boldsymbol{D}(v,\omega) = \boldsymbol{D}(e,\omega)$ and $\boldsymbol{D}(v,\omega) = \boldsymbol{D}(1-e,\omega)$ whenever $v \le e$ and $v \ge 1-e$ respectively. Under the hypothesis of no structural break, i.e. $K = 0$, we have $D = 0$, while D is strictly positive if structural breaks occur. In order to obtain a test for the null hypothesis

$$H_0 : \quad K = 0, \tag{5}$$

it is therefore natural to estimate D and to reject the null hypothesis for 'large' values of the estimator. An empirical version of D is obtained by choosing some even integer $N \le T/2$, taking the local periodogram

$$\boldsymbol{I}_N(u,\lambda) := \frac{1}{2\pi N}\sum_{r,s=0}^{N-1} \boldsymbol{X}_{\lfloor uT\rfloor - N/2+1+s,T}\boldsymbol{X}^T_{\lfloor uT\rfloor - N/2+1+r,T}\exp(-i\lambda(s-r)), \tag{6}$$

[where we set $\boldsymbol{X}_{j,T} = 0$ whenever $j \notin \{1,\dots,T\}$] as an estimator for $\boldsymbol{f}(u,\lambda)$ and considering

$$\hat{\boldsymbol{D}}_T(v,\omega) := \frac{1}{N}\sum_{k=1}^{\lfloor \omega N/2\rfloor}\Big(\boldsymbol{I}_N\big(v+N/(2T),\lambda_k\big) - \boldsymbol{I}_N\big(v-N/(2T),\lambda_k\big)\Big), \tag{7}$$

if $v \in [\frac{N}{T}, 1-\frac{N}{T}]$ where $\lambda_k = 2\pi k/N$ denote the Fourier frequencies. On the intervals $[0,\frac{N}{T})$ and $(1-\frac{N}{T},1]$ we define $\hat{\boldsymbol{D}}_T(v,\omega)$ as $\hat{\boldsymbol{D}}_T(\frac{N}{T},\omega)$ and $\hat{\boldsymbol{D}}_T(1-\frac{N}{T},\omega)$ respectively. So roughly speaking we construct an estimator of $\boldsymbol{D}(v,\omega)$ by replacing the integral by a Riemann sum, where the averaged time varying spectral density matrices $\frac{1}{e}\int_v^{v+e}\boldsymbol{f}(u,\lambda)du$ and $\frac{1}{e}\int_{v-e}^{v}\boldsymbol{f}(u,\lambda)du$ on the intervals $[v,v+e]$ and $[v-e,v]$ are replaced by the local periodograms $\boldsymbol{I}_N(v+N/(2T),\lambda)$ and $\boldsymbol{I}_N(v-N/(2T),\lambda)$. The final estimate of the quantity D in (3) is then defined by

$$\hat{D}_T := \sup_{(v,\omega)\in[0,1]^2} ||\hat{\boldsymbol{D}}_T(v,\omega)||_\infty = \max_{v\in[N/T,1-N/T]}\sup_{\omega\in[0,1]} ||\hat{\boldsymbol{D}}_T(v,\omega)||_\infty. \tag{8}$$

In [Preuß et al. (2013)] we derive crucial asymptotic results for $\hat{D}_T$ and prove that the corresponding quantiles under H_0 can be approximated sufficiently well by an extension of the AR(∞)-bootstrap of [Kreiß (1988)], resulting in a consistent level-α test for the null hypothesis of no structural breaks.

2. Detecting the number and location of break points

If structural breaks have been detected by the above described test it is of futher interest to estimate the number and location of possible break points and to identify the components responsible for these changes in the regime. In the following discussion we will briefly describe a procedure which consists of two steps and detects simultaneously the number, location and corresponding components of multiple structural breaks [cf. [Preuß et al. (2013)] for all details]. In the first step we estimate (shrinking if $N/T \to 0$) sets, which may contain potential break points. Roughly speaking these sets contain all points where the components of the spectral density estimate indicate a structural break. In a second step these sets are reduced to identify the relevant structural breaks and corresponding components which are responsible for these breaks. For this purpose we recall the definition (7), choose some constant $0 < \gamma < 1/2$ [a recommendation for this choice is given in [Preuß et al. (2013)]] and proceed as follows.

Step I (identification of sets containing break points) We consider a point $v \in \{\frac{N}{T}, \frac{N+1}{T}, ..., \frac{T-N}{T}\}$ as a candidate for a structural break in the component (a, b) if the inequality

$$N^{\gamma} \sup_{\omega \in [0,1]} |[\hat{\boldsymbol{D}}_T(v, \omega)]_{a,b}| > \boldsymbol{\varepsilon}_{T,a,b}(v) \tag{9}$$

holds, where $\boldsymbol{\varepsilon}_{T,a,b}(v)$ is a threshold satisfying $\liminf_{T\to\infty} \boldsymbol{\varepsilon}_{T,a,b}(v) \geq C > 0$ for some constant C and $\boldsymbol{\varepsilon}_{T,a,b}(v) = o(N^{\gamma})$ uniformly in $v \in [0, 1]$. A data driven rule for the choice of the threshold $\boldsymbol{\varepsilon}_{T,a,b}(v)$ with good finite sample properties is given in [Preuß et al. (2013)].

The decision rule (9) identifies subsets $R_1, ..., R_{K_T} \subset \{N/T, ..., 1-N/T\}$ where possible break points in the components of the spectral density matrix may occur. The goal of the second step is to reduce this set significantly in order to end up with the final estimators of the break points. This step (roughly speaking) works as follows: For every set R_j of points in $\{N/T, ..., 1 - N/T\}$ satisfying (9) for at least one pair $(a, b) \in \{1, ..., d\}^2$ we identify a point $\tilde{b} \in R_j$ for which the local deviation from stationarity is maximal and then remove all points of the interval $[\tilde{b} - \frac{N}{T}, \tilde{b} + \frac{N}{T}]$ from the set R_j. We then say that there is a structural break at $\tilde{b}$ for all components which are above the corresponding threshold at this point.

References

[Kreiß (1988)] Kreiß, J.-P. (1988). Asymptotic statistical inference for a class of stochastic processes. Habilitationsschrift, Fachbereich Mathematik, Universität Hamburg.

[Preuß et al. (2013)] Preuß, P. and Puchstein, R. and Dette, H. (2013). Detection of multiple structural breaks in multivariate time series. arXiv:1309.1309.

Adaptive online forecasting of a locally stationary time varying autoregressive process

François Roueff
(joint work with Andrés Sánchez-Pérez, Christophe Giraud)

This report gathers some results of the ongoing Phd thesis of Andrés Sánchez-Pérez that I have been supervising at the LTCI (Institut Mines-Télécom; CNRS; Télécom ParisTech) with Christophe Giraud from Université Paris-Sud. In this work, we study the problem of online adaptive forecasting for locally stationary Time Varying Autoregressive processes (TVAR). The Normalized Mean Least Squares algorithm (NMLS) is an online stochastic gradient method which has been shown to perform efficiently, provided that the gradient step size is well chosen. This choice highly depends on the smoothness exponent of the evolving parameters. In this contribution, we show that a sequential aggregation of several NLMS estimators at various gradient step sizes is able to adapt to an unknown smoothness, resulting in an online adaptive predictor.

1. Introduction

In many applications where high frequency data is collected, one wishes to predict the next values of an observed time series through an online predictor learning algorithm, allowing one to process a large amount of data. However, as a counterpart, the usual stationarity assumption has to be weakened to take into account some smooth evolution of the environment. An interesting approach to cope with this non-stationarity issue is to rely on a *local stationarity* assumption. We refer to [2] and the references therein for a recent general view about statistical inference for locally stationary processes. Here we focus on a particular model, which is obtain by apply this approach to a time-varying autoregressive process.

Definition 1 (Time-varying autoregressive process (TVAR)) *The T-sample $X_{1,T}, \ldots, X_{T,T}$ of a TVAR process or order d satisfies*

$$X_{t,T} = \sum_{j=1}^{d} \theta_j \left(\frac{t-1}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t \,, \tag{1}$$

where the ξ_t are i.i.d. with $\mathbb{E}\xi_t = 0$ and θ_j are the time-varying autoregressive coefficients rescaled on the interval $[0,1]$.

Initial conditions should be added but we omit the details here for brevity. Nevertheless it is important to mention that they imply that ξ_t is independent of the past of $X_{s,T}$ up to $s = t-1$. Hence the best predictor of $X_{t,T}$ given its past is $\boldsymbol{\theta}'\left(\frac{t-1}{T}\right)\underline{X}_{t-1,T}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ and $\underline{X}_{t-1,T} = (X_{t-1,T}, \ldots, X_{t-d,T})'$. Here and in the following, we let A' denote the transpose of the matrix A.

The local autoregressive polynomial of the TVAR at rescaled time $u \in [0,1]$ is defined as $\boldsymbol{\theta}(z;u) = 1 - \sum_{j=1}^{d} \theta_j(u) z^j$.

For $\zeta > 0$ we define by $s_d(\zeta) = \left\{ \boldsymbol{\theta} : [0,1] \to \mathbb{R}^d, \boldsymbol{\theta}(z;u) \neq 0, \forall |z| < \zeta^{-1}, u \in [0,1] \right\}$.

Following [3], a TVAR process is *locally stationary* if $\boldsymbol{\theta}$ and σ satisfy some smoothness conditions and $\boldsymbol{\theta} \in s_d(\delta)$ for some $\delta \in (0,1)$. These conditions with adequate initial conditions imply that there exists a solution with representation

$$X_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j) \xi_{t-j} \ , \tag{2}$$

and that there exist $\bar{K} > 0$ and $\rho \in [0,1)$ such that $\sup_{t,T} |a_{t,T}(j)| \leq \bar{K} \rho^j$.

In this contribution, we shall use $\beta-$ Lipschitz smoothness conditions. For any $\beta \in (0,1]$, the $\beta-$ Lipschitz semi-norm of a function $f : [0,1] \to \mathbb{R}^d$ is defined as $|f|_{\Lambda,\beta} = \sup_{s_1 \neq s_2} \frac{f(s_1) - f(s_2)}{|s_1 - s_2|^\beta}$. For $L \in \mathbb{R}_+^*$ and $\beta > 0$, let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be such that $\beta = k + \alpha$. The $\beta-$ Lipschitz ball of radius L is denoted by $\Lambda_d(\beta, L)$.

2. NLMS estimators

In [4], the normalized least mean squares algorithm (NLMS) estimator of the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ is studied for locally stationary TVAR processes. We will use basically the same estimators but with a slight modification (Eq. (4) below). For a given gradient step size $\mu > 0$, our modified NLMS estimator is defined recursively by

$$\tilde{\boldsymbol{\theta}}_{t,T}(\mu) = \tilde{\boldsymbol{\theta}}_{t-1,T}(\mu) + \mu \left(X_{t,T} - \tilde{\boldsymbol{\theta}}'_{t-1,T}(\mu) \underline{X}_{t-1,T} \right) \frac{\underline{X}_{t-1,T}}{1 + \mu \left| \underline{X}_{t-1,T} \right|_2^2} \ , \tag{3}$$

$$\hat{\boldsymbol{\theta}}_{t,T}(\mu) = \begin{cases} \tilde{\boldsymbol{\theta}}_{t,T}(\mu) & \text{if } \left| \tilde{\boldsymbol{\theta}}_{t,T}(\mu) \right|_2 \leq 2^d - 1 \ , \\ \frac{2^d - 1}{\left| \tilde{\boldsymbol{\theta}}_{t,T}(\mu) \right|_2} \tilde{\boldsymbol{\theta}}_{t,T}(\mu) & \text{otherwise.} \end{cases} \tag{4}$$

Here $|\cdot|_2$ stands for the Euclidean norm. The additional step (4) is a projection on the ball of radius $2^d - 1$ which guaranties our estimators to be bounded. The statistic $\hat{\boldsymbol{\theta}}_{t-1,T}(\mu)$ is our estimation for $\boldsymbol{\theta}\left(\frac{t-1}{T}\right)$, from which we obtain the predictor $\hat{\boldsymbol{\theta}}'_{t-1,T}(\mu) \underline{X}_{t-1,T}$ of $X_{t,T}$. The precision of the prediction is evaluated by the cumulative loss

$$\sum_{t=1}^{T} \ell(\hat{\boldsymbol{\theta}}'_{t-1,T}(\mu) \underline{X}_{t-1,T}, X_{t,T}) \ ,$$

for the loss function $\ell(x,y) = |x - y|^q$ for some $q = 1, 2, 3, \ldots$ We shall focus on the case $q = 2$ in the following.

3. Aggregation of predictors

From a collection $\left\{\hat{\boldsymbol{\theta}}^{(j)}_{t-1,T}\right\}_{1\le j\le N}$ of N estimators of $\boldsymbol{\theta}$ we obtain predictors of $X_{t,T}$ defined by $f_{j,t}=\hat{\boldsymbol{\theta}}^{(j)\prime}_{t-1,T}\underline{X}_{t-1,T}$, $1\le j\le N$. In particular, each index j may correspond to a NLMS estimator obtained with a given μ_j. In aggregation language, the $f_{j,t}$s are called expert's predictions or forecasts. The strategy used in a different context (bounded observations) by [1] suggest to combine all possible expert's predictions as follows

$$\check{\boldsymbol{\theta}}_{t-1,T}=\sum_{j=1}^{N}\bar{\alpha}_{j,t}\hat{\boldsymbol{\theta}}^{(j)}_{t-1,T}\quad\text{with}\quad\bar{\alpha}_{j,t}=\frac{\exp\left(-\eta\sum_{s=1}^{t-1}\tilde{\ell}_{j,s}\right)}{\sum_{i=1}^{N}\exp\left(-\eta\sum_{s=1}^{t-1}\tilde{\ell}_{i,s}\right)},$$

where $\tilde{\ell}_{j,t}=\nabla_x\ell\left(\sum_{i=1}^{N}\alpha_{i,t}f_{i,t},X_{t,T}\right)\cdot f_{j,t}$ and with the convention that a sum over no element is null, i.e. $\bar{\alpha}_{j,1}=\frac{1}{N}$ for all j. The symbol $\nabla_x\ell$ denotes the subgradient of ℓ taken with respect to the first coordinate. The parameter $\eta>0$ will be specified below. Based on sequential aggregation techniques (see [1]), and [4, Theorem 2], we obtain the following result.

Theorem 1 *Suppose that $\mathbb{E}[|\xi_0|^r]<\infty$ for some $r>8$. For $j=1,\dots,N$ let $\hat{\theta}^{(j)}_{t,T}=\hat{\boldsymbol{\theta}}_{t,T}\left(T^{-2j/(2j+N)}\right)$. Define the aggregated estimator $\check{\boldsymbol{\theta}}_{t,T}$ as above with $\eta\propto(\log N/T)^{1/2}$. Suppose moreover that $\boldsymbol{\theta}\in s_d(\delta)\cap\Lambda_d(\beta,L)$ for $L>0$ and $\beta\in(0,1]$, and that σ is bounded between two positive values. Then we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\left|\check{\boldsymbol{\theta}}'_{t-1,T}\underline{X}_{t-1,T}-X_{t,T}\right|^2-\sigma^2\left(\frac{t}{T}\right)\,\mathbb{E}\left[|\xi_0|^2\right]\right)\right]$$
$$=O\left(T^{1/(1+2\beta)}+\sqrt{T\log N}\right),$$

provided that $\frac{\log(T)}{N}=O(1)$ as $T\to\infty$.

Observe that, for $\beta<1/2$, the convergence rate is the same as the one in [4, Theorem 2] when μ is optimized according to the value of β. Here, in contrast, the estimator does not require the knowledge of β to achieve this rate.

Acknowledgements

This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012 - 2015 and by the Labex LMH (ANR-11-IDEX-003-02).

References

[1] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

[2] Rainer Dahlhaus. Local inference for locally stationary time series based on the empirical spectral measure. *J. Econometrics*, 151(2):101–112, 2009.

[3] Rainer Dahlhaus and Wolfgang Polonik. Empirical spectral processes for locally stationary time series. *Bernoulli*, 15(1):1–39, 2009.

[4] Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.

Structured wavelet estimation of time-varying spectra

Rainer von Sachs

(joint work with Jean-Marc Freyermuth)

In this work we revisit nonparametric estimation of the time-varying spectral density $f(t/T, \omega)$ of a (zero mean) locally stationary time series $\{X_t\}_{t=1,\dots,T}$. Hereby we want to avoid using any (prior) segmentation of the data over time but work under the paradigm of the (non-linear) time-frequency smoothing method of [1]. This approach suggests to replace classical linear smoothing of the (empirical) Wigner-Ville (WV) spectrum - or preperiodogram - by some more adaptive wavelet denoising method which is akin non-linear (hard) thresholding of data observed on an equidistant grid of a two-dimensional curve plus noise (see equation (2) below). As such, it avoids the usually non-adaptive (though asymptotically consistent) time-frequency smoothing of segmented periodograms of equal segment length. In this alternative approach, the challenge lies in the finite sample behaviour of the preperiodogram which is basically the Fourier transform of a very local autocovariance estimation:

$$I_{t,T}(\omega) = \sum_{k:1\le t-k/2,t+k/2\le T} X_{[t-k/2],T}\, X_{[t+k/2],T}\ \exp(-2\pi i k\omega) \tag{1}$$

As such it is a highly around zero oscillating object, and unlike the classical periodogram it is not non-negative and suffers from inherent cross-interference terms: the preperiodogram is a bilinear function of the time series data and hence the quadratic superposition principle applies. In particular time data which come from non-overlapping regions of the time domain find themselves mingled in the non-vanishing cross-terms, hence causing a potentially important bias due to non-stationarity of the data. There is indeed a vast engineering literature which tries to best reduce this interference terms by investigating the behaviour of *linear* kernel smoothers. However, our approach builds on *non-linearly thresholding* the 2-dimensional projection coefficients of the preperiodogram onto a separable (or hyperbolic) wavelet basis $\{\psi_{j_1k_1}(t/T)\,\psi_{j_2k_2}(\omega)\}$. Here, we recall the role of increasing multiresolution scales $j_1 \ge 0$ in time and $j_2 \ge 0$ in frequency, providing, in a separable way, finer and finer approximations of the local time-frequency structure.

In this work we address, both from a theoretical and a practical point of view, a series of questions that arise throughout this research agenda: i) as an alternative to the plug-in thresholds suggested in [1], we investigate how to optimally calibrate the level-and-location dependent thresholds in our very noisy and highly heteroscedastic curve estimation context; ii) we study which combination of wavelet functions forms the hyperbolic basis to best reduce the inherent cross-terms in the preperiodogram; iii) we investigate "Tree-structured" wavelet estimation ([3]) where we compare fully non-linear hard-thresholding (as in [1]) by "hard-tree" thresholding under a hereditary constraint: a "children" coefficient in the empirical wavelet tree can only survive our keep-or-kill rule if all its ancestors remain activated; and finally iv) we study how the well-known "Heisenberg constraint" translates into another structural constraint of our thresholding scheme: given T^2 preperiodogram data $(I_{t,T}(\omega_n))_{t,n=1,\ldots,T}$ the effective support of any (smoothing) time-frequency window needs to have as minimal area $2\pi T$ (or $2\pi/T$, respectively, in rescaled time of equation (2) below). This leads us to work in a *constrained* 2-dimensional curve estimation (denoising) problem

$$(2) \quad I_{t,T}(\omega_n) \approx f(t/T,\omega_n) + \sqrt{f(t/T,\omega_n)}\,\varepsilon_{t,n}\ , \quad t=1,\ldots,T,\ n=1,\ldots,T,$$

where we reduce the redundancy in this "oversampled" problem of T^2 preperiodograms, resulting from only T time series values $X_1,\ldots,X_T$, by the following structural constraint on *tree-structured hard thresholding* of the coefficients

$$(3) \qquad \widehat{\theta}_I = < I_{t,T}(\omega_n), \psi_{j_1k_1}(t/T)\ \psi_{j_2k_2}(\omega_n) > ,$$

given by

$$(4) \qquad 2^{j_1+j_2} < 2^J = T \qquad \text{(Heisenberg constraint)}.$$

This natural constraint balances time and frequency resolution being reciprocal to each other as the temporal and frequential support of the 2-d wavelets have lengths proportional to 2^{-j_1} and to 2^{-j_2}, respectively. The resulting threshold estimator is supposed to provide the right time-frequency "support", i.e. to adapt to the time-frequency structure of the true underlying spectrum, provided the thresholds are correctly determined to suppress all those "time-frequency windows" which are not significant (i.e. not matching the local time-frequency spectral structure).

Tree-structured wavelet (hard) thresholding ([3], [4]) is motivated from two observations for denoising curves: a) the use of the Haar basis in time for $\{\psi_{j_1k_1}(t/T)\}$ allows non-dyadic breaks in time to be better resolved than with classical hard thresholding, which is ideal for our goal of achieving implicitly an optimal time-segmentation; b) in a general multivariate context [4] showed that under certain *structural conditions* the *maxiset* of Hard tree thresholding is larger than those of Hard thresholding. Hereby, we understand by *maxiset* the largest set of functions that can be estimated with a given rate of convergence, usually chosen to be the (near-) optimal nonparametric rate of convergence of the associated L_2-risk over standard function spaces (here anisotropic Besov classes).

Tansfering this maxiset result to spectral estimation under the model and conditions of [1], we show that the control of the variance of the Hard tree estimator is achieved via the Heisenberg constraint of (4), in that all coefficients on scales not respecting this constraint need to be eliminated from the reconstruction.

For calibrating thresholding of the remaining empirical coefficients we propose to borrow strength from the paradigm behind what [2] call "wavelet-Fisz" noise free thresholds: here the idea is to calculate "Fisz-ratios" by studentizing each empirical coefficient $\widehat{\theta}_I$ by an estimator of its standard deviation, in order to avoid plug-in estimation of the asymptotic variance of $\widehat{\theta}_I$ (as proposed by [1]). One possibility for the denominator of this Fisz-ratio is to project the preperiodogram onto a time-frequency "mod-wavelet", essentially an L_1-normalized basis $\{|\psi_{j_1k_1}(t/T)|\ |\psi_{j_2k_2}(\omega)|\}$. Subsequently, following the paradigm of wavelet thresholding to suppress (asymptotically) all noise in the coefficient domain, we perform Monte Carlo simulations to find appropriate quantiles in the tails of the Fisz-ratio, under the null of a constant spectrum in time and frequency.

And finally, to address the problem of reducing the interference terms in preperiodogram smoothing, in work in progress we study which type of wavelets $\{\psi_{j_2k_2}(\omega)\}$ in frequency provide for the best finite-sample performance. Using as a benchmark an ideal piecewise stationary approximation to the underlying locally stationary time series, we investigate the amount of cross-interference contribution supposedly minimal for frequency wavelets with the shortest possible support in the Fourier domain, i.e. the domain of the autocovariance function of the time series. Natural candidates are wavelets with a compact support in the Fourier domain, such as Shannon wavelets and members of the family of Meyer wavelets (the latter one having better decay properties in the physical, i.e. frequency domain). Under the Heisenberg constraint (4), these frequency wavelets provide, locally for a given time point within the support of the time domain wavelet of length $N_{j_1} = T/2^{j_1}$, some kind of lag-window smoothing of the (local) autocovariances by using as few observations as possible coming from outside the given time segment.

We conclude by repeating the essential steps of our proposed *constrained* Hard tree estimator, combining the ingredients of our research agenda i)-iv):

- Project the preperiodogram onto a 2-dimensional hyperbolic wavelet basis, suited for estimation of spectra of anisotropic smoothness over time (being rather "unsmooth") and frequency (being typically more regular).
- Impose the Heisenberg constraint: $2^{j_1+j_2} < 2^J = T$. This reduces (statistical) redundancy in the preperiodogram and is essential to regularise the bivariate thresholding estimator.
- Apply *tree-structured hard thresholding* on $\widehat{\theta}_I$ from (3), using a hereditary structure for hard thresholding all wavelet coefficients on all constrained scales, and using correctly calibrated thresholds, based on "Fisz-ratios".

The resulting estimator is supposed to provide the right time-frequency "support", i.e. to adapt to the time-frequency structure of the true spectrum. Under some

employed structural constraints, it achieves equally well the optimal rates of convergence over anisotropic smoothness classes as the estimator of [1]. Moreover, its maxisets can be shown to be larger than those of classical hard thresholding.

References

[1] M. Neumann and R. von Sachs. *Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra.* Ann. Statist., **25** (1997), 38-76.
[2] P. Fryzlewicz, G. Nason and R. von Sachs. *A wavelet-Fisz approach to spectrum estimation.* J. Time Series Analysis **29** (2008), 868-880,
[3] F. Autin, J.-M. Freyermuth and R. von Sachs. *Ideal Denoising within a Family of Tree-Structured Wavelet Estimators.* Electron. J. Statist., **5** (2011), 829-855.
[4] F. Autin, G. Claeskens and J.-M. Freyermuth. *Hyperbolic wavelet thresholding methods and the curse of dimensionality through the maxiset approach.* Appl. Comp. Harmonic Analysis (2013), electronically published (9 May 2013).

Feature Matching in Time Series Modelling

Howell Tong
(joint work with Yingcun Xia)

This talk is based on joint work with Professor Yingcun Xia of the National University of Singapore.

Using a time series model to mimic an observed time series has a long history. However, with regard to this objective, conventional estimation methods for discrete-time dynamical models are frequently found to be wanting. In fact, they are characteristically misguided in at least two respects: (i) assuming that there is a true model; (ii) evaluating the efficacy of the estimation as if the postulated model is true. There are numerous examples of models, when fitted by conventional methods, that fail to capture some of the most basic global features of the data, such as cycles with good matching periods, singularities of spectral density functions (especially at the origin) and others. We argue that the shortcomings need not always be due to the model formulation but the inadequacy of the conventional fitting methods. After all, all models are wrong, but some are useful if they are fitted properly. The practical issue becomes one of how to best fit the model to data. Thus, in the absence of a true model, we prefer an alternative approach to conventional model fitting that typically involves one-step-ahead prediction errors. Our primary aim is to match the joint probability distribution of the observable time series, including long-term features of the dynamics that underpin the data, such as cycles, long memory and others, rather than short term prediction. For want of a better name, we call this specific aim feature matching. The challenges of model misspecification, measurement errors and the scarcity of data are forever present in real time series modelling. In this paper, by synthesizing earlier attempts into an extended-likelihood, we develop a systematic approach to empirical time series analysis to address these challenges and to aim at achieving better feature matching. Numerical results, based on both simulations and real data,

suggest that the proposed catch-all approach has several advantages over the conventional methods, especially when the time series is short or with strong cyclical fluctuations. We conclude with listing directions that require further development.

We shall use letters y and x to signify respectively the real time series under study and the time series generated by the postulated model. The adjective observable is reserved for a stochastic process. An observed time series consisting of observations constitutes (possibly part of) a realization of a stochastic process. Consider the postulated model

$$x(t) = g(x(t-1), ..., x(t-p); \theta) + e(t),$$

where $e(t)$ is the innovation and the function $g(\cdot; \theta)$ is known up to parameters θ. To indicate the dependence of $x(t)$ on θ, we also write it as $x(t; \theta)$. In order for the model to be able to approximate an observable $\{y(t) : t = 1, 2, ...\}$ well, it is natural to require that the state space of $\{x(t; \theta) : t = 1, 2, ...\}$ covers that of the observable $\{y(t) : t = 1, 2, ...\}$. For simplicity of exposition, let $p = 1$. Starting from $x(0; \theta) = y(0)$, the postulated model is said to match an observable time series under study perfectly if their conditional distributions are the same, namely,

$$P\{x(1; \theta_0) < u(1), ..., x(n; \theta_0) < u(n) | x(0; \theta_0) = y(0)\} \\ = P\{y(1) < u(1), ..., y(n) < u(n) | y(0)\}$$

almost surely for some θ_0 and any n and any real values $u(1), ..., u(n)$. We call the approach based on the above model, including all its weaker versions, some of which will be described in the full paper collectively by the name catch-all approach.

In our paper, we adhere to Boxs dictum and abandon, right from the very beginning, the assumption of either the postulated parametric model being true or the observations being error-free. Instead, we focus on ways to improve the feature matching of a postulated parametric model to the observable time series. We have introduced the notion of an optimal parameter in the absence of a true model and defined a new form of consistency. In particular, we have synthesized earlier attempts into a systematic approach of estimation of the optimal parameter, by reference to up-to-m-step-ahead predictions of the postulated model. We have also developed some general results with proofs. Conventional methods of estimation are typically based on just the one-step-ahead prediction. Our analysis, simulation study and real applications have convinced us that they are often found wanting in many situations, for example, the absence of a true model, short data sets, observation errors, highly cyclical data and others. Our stated primary objective is feature matching. Prediction is secondary here. However, we have evidence to suggest that a model with good feature matching can stand a better chance of enjoying good medium- to long-term prediction. Of course, if the aim is prediction at a specified horizon, then our approach has almost nothing to offer.

Discriminating between long-range dependence and non stationarity

Mathias Vetter
(joint work with Philip Preuß)

In this talk we give a partial answer to the problem of discrimination between a stationary long-range dependent model and a non stationary process.

In (econometric) practice, there are various models accounting for deviations from the expected behaviour of the autocorrelation function if the underlying time series was a stationary short memory one. In earlier days, the lack of an exponential decay of the ACF forced authors to advocate the use of long memory models as in [2] and [3] instead, while nowadays people point in the direction of using non-stationary models; see [1], [4] or [7]. Our task therefore is to give a hint on how to decide whether the observed process comes from one or the other model. In particular, we propose a nonparametric test for stationarity within a general non stationary long memory framework.

The proposed procedure in the talk is based on [6]: Suppose the underlying process is a locally stationary one, i.e. it is of the form

$$X_{t,T} = \sum_{l=0}^{\infty} \psi_{t,T,l} Z_{t-l}, \quad t = 1, \dots, T,$$

for certain coefficients $\psi_{t,T,l}$ and standard normal variables Z_t. We assume that these coefficients can (in an appropriate sense) be approximated by smooth functions $\psi_l(t/T)$ which satisfy

$$\psi_l(u) = a(u) I(l)^{d(u)-1} + O(I(l)^{D-2})$$

for twice differentiable functions $a, d : [0,1] \to \mathbb{R}_+$ and where $D := \sup_u |d(u)| < 1/2$ is the maximal long range dependence parameter. The time varying spectral density is given by

$$f(u,\lambda) := \frac{1}{2\pi} \Big| \sum_{l=0}^{\infty} \psi_l(u) \exp(-i\lambda l) \Big|^2.$$

In case $X_{t,T}$ is actually stationary, the spectral density does not depend on the time u. Therefore, our test is based on the intuition that

$$E(v,\omega) := \frac{1}{2\pi} \Big(\int_0^v \int_0^{\pi\omega} f(u,\lambda) d\lambda du - v \int_0^{\pi\omega} \int_0^1 f(u,\lambda) du d\lambda \Big), \quad (v,\omega) \in [0,1]^2,$$

is equal to zero for all choices of v and ω for stationary processes, while it deviates from zero otherwise. Therefore,

$$E := \sup_{(v,\omega)\in[0,1]^2} |E(v,\omega)|$$

serves as a Kolmogorov-Smirnov type distance between the time varying spectral density and its best approximation through a stationary spectral density. We use

estimates for the integrated spectral density via Riemann sums of periodograms. Precisely, if

$$I_N(u,\lambda) := \frac{1}{2\pi N}\Big|\sum_{s=0}^{N-1} X_{\lfloor uT\rfloor - N/2+1+s,T}\exp(-i\lambda s)\Big|^2$$

denotes the periodogram computed from N observations around time uT, then the empirical version of $E(v,\omega)$ is given by

$$\hat{E}_T(v,\omega) := \frac{1}{T}\sum_{j=1}^{\lfloor vM\rfloor}\sum_{k=1}^{\lfloor \omega\frac{N}{2}\rfloor} I_N(u_j,\lambda_k) - \frac{\lfloor vM\rfloor}{M}\frac{1}{T}\sum_{j=1}^{M}\sum_{k=1}^{\lfloor \omega\frac{N}{2}\rfloor} I_N(u_j,\lambda_k),$$

where $u_j = (N(j-1)+N/2)/T$, $\lambda_k := 2\pi k/N$ are the regular Fourier frequencies and $M = T/N$.

Setting

$$\hat{G}_T(v,\omega) = \hat{E}_T(v,\omega) - E\left(\frac{\lfloor vM\rfloor}{M}, \frac{\lfloor \omega N/2\rfloor}{N/2}\right),$$

we have weak convergence in a process sense under the null hypothesis, that is

$$\sqrt{T}(\hat{G}_T(v,\omega))_{(v,\omega)\in[0,1]^2} \Rightarrow (G(v,\omega))_{(v,\omega)\in[0,1]^2},$$

where the limiting Gaussian process is the same one as in the short memory case [see [5]], at least if the (possibly time varying) long memory parameter is smaller than $1/4$. As a consequence, $\sqrt{T}\sup_{(v,\omega)\in[0,1]^2}|\hat{E}_T(v,\omega)|$ converges in distribution to $\sup_{(v,\omega)\in[0,1]^2}|G(v,\omega)|$, if the underlying process is indeed stationary. Consistency under the alternative is provided as well. If the long memory parameter D exceeds the boundary $1/4$, the limiting process is non-Gaussian and different from $(G(v,\omega))_{(v,\omega)\in[0,1]^2}$ above.

In any case, it is difficult to assess the disctribution of $\sup_{(v,\omega)\in[0,1]^2}|G(v,\omega)|$ even if $D < 1/4$. For this reason, the novel FARI(∞) bootstrap is introduced which provides a bootstrap-based test for stationarity which shows good empirical properties if the long memory parameter is smaller than $1/2$ which is the usual restriction in the framework of long-range dependent time series. We investigate the finite sample properties of our approach in a comprehensive simulation study and employ the new test in an analysis of two data sets.

References

[1] Fryzlewicz, P., Sapatinas, T., Subba Rao, S. *A Haar-Fisz technique for locally stationary volatility estimation*, Biometrika **93** (2006), 687–704.

[2] Granger, C., Joyeux, R. *An introduction to long-memory time series models and fractional differencing*, Journal of Time Series Analysis **1** (1980), 15–29.

[3] Hosking, J. *Fractional differencing*, Biometrika **68** (1981), 165–176.

[4] Mikosch, T., Starica, C. *Non-stationarities in financial time series, the long range dependence and the IGARCH effects*, The Review of Economics and Statistics **86** (2004), 378–390.

[5] Preuß, P., Vetter, M., Dette, H. *A test for stationarity based on empirical processes*, Bernoulli (2013+), to appear.

[6] Preuß, P., Vetter, M. *Discriminating between long-range dependence and non stationarity*, Electronic Journal of Statistics **7** (2013), 2241–2297.

[7] Starica, C., Granger, C. *Nonstationarities in stock returns*, The Review of Economics and Statistics **87** (2005), 503–522.

Detecting Smooth Changes in Locally Stationary Processes

Michael Vogt
(joint work with Holger Dette)

In many applications, the stochastic properties of the observed time series change over time. It is often realistic to assume that the properties are approximately the same over short time periods and then gradually start to vary. This behaviour is well modelled by locally stationary processes introduced in Dahlhaus [1]. Roughly speaking, we call a process $\{X_{t,T}\}$ locally stationary if it can be approximated by a stationary process $\{X_t(u)\}$ locally around each rescaled time point $u \in [0,1]$. For a rigorous definition, see Vogt & Dette [5].

In what follows, we investigate the question how to estimate time spans where the stochastic properties of a locally stationary time series $\{X_{t,T}\}$ are (approximately) the same. More specifically, let $\lambda_{t,T}$ be a stochastic feature of $X_{t,T}$ such as the mean $\mathbb{E}[X_{t,T}]$, the variance $\mathrm{Var}(X_{t,T})$ or the distribution function $F_{t,T}(\cdot) = \mathbb{P}(X_{t,T} \leq \cdot\,)$. Suppose we are interested in the behaviour of $\lambda_{t,T}$ around the time point t^*, or equivalently, around the rescaled time point $u^* = t^*/T$. Moreover, let $u_0 < u^* < u_1$ and assume that $\lambda_{t,T}$ does not vary within the interval $[u_0, u_1]$ but gradually starts to vary outside it. Our goal is to estimate the time span $[u_0, u_1]$.

To handle this estimation problem, we slightly reformulate it: Let λ_u be the stochastic feature of the approximating process $\{X_t(u)\}$ which corresponds to $\lambda_{t,T}$. Under mild conditions, time-variation in $\lambda_{t,T}$ is asymptotically equivalent to time-variation in λ_u. Our estimation problem can thus be formulated in terms of λ_u rather than $\lambda_{t,T}$: Suppose that λ_u does not vary within the rescaled time interval $[u_0, u_1]$ but smoothly varies outside it. Our aim is to estimate the time points u_0 and u_1 where λ_u starts to change over time.

To keep the exposition as simple as possible, we restrict attention to the case where the time point of interest u^* is equal to 1, i.e., $[u_0, u_1] = [u_0, 1]$. This case occurs quite frequently in applications. When performing forecasts, for example, we would ideally like to know the interval $[u_0, 1]$ where the stochastic properties of the data are stable over time.

To estimate the time point u_0, we proceed as follows:

Step 1: We first set up a function $\mathcal{D} : [0,1] \to \mathbb{R}_{\geq 0}$ which measures the amount of time-variation in the feature of interest λ_w. It is constructed such that

$$\mathcal{D}(u) \begin{cases} = 0 & \text{if } \lambda_w \text{ does not vary on } [u,1] \\ > 0 & \text{if } \lambda_w \text{ varies on } [u,1] \end{cases} \tag{1}$$

and is called a measure of time-variation. (1) immediately implies that $\mathcal{D}(u) = 0$ for $u \geq u_0$ and $\mathcal{D}(u) > 0$ for $u < u_0$. Hence, u_0 is characterized as the time point

where the measure of time-variation $\mathcal{D}$ starts to deviate from zero. Importantly, $\mathcal{D}$ does not have a jump at u_0 in general, but smoothly deviates from zero at this point. In particular, its degree of smoothness depends on how smoothly λ_w varies over time.

Before we describe how to construct $\mathcal{D}$, we specify the class of stochastic features we work with. We allow for any feature λ_w which has the following property:

(P) λ_w is uniquely determined by the set of moments $\{\mathbb{E}[f(X_t(w))] : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a family of real-valued measurable functions f.

(P) is a fairly weak property satisfied by a wide range of features. If $\lambda_w = \mathbb{E}[X_t(w)]$ for instance, then simply $\mathcal{F} = \{\mathrm{id}\}$. As another example, suppose that λ_w is the distribution function of $X_t(w)$, i.e., $\lambda_w = F(w, \cdot) = \mathbb{E}[1(X_t(w) \leq \cdot)]$. Here, $\mathcal{F}$ is the class of indicator functions $\{1(\,\cdot \leq x) : x \in \mathbb{R}\}$, where for vectors the inequality sign is understood componentwise.

Our construction of $\mathcal{D}$ is based on the following idea: By (P), the feature λ_w is fully characterized by the values $\mathbb{E}[f(X_t(w))]$ with f running over all functions in the family $\mathcal{F}$. This implies that time-variation in λ_w is equivalent to time-variation in the moments $\mathbb{E}[f(X_t(w))]$ for some $f \in \mathcal{F}$. To detect changes in λ_w over time, we may thus set up a function which captures time-variations in the quantities $\mathbb{E}[f(X_t(w))]$ for any $f \in \mathcal{F}$. This idea underlies the following definition:

$$\mathcal{D}(u) = \sup_{f \in \mathcal{F}, v \in [u,1]} \big| D(u, v, f) \big|,$$

where

$$D(u, v, f) = \int_v^1 \mathbb{E}[f(X_t(w))] dw - \Big(\frac{1-v}{1-u}\Big) \int_u^1 \mathbb{E}[f(X_t(w))] dw.$$

Since $\mathcal{D}$ depends on the unobserved feature λ_w, we replace it by an estimator $\hat{\mathcal{D}}_T$. In particular, we set

$$\hat{\mathcal{D}}_T(u) = \sup_{f \in \mathcal{F}, v \in [u,1]} \big| \hat{D}_T(u, v, f) \big|$$

along with

$$\hat{D}_T(u, v, f) = \frac{1}{T} \sum_{t=\lceil vT+1 \rceil}^{T} f(X_{t,T}) - \Big(\frac{1-v}{1-u}\Big) \frac{1}{T} \sum_{t=\lceil uT+1 \rceil}^{T} f(X_{t,T}).$$

Step 2: Our estimator of u_0 is based on the observation that

$$\sqrt{T}\mathcal{D}(u) \begin{cases} = 0 & \text{for } u \geq u_0 \\ \to \infty & \text{for } u < u_0 \end{cases}$$

as $T \to \infty$. As the statistic $\hat{\mathcal{D}}_T$ estimates the measure $\mathcal{D}$, its scaled version $\sqrt{T}\hat{\mathcal{D}}_T$ should exhibit a similar behaviour. Indeed, one can show that

$$\sqrt{T}\hat{\mathcal{D}}_T(u) \begin{cases} = O_p(1) & \text{for } u \geq u_0 \\ \xrightarrow{p} \infty & \text{for } u < u_0. \end{cases}$$

We now exploit this dichotomous behaviour of $\sqrt{T}\hat{\mathcal{D}}_T$.

To do so, we first transform the statistic $\sqrt{T}\hat{\mathcal{D}}_T$ to behave approximately like a function that has a jump at u_0. Define $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to be a strictly increasing function which is normalized to satisfy $\Phi(0) = 0$ and $\lim_{x\to\infty} \Phi(x) = 1$. Moreover, let $\{\rho_T\}$ be a sequence of positive constants which slowly converges to zero, in particular much slower than $O(T^{-1/2})$. Premultiplying $\sqrt{T}\hat{\mathcal{D}}_T(u)$ with the shrinkage factor ρ_T and then applying the function Φ yields the quantity $\hat{q}_T(u) = \Phi(\rho_T\sqrt{T}\hat{\mathcal{D}}_T(u))$ which has the property that

$$\hat{q}_T(u) \xrightarrow{p} \begin{cases} 0 & \text{for } u \geq u_0 \\ 1 & \text{for } u < u_0. \end{cases}$$

Hence, $\hat{q}_T(\cdot)$ behaves approximately like the step function $1(\cdot < u_0)$ which has a jump at the point u_0.

We next use the quantity $\hat{q}_T$ to construct a criterion function which is minimized approximately at u_0. In particular, we define

$$\hat{Q}_T(u) = u + (1-u)\hat{q}_T(u).$$

Since $\hat{q}_T(\cdot)$ roughly behaves like the indicator $1(\cdot < u_0)$, the function $\hat{Q}_T(u)$ should be minimized at a point close to u_0. Indeed, the asymptotic counterpart $Q(u) = u + (1-u)1(u < u_0)$ of $\hat{Q}_T(u)$ is easily seen to take its minimum exactly at u_0. These considerations suggest to estimate u_0 by

$$\hat{u}_0 := \operatorname*{argmin}_{u\in[0,1]} \hat{Q}_T(u).$$

The estimator $\hat{u}_0$ implicitly depends on the choice of the transformation function Φ and the shrinkage factor ρ_T. A natural choice of Φ and ρ_T is suggested by our asymptotic results on $\hat{u}_0$ which show that $\hat{u}_0$ consistently estimates u_0 and provide its convergence rate. The details can be found in Vogt & Dette [5].

Our estimation method is very general in nature and allows to deal with a wide variety of stochastic features including the mean, covariances, higher moments and the distribution function of the process under consideration. For some special cases, in particular for the case that $\lambda_{t,T} = \mathbb{E}[X_{t,T}]$, the literature provides some alternative approaches to estimate u_0. To start with, it is possible to use change point techniques; cp. Müller [4] among others. Moreover, Mallik et al. [2, 3] propose a p-value based procedure in this special case which relies on a similar idea as our method. Finally, it is also possible to apply sequential testing ideas.

References

[1] R. Dahlhaus, *Fitting time series models to nonstationary processes*, Annals of Statistics **25** (1997), 1–37.

[2] A. Mallik, B. Sen, M. Banerjee and G. Michailidis, *Threshold estimation based on a p-value framework in dose-response and regression settings*, Biometrika **98** (2011), 887–900.

[3] A. Mallik, M. Banerjee and B. Sen, *Asymptotics for p-value based threshold estimation in regression settings*, Preprint (2013).

[4] H.-G. Müller, *Change-points in nonparametric regression analysis*, Annals of Statistics **20** (1992), 737–761.
[5] M. Vogt and H. Dette, *Detecting smooth changes in locally stationary processes*, Preprint (2013).

Large Volatility Matrix Estimation for High-Frequency Financial Data

Yazhen Wang
(joint work with Donggyu Kim)

High-frequency financial data on assets' prices are often modeled by diffusion processes with micro-structure noise, and multi-scale realized volatility, realized kernel estimator and pre-averaging estimator are common estimators of the integrated volatility matrix. For problems involving a large number of assets, we want to estimate volatility matrices of large size. These existing volatility estimators work well for a small number of assets but are inconsistent when both the number, p, of the assets and the average sample size, n, of the price data on the p assets go to infinity. We propose a new type of estimators for the integrated volatility matrix by thresholding multi-scale realized volatility, realized kernel estimator and pre-averaging estimator. We establishes asymptotic theory for the proposed large matrix estimators in the framework that allows both n and p to approach to infinity. The established theory demonstrates that the proposed estimators achieve high convergence rates under a sparsity assumption on the integrated volatility matrix. The numerical studies illustrate that the proposed estimators perform well for large p and complex price and volatility models.

References

[1] O. E. Barndorff-Nielsen, P.R. Hansen, A. Lunde, and N. Shephard, *Multivariate realized kernels: Consistent positive semi-denite estimators of the covariation of equity prices with noise and nonsynchronous trading,* Journal of Econometrics **162**(2011), 149–169.
[2] K. Christensen, S. Kinnebrock, and M. Podolskij, *Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data,* Journal of Econometrics **159**(2010), 116–133.
[3] M. Tao, Y. Wang, and X. Chen, *Fast convergence rates in estimating large volatility matrices using high-frequency financial data,* Econometric Theory **29** (2013), 838–856.
[4] M. Tao, Y. Wang, Q. Yao, and J. Zou, *Large volatility matrix inference via combining low-frequency and high-frequency approaches,* Journal of the American Statistical Association **106**(2011), 1025–1040.
[5] M. Tao, Y. Wang, and H. H. Zhou, *Optimal sparse volatility matrix estimation for high dimensional Itô processes with measurement errors,* Annals of Statistics **41**, 1816–1864.
[6] Y. Wang and J. Zou, *Vast volatility matrix estimation for high-frequency financial data,* Annals of Statistics **38** (2010), 943–978.

Simultaneous Inferences on Sample Covariances

Han Xiao
(joint work with Wei Biao Wu)

Testing for serial correlation has been extensively studied in both statistics and econometrics, and it is a standard diagnostic procedure after a model is fitted to a time series. A natural omnibus choice is to use the maximum sample autocorrelation as the test statistic. We consider a more general problem: the limiting distribution of the maximum deviation

$$\max_{1\le k\le n} \sqrt{n}|\hat{\gamma}_k - \gamma_k|,$$

where γ_k is the autocovariance at lag k, $\hat{\gamma}_k$ is the sample version, n is the length of the time series, and s_n satisfies $s_n \to \infty$ and $s_n/n \to= 0$. The problem is also related to the uniform convergence rate of the sample autocovariances, which is useful in determining the order of a linear system [1] and bandwidth selection for spectral density estimation [2, 3]. Recently, [4] obtained the limiting distribution for linear processes, with s_n growing with a logarithmic speed. Under the general framework of causal representation and the associated physical dependence measures [5], we show that the asymptotic distribution is Gumbel for general nonlinear processes, allowing s_n to grow as a power of n.

Theorem 1 [6] *Under suitable conditions, if s_n satisfies $s_n \to \infty$ and $s_n = O(n^\eta)$, where η depends on the moment condition and the dependence of the underlying process, then for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\max_{1\le k\le s_n} |\sqrt{n}[\hat{\gamma}_k - (1-k/n)\gamma_k]| \le \sqrt{\sigma_0}(a_{2s_n}\, x + b_{2s_n})\right) = \exp\{-e^{-x}\},$$

where $\sigma_0 = \sum_k \gamma_k^2$, and

$$a_n = (2\log n)^{-1/2} \qquad b_n = (2\log n)^{1/2} - (8\log n)^{-1/2}(\log\log n + \log 4\pi).$$

The main techniques for proving this theorem are three approximations: m-dependence approximation, Poisson approximation and Gaussian approximation. It turns out the later two approximations can be used to study the maximum deviation of sample covariances, under the setting of high dimensional statistics.

Let $\boldsymbol{X}_n = (X_{ij})$ be a $n \times m$ data matrix whose n rows are independent and identically distributed as some population distribution with mean vector $\boldsymbol{\mu}_n$ and covariance matrix $\Sigma = (\sigma_{ij})$. We consider the high dimensional paradigm where $m = m_n$ grows to infinity as n does. Motivated by testing high dimensional covariance structure, we study the asymptotic distribution of the following maximum deviation

$$M_n = \max_{1\le i<j\le m} \frac{|\hat{\sigma}_{ij} - \sigma_{ij}|}{\sqrt{\hat{\tau}_{ij}}},$$

where $\hat{\tau}_{ij}$ is the sample estimate of $\mathrm{Var}[(X_{1i} - \mu_i)(X_{1j} - \mu_j)]$. The problem was first studied by [7], and followed by many other authors, see [8] and references therein. Most of these works assumed that entries of $\boldsymbol{X}$ are i.i.d. [8] also considered the maximum sample covariance outside a band around the main diagonal, assuming the population distribution is Gaussian, and Σ has a banded structure. We allow Σ to have a more general dependence structure, and do not require the population distribution to be Gaussian.

Theorem 2 [9] *Under suitable moment conditions and weak dependence conditions, if m_n grows to infinity with a suitable speed, we have for any $y \in \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(nM_n^2 - 4\log m + \log(\log m) + \log(8\pi) \le y\right) = \exp\left(-e^{-y/2}\right).$$

The growth speed of $m = m_n$ depends on the moment condition on the entries of $\boldsymbol{X}$. If they have uniform finite moments, then m_n can grow as a power of n, and if the elements of $\boldsymbol{X}$ have uniform exponential moments, m_n is allowed to grow exponentially. The weak dependence condition requires that the dependence of the true population distribution, as reflected by Σ, cannot be too strong. For example, an important application of our result is on testing for stationarity. If the population distribution is given by stationary process, we only require the autocovariance $\gamma_k = o(1/\log k)$. So in fact, our "weak dependence" condition is very mild.

This result can be used to test whether the population distribution is independent, whether Σ is banded, and whether Σ is Toeplitz. It is also related to the problem of testing whether two covariance matrices are equal.

References

[1] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. John Wiley & Sons Inc., New York, 1988.

[2] D. Politis. Adaptive bandwidth choice. *J. Nonparametr. Stat.*, 15: 517–533, 2003.

[3] E. Paparoditis and D. Politis. Nonlinear spectral density estimation: thresholding the correlogram. *J. Time Series Anal.*, 33: 386–397, 2012.

[4] M. Jirak. On the maximum of covariance estimators. *Journal of Multivariate Analysis*, 102: 1032 – 1046, 2011.

[5] W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102: 14150–14154, 2005.

[6] H. Xiao and W. B. Wu. Portmanteau test and simultaneous inference for serial covariances. *Statistica Sinica*, to appear.

[7] T. Jiang. The asymptotic distributions of the largest entries of sample correlation matrices. *Ann. Appl. Probab.*, 14(2):865–880, 2004.

[8] T. Cai and T. Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.*, 39(3):1496–1525, 2011.

[9] H. Xiao and W. B. Wu. Asymptotic theory for maximum deviations of sample covariance matrix estimates. *Stochastic Process. Appl.*, 123 (7): 2899–2920, 2013.

Heteroscedasticity and autocorrelation robust structural change detection

Zhou Zhou

Structural stability over time is important in many scientific endeavors. For most of the frequently used statistical tests of structural change, the assumption of (weak) stationarity under the null hypothesis is crucial for their validity. However, the stationarity assumption has become restrictive for many contemporary structural change analysis. To simplify discussion, let us consider the test of structural change in mean where we observe time series $\{X_i\}_{i=1}^n$ with $\mathbb{E}[X_i] = \mu_i$ and we are interested in testing whether μ_i remains constant over time; namely testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu, \quad \longleftrightarrow \quad H_a : \mu_i \neq \mu_j \tag{1}$$

for some $1 \leq i < j \leq n$. For most of conventional tests of H_0, the covariance structure of $\{X_i\}$ should remain unchanged over time. In other words, the latter tests are applicable to time series of the form $X_i = \mu_i + e_i$, where $\{e_i\}$ is a zero-mean weakly stationary sequence. Nevertheless, the stationarity assumption is unrealistic in many important current applications.

When the covariance structure of the time series is varying, it is shown in this talk that most of conventional tests of H_0 are inconsistent and can lead to biased testing results. To understand this, think of the classic cumulative sum (CUSUM) test

$$T_n = \max_{1 \leq i \leq n} |S_i - t_i S_n|, \text{ where } S_i = \sum_{j=1}^{i} X_j \text{ and } t_i = i/n. \tag{2}$$

The classic idea to perform this test as well as most other tests of structural change is normalization. More specifically, one normalizes T_n by a consistent or inconsistent estimator of $\mathrm{Cov}(S_n)/n$ to make the test asymptotically pivotal. Critical values of the test can then be obtained accordingly. Nevertheless, when $\{X_i\}$ is second order non-stationary, it is found in this paper that the behavior of T_n under H_0 is determined by a centered Gaussian process with very complex covariance structure. As a consequence it is generally impossible to make T_n pivotal by normalizing it with one or even a sequence of covariance estimators. The complicated non-stationary dynamics in the second order structure in time series has posted new challenges to the classic problem of structural change detection. To date, little progress has been made toward structural change tests that are robust to heteroscedasticity and autocorrelation of general forms.

The contents of the talk are mostly based on the results of Zhou (2013). We propose a simple bootstrap procedure that is shown to be consistent under general forms of abrupt and smooth changes in the temporal dynamics of the time series. More specifically, we discover and utilize a somewhat surprising observation that, for a wide class of non-stationary times series, progressive convolutions of their block sums and i.i.d. standard normal random variables consistently mimic the complex joint probabilistic behavior of their partial sum processes. Hence structural change tests for non-stationary time series can be easily performed by

generating large samples of the latter convolutions. While remaining consistent for a much larger class of time series, the proposed bootstrap procedure is shown to have the same rate of accuracy (in terms of estimating the true covariance structure) and can detect local alternatives with the same $\sqrt{n}$ parametric rate as the conventional tests. The above theoretical findings are supported by our finite sample Monte Carlo experiments in which it is found that our bootstrap enjoys similar accuracy and power to the conventional tests when the time series is second order stationary. However, for second order non-stationary time series, our Monte Carlo simulations show that the robust bootstrap remains accurate while the conventional tests are invalid, as indicated by our theoretical findings.

References

[1] Z. Zhou, *Heteroscedasticity and Autocorrelation Robust Structural Change Detection*, Journal of the American Statistical Association **108** (2013), 726-740.

Reporter: Rainer Dahlhaus

Participants

Prof. Dr. Alexander Aue
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
UNITED STATES

Prof. Dr. Peter J. Brockwell
Department of Statistics
Colorado State University
Fort Collins CO 80523-1877
UNITED STATES

Prof. Dr. Rong Chen
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854-8019
UNITED STATES

Prof. Dr. Rainer Dahlhaus
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Prof. Dr. Richard A. Davis
Department of Statistics
Columbia University
1255 Amsterdam Ave., MC 4690
New York, NY 10027
UNITED STATES

Prof. Dr. Herold Dehling
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

Prof. Dr. Manfred Deistler
Institut für Ökonometrie und
Operations Research
Technische Universität
Argentinierstraße 8
1040 Wien
AUSTRIA

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

Prof. Dr. Paul Doukhan
Département de Mathématiques
Université de Cergy-Pontoise
Site Saint-Martin, BP 222
2, avenue Adolphe Chauvin
95302 Cergy-Pontoise Cedex
FRANCE

Prof. Dr. Michael Eichler
Dept. of Quantitative Economics
University of Maastricht
Postbus 616
6200 MD Maastricht
NETHERLANDS

Prof. Dr. Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton NJ 08544
UNITED STATES

Prof. Dr. Jürgen Franke
Fachbereich Mathematik
T.U. Kaiserslautern
Erwin-Schrödinger-Straße
67653 Kaiserslautern
GERMANY

Prof. Dr. Piotr Fryzlewicz
Department of Statistics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Prof. Dr. Yulia R. Gel
Department of Statistics and
Actuarial Science
University of Waterloo
Faculty of Mathematics
Waterloo Ontario N2L 3G1
CANADA

Prof. Dr. Wolfgang Karl Härdle
Wirtschaftswissenschaftliche Fakultät
Ladislaus v. Bortkiewicz Chair of
Statistics CASE
Humboldt-Universität zu Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Jean Jacod
IMJ
Université P. et M. Curie
4, Place Jussieu
75252 Paris Cedex 05
FRANCE

Dr. Carsten Jentsch
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68161 Mannheim
GERMANY

Dr. Claudia Kirch
Institut für Stochastik
Karlsruher Institut f. Technologie (KIT)
Kaiserstr. 89-93
76133 Karlsruhe
GERMANY

Prof. Dr. Claudia Klüppelberg
Zentrum Mathematik
TU München
Boltzmannstr. 3
85748 Garching b. München
GERMANY

Prof. Dr. Jens-Peter Kreiß
Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig
GERMANY

Prof. Dr. Clifford Lam
Department of Statistics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Prof. Dr. Anne Leucht
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68131 Mannheim
GERMANY

Prof. Dr. Alexander Lindner
Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig
GERMANY

Prof. Dr. Oliver Linton
Faculty of Economics
University of Cambridge
Sydewick Avenue
Cambridge CB3 9DD
UNITED KINGDOM

Prof. Dr. Enno Mammen
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68161 Mannheim
GERMANY

Prof. Dr. Maria E. Mancino
Dept. of Economics and Management
Universita di Firenze
Via Delle Pandette 9
50134 Firenze
ITALY

Prof. Dr. Thomas Mikosch
Department of Mathematics
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
DENMARK

Prof. Dr. Eric Moulines
Departement TSI
E.N.S. Telecommunications
46, rue Barrault
75634 Paris Cedex
FRANCE

Prof. Dr. Per Mykland
Department of Statistics
The University of Chicago
5734 University Avenue
Chicago, IL 60637-1514
UNITED STATES

Prof. Dr. Michael Neumann
Institut für Mathematik
Friedrich-Schiller-Universität
Ernst-Abbe-Platz 2
07743 Jena
GERMANY

Prof. Dr. Efstathios Paparoditis
Dept. of Mathematics & Statistics
University of Cyprus
P.O. Box 20537
1678 Nicosia
CYPRUS

Prof. Dr. Mark Podolskij
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Prof. Dr. Wolfgang Polonik
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
UNITED STATES

Prof. Dr. Mohsen Pourahmadi
Department of Statistics
Texas A & M University
College Station, TX 77843-3368
UNITED STATES

Dr. Philip Preuß
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

Prof. Dr. Suhasini Subba Rao
Department of Statistics
Texas A & M University
College Station, TX 77843-3143
UNITED STATES

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Francois Roueff
Departement TSI
E.N.S. Telecommunications
46, rue Barrault
75634 Paris Cedex
FRANCE

Christian Schmidt
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Prof. Dr. Vladimir G. Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin
GERMANY

Prof. Dr. David S. Stoffer
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260
UNITED STATES

Prof. Dr. Howell Tong
Department of Statistics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Sophon Tunyavetchakit
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Dr. Mathias Vetter
Ruhr-Universität Bochum
Fakultät für Mathematik
Lehrstuhl für Stochastik
44780 Bochum
GERMANY

Dr. Michael Vogt
Faculty of Economics
University of Cambridge
Sydewick Avenue
Cambridge CB3 9DD
UNITED KINGDOM

Prof. Dr. Rainer von Sachs
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 20
1348 Louvain-la-Neuve
BELGIUM

Prof. Dr. Yazhen Wang
Department of Statistics
University of Wisconsin
Medical Science Center
1300 University Avenue
Madison, WI 53706
UNITED STATES

Prof. Dr. Wei-Biao Wu
Department of Statistics
The University of Chicago
5734 University Avenue
Chicago, IL 60637-1514
UNITED STATES

Han Xiao
Dept. of Statistics & Biostatistics
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854-8019
UNITED STATES

Prof. Dr. Qiwei Yao
Department of Statistics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Prof. Dr. Zhou Zhou
Department of Statistics
University of Toronto
100 St. George Street
Toronto ONT M5S 1A1
CANADA