Mathematisches Forschungsinstitut Oberwolfach

Report No. 12/2016

DOI: 10.4171/OWR/2016/12

# New Developments in Functional and Highly Multivariate Statistical Methodology

Organised by
Gerda Claeskens, Leuven
Holger Dette, Bochum
Irène Gijbels, Leuven
Peter Hall, Parkville VIC

21 February – 27 February 2016

Abstract. The central focus of the workshop was on recent developments in statistical techniques for highly multivariate data and functional data. The programme delivered talks on state-of-the-art research in the area, with a number of talks on highly-dimensional multivariate settings as well as talks dealing with functional data. The talks were followed by lively discussions on how to tackle difficult issues in the statistical methodology for such complex data.

*Mathematics Subject Classification (2010):* 62Hxx, 62Gxx, 62Mxx and 62Pxx.

## Introduction by the Organisers

The workshop "New Developments in Functional and Highly Multivariate Statistical Methodology" organized by Gerda Claeskens (KU Leuven), Holger Dette (Ruhr-Universität Bochum), Irène Gijbels (KU Leuven) and Peter Hall (University of Melbourne) was held from February 22 till February 27, 2016. The meeting was very well attended with 26 participants, coming from all over the globe: Europe, North America, and Australia. During the workshop we had 17 talks of about 50 minutes each followed by a discussion of about 10 to 15 minutes. We also had four talks by young researchers on Tuesday afternoon, each of them presenting their research results in this area during 25 minutes, followed by a short discussion. One of the senior participants could be present thanks to a *Simons Visiting Professorship*, and three of the young researcher participants got travel support

through special grants (OWLG, NSF). In addition the workshop was attended by one of the annual laureates of the DMV-students conference.

During the workshop we also organized a "floor discussion". During this session on Thursday late afternoon many thoughts on interesting future research directions were formulated. Some uncharted challenging areas were highlighted, which require attention to develop novel statistical methodology. The session unfolded in a lively discussion with a lot of brainstorming on further scientific progress.

Unfortunately one of the organizers, Professor Peter Hall passed away on January 9, 2016. Since many of the workshop's participants collaborated with Peter, we organized a small memorial session, in which attention went to his outstanding qualities as a scientist that we were so fortunate to witness and learn from. During the many small contributions to this session the main focus was on his extraordinary human personality that we all knew and will dearly miss.

The focus of the workshop was on recent developments in statistical techniques for highly multivariate data and functional data. The key issue is to subtract from the data valid conclusions regarding the stochastic process that led to the observed data, and further, amongst other objectives, to perform good predictions for some quantities of interest. Standard statistical techniques can be inappropriate for the analysis of highly multivariate or complex data. There were several talks on flexible regression models, for functional data as well as for high-dimensional multivariate data. This also included classification problems and statistical methodology when dealing with the complication of having incomplete data. Other talks discussed specific issues of variable selection and model selection. Classical dimension reduction methods, such as principal component analysis, and how they are adopted to functional data as well as to vector time series led to interesting scientific discussions. Other talks dealt with recent developments in inverse or deconvolution problems in the context of high-dimensional or functional data. All talks contributed to initiate lively discussions on how to arrive at valid conclusions regarding the data generating stochastic process.

The excellent scientific programme was complemented by the Wednesday afternoon hike to St. Roman, that was done jointly with participants of the workshop on "Asymptotic Geometric Analysis" that took place the same week. In fact there were some common scientific elements of interest between the two workshops, in particular with respect to probability theory (characteristic inequalities for Gaussian processes, small ball probabilities, depths in functional spaces). Participants from both workshops attended some talks at the other workshop.

*Acknowledgement:* The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Alexander Aue in the "Simons Visiting Professors" program at the MFO.

## Workshop: New Developments in Functional and Highly Multivariate Statistical Methodology

## Table of Contents

# Abstracts

## Fréchet Regression for Random Objects

Hans-Georg Müller
(joint work with Alexander Petersen)

We introduce the notion of Random Objects as complex data points that are elements of a bounded metric space that is not necessarily a vector space. Therefore, Hilbert space based methods that are commonly used in Functional Data Analysis, such as functional principal components, cannot be used. Some Random Objects can be transformed into a Hilbert space and therefore can be considered to be "quasi-Hilbert".

An example for such "quasi-Hilbert" data are random samples of density functions, due to the constraint that they are non-negative and integrate to 1. The analysis of random density functions was pioneered by [1]. Due to their inherent constraints, densities do not live in a vector space and therefore commonly used Hilbert space based methods of functional data analysis are not applicable. To address this problem, we consider a transformation approach, mapping probability densities to a Hilbert space of functions by applying a continuous and invertible map.

Basic methods of functional data analysis, such as the construction of functional modes of variation, functional regression or classification, can then be implemented by using representations of the densities in this linear space. Representations of the densities themselves are obtained by applying the inverse map from the linear functional space to the density space. Transformations of interest include log quantile density and log hazard transformations, among others. Rates of convergence are derived for the representations that are obtained for a general class of transformations under certain structural properties. If the subject-specific densities need to be estimated from data, these rates correspond to the optimal rates of convergence for density estimation. The proposed methods are illustrated through simulations and applications in brain imaging, see [2].

Next we consider the case where the random objects are not quasi-Hilbert, i.e., there is no smooth and invertible transformation to a Hilbert space available. For this case, we consider a regression scenario where the predictors are scalars or vectors and the responses are random objects. This will address the increasing need for statistical tools to analyze complex data that are non-Euclidean and specifically do not lie in a vector space.

To address the need for statistical methods for this regression scenario, we introduce the concept of Fréchet regression. This is a general approach to regression when responses are complex random objects in a metric space and predictors are in $\mathcal{R}^p$. We develop generalized versions of both global least squares regression and local weighted least squares smoothing.

The target quantities are appropriately defined population versions of global and local regression for the case of metric-space valued responses. We derive

asymptotic rates of convergence for the corresponding sample based fitted regressions to the population targets under suitable regularity conditions by applying empirical process methods.

For the special case of random objects that reside in a Hilbert space, for example regression models with vector predictors and functional data or more general Hilbert space valued data as responses, we obtain a limit distribution. The proposed methods have broad applicability. Illustrative examples include responses that consist of probability distributions and correlation matrices, and we demonstrate the proposed Fréchet regression for demographic and brain imaging data.

### References

[1] A. Kneip and K.J. Utikal, *Inference for density families using functional principal component analysis*, Journal of the American Statistical Association **96** (2001), 519–542.

[2] A. Petersen and H.G. Müller, *Functional data analysis for density functions by transformation to a Hilbert space*, Annals of Statistics **44** (2016), 183–218.

## Nonparametric Registration to Low-Dimensional Functionspaces

Alois Kneip

(joint work with Heiko Wagner)

This report provides some resume of the paper of Wagner and Kneip (2016). The data that we consider are a sample of i.i.d. smooth random functions $x_1, \ldots, x_n$ defined over a closed interval on the real line. Registration literature focuses on the situation that all functions share a common set of shape features, such as peaks and valleys. The sizes of the features vary, and we refer to this as *amplitude variation*. The locations of the features also vary from curve to curve, which indicates the existence of *phase variation*. Generally speaking, registration deals with separating amplitude and phase variation in a statistically meaningful way. The aim is to search for a set of smooth strictly monotonic functions $h_i$, called *warping function*, which eliminate phase variation such that the *registered* functions $y_i(t)$ of the form $y_i(t) = x_i[h_i(t)] = (x_i \circ h_i)(t)$ represent amplitude variation. Since monotone transformations do not destroy shape features the registered functions will possess the same sequences of peaks and valleys as the original functions $x_i$.

It is well-known that phase variation is present in many important applications, and it poses severe problems for the application of functional versions of commonly used multivariate data analyses such as computing pointwise means, variances and correlations; principal components analysis and canonical correlation analyses.

Traditional literature on the registration problem aims to define warping functions in such a way that registered functions $y_i$ have all shape features aligned. A common property of the most important methods proposed in this context is to determine warping functions $h_i$ by minimizing a distance $d(x_i \circ h_i, \gamma)$ between registered functions $y_i(t) = x_i(h_i(t))$ and a template $\gamma(t)$. There is a considerable literature proposing algorithms which aim to minimize the distance $d_2(x_i \circ h_i, \gamma) = \|x_i \circ h_i - \gamma\|_2$, where $\|\cdot\|_2$ denotes the $L^2$-distance. Well-known

problems with these techniques have lead to the development of more sophisticated techniques based on alternative distance measures. For example, [3], [5, 6, 4], or [2] propose to minimize semi-metrics with the property that $d(x_i(h_i), \gamma) = 0$ if $x_i \circ h_i = a_i\gamma$ for some $a_i \in \mathbb{R}$.

All these methods share a common point of view. The success of a registration method is assessed in terms of how well it is able to align visible features. Templates are often determined iteratively from the sample and their construction aims to establish a "structural mean" which possess all common shape features at mean locations and with mean amplitude. Hence, traditionally registration tends to concentrate on establishing a most informative mean curve summarizing the sample functions.

In our paper we consider registration from a more general point of view. Registration may be used as a tool for statistical analysis whenever the random functions $x_i$ possess "bounded shape variation", i.e. there exists a fixed value $\mathbf{q} < \infty$ such that with probability 1 the number of shape features to be found within each possible realization does not exceed $\mathbf{q}$. Our approach is based on an observation already made by [1] that for random functions with bounded shape variation there exists a finite $K$ and warping functions $h_i$ such that with probability 1 have

$$x_i(h_i(t)) = \sum_{j=1}^{K} a_{ij}\gamma_j(t) \tag{1}$$

for some basis functions $\gamma_1, ..., \gamma_K$ and individually different coefficients $a_{i1}, ..., a_{iK}$.

For random functions with bounded shape variation there exists a minimal dimension $\mathbf{K}_0 < \infty$ such that (1) holds for all $K \geq \mathbf{K}_0$. Conventional registration procedures based on minimizing semi-metrics with respect to a template function are implicitly based on assuming $\mathbf{K}_0 = 1$.

In our paper we are going beyond [1] by studying decomposition (1) from a theory-guided, conceptional point of view and by deriving some basic inference results for situations, where the true functions have to be reconstructed from discrete, noisy observations. Appropriate values of $K$ depend on the structure of $x_i$, and possible non-uniqueness of solutions to (1) are resolved by selecting the registration procedure with the least complex warping functions. Furthermore, we present a new algorithm which estimates the components of (1) for all possible values of $K$ and seems to work well for many applications.

The generality of our approach is illustrated by simulated random functions in Figure 1. These functions do not possess a clearly visible "common shape". But a closer look at the unregistered curves shows that there exist structurally similar curves which quite obviously exhibit some phase variation. Nevertheless, these are not the type of data that may be registered by any conventional method. Indeed, fitting a $K = 1$ dimensional model leads to unreasonable results with extreme warping functions. On the other hand, the number of local extrema of these functions varies between 1 and 3, and these random functions are of bounded shape variation. Indeed the true minimal dimension is $K = 2$, and the $K = 2$ dimensional registration presented in the figure rests upon structurally

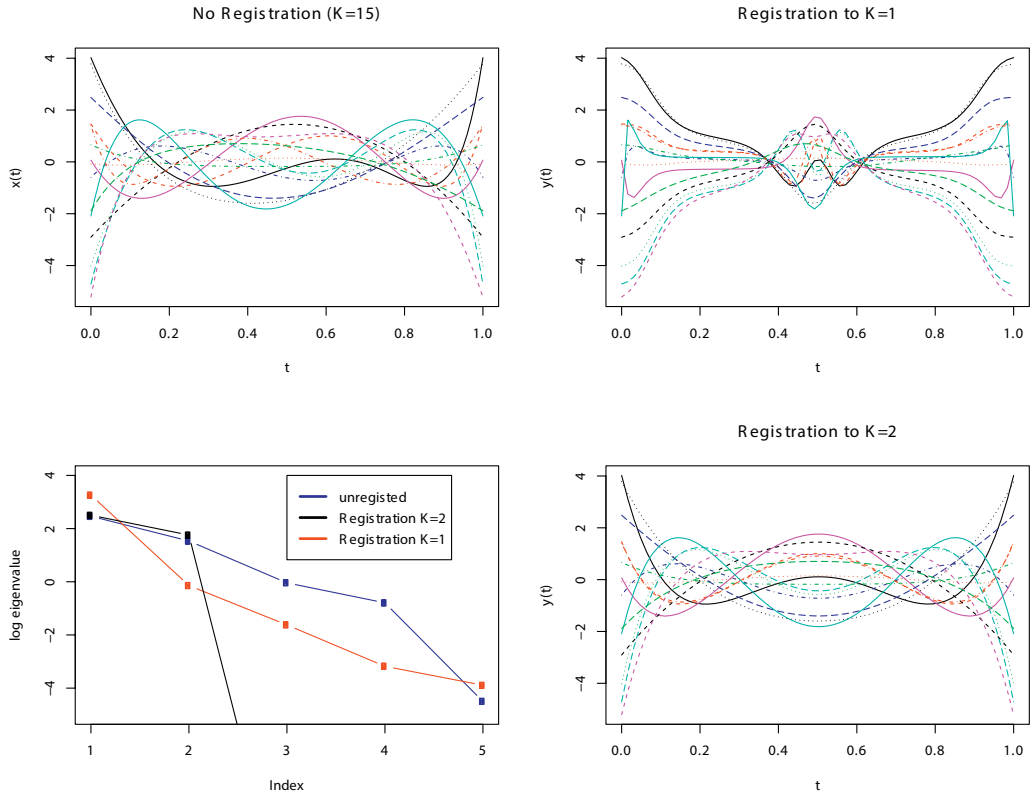


FIGURE 1. Random functions with bounded shape variation

simple warping functions (which can themselves be described by a one dimensional model).

Assuming that functional shapes are of bounded complexity does not seem to be restrictive in important applications biomedicine, technics, chemometrics, etc., and often the presence of phase variation is already imposed from a substantial point of view (different reactions times, etc.). Our approach then generalizes the rather limited range of applicability of traditional registration techniques. Together with a suitable analysis of warping functions, the method allows to decompose functional data in a way that might be more informative than standard functional principal component analysis (FPCA). In our paper the approach is illustrated by applications to human growth curves and gene expression data.

## REFERENCES

[1] A. Kneip and J.O. Ramsay, *Combining Registration and Fitting for Functional Models*, Journal of the American Statistical Association **103** (2008), 1155–1165.

[2] A. Srivastava, W. Wu, S. Kurtek, E. Klassen and J.S. Marron, *Registration of Functional Data Using Fisher-Rao Metric*, (2011) http://arxiv.org/pdf/1103.3817

[3] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer Series in Statistics, 2005

[4] K. Wang and T. Gasser, *Synchronizing sample curves nonparametrically*, The Annals of Statistics, **27** (1999), 439–460

[5] K. Wang and T. Gasser, *Alignment of Curves by Dynamic Time Warping*, The Annals of Statistics, **25** (1997), 1251–1276
[6] K. Wang and T. Gasser, *Asymptotic and bootstrap confidence bounds for the structural average of curves*, The Annals of Statistics, **26** (1998), 972–991

## Optimal classification and nonparametric regression for functional data

Alexander Meister

We establish minimax convergence rates for classification of functional data and for nonparametric regression with functional design variables. The optimal rates are of logarithmic type under smoothness constraints on the functional density and the regression mapping, respectively. These asymptotic properties are attainable by conventional kernel procedures. The bandwidth selector is automatically adaptive. In this work the functional data are considered as realisations of random variables which take their values in a general Polish metric space. We impose certain metric entropy constraints on this space; but no algebraic properties are required.

### References

[1] A. Meister, *Optimal classification and nonparametric regression for functional data*, Bernoulli, to appear.

## High Dimensional M-estimation and Inference for Left-Censored Models

Jelena Bradic
(joint work with Jiaqi Guo)

Left-censored data is characteristic of many datasets, due to the inherent limit of detection and the limit of quantitation in the measurements. Estimation in the left-censored models has been studied since the 1950's (e.g., [14]). The most common approach is to consider a data transformation model and then impose a class of distributions for the resulting model errors. However, as Zellner noted in [16], knowledge of the underlying data generating mechanism is seldom available, and thus models with parametric distributions may be subject to distributional misspecifications, which in turn lead to inconsistent or inefficient estimates.

[9, 10, 6] pioneered development of robust inference procedures for the left-censored data. A series of work proceeded [8], [3], [4], [19], [12], [18], and developed rank based and estimating equations estimators with the number of parameters, $p$, smaller than the number of observations, $n$. However, in the past decade, in order to build better models of high-dimensional datasets, the trend has been to use models with $p \gg n$. [5] offered a penalized version of the Powell's estimator and its oracle estimation bounds. However, substantially smaller efforts have been made toward high-dimensional inference, namely confidence interval construction and statistical testing in the high-dimensional setting. Recently, [15], [1] and [17], [13], [11] have corrected the bias of high-dimensional regularized estimators by a

two-step of a three-step bias correction technique. However, the fundamental limit of the precision of estimation in the Tobit-I models is largely unknown, let alone an adaptive procedure to achieve that limit.

In the case of simple linear models [2] use generalized M-estimators to circumvent the problem of misspecifications of the error distribution, but the idea of censored, robust one-step estimators has never been formally developed. We formally develop the methodology for robust Mallow's, Schweppe's and Hill-Ryan's estimators that adapts to the unknown censoring. Censored robust one-step estimators significantly expand the methodology of one-step estimators and censored estimators. While it is generally understood that under certain conditions the use of debiasing will lead to consistent estimators, deriving the form of the corresponding debiasing step for non-differentiable losses is not apparent; nor is the method of obtaining its asymptotic covariance matrix, as required by the inference.

Without loss of generality, we focus on the zero-censored model

$$y_i = \max\left\{0, x_i\boldsymbol{\beta}^* + \varepsilon_i\right\}. \tag{1}$$

We assume that the following function is uniquely defined,

$$\arg\min_a \mathbb{E}\left[|y_i - a| - |y_i|\big|x_i\right],$$

and the unique minimizer is a median of the response vector $Y$. We assume that this median takes on a form $\max\{0, x_i\boldsymbol{\beta}^*\}$. As initial estimator we consider the penalized censored least absolute deviation (P-CLAD) estimator

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \left\{ n^{-1}\sum_{i=1}^{n} |y_i - \max\{0, x_i\boldsymbol{\beta}\}| + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{2}$$

In the theory of M-estimation, a one-step improvement typically takes the form

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - HS, \tag{3}$$

where a vector $S \in \mathbb{R}^p$ is the direction of the projection, a matrix $H \in \mathbb{R}^{p\times p}$ is the projection length and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda)$ is an initial estimator. The projection direction should be chosen as close as possible to the efficient score vector. As our loss takes the form $n^{-1}\sum_{i=1}^n |y_i - \max\{0, x_i\boldsymbol{\beta}\}|$, we consider a sample plug-in approximation of the form $-n^{-1}\sum_{i=1}^n \operatorname{sign}\left(y_i - \max\{0, x_i\hat{\boldsymbol{\beta}}\}\right) w_i^\top(\hat{\boldsymbol{\beta}})$. where $w_i(\boldsymbol{\beta}) = x_i\mathbf{1}\{x_i\boldsymbol{\beta} > 0\}$. Estimation of the precision matrix in the presence of missing data is particularly difficult. Typical approaches consisting of careful substitution or imputation are not suitable for the high-dimensional data. We develop new methodology that directly explores the zero pattern in the covariance matrix by establishing new connections to the double censored regression framework. Instead of smoothing out the observed zeros, we propose to model them directly and therefore obtain an adaptive inference that is more stable when censoring gets high. This approach leads to the efficient estimation of the uncertainty of the one-step estimates. It is well known that assessing the accuracy of the estimation process in the censored regression is extremely challenging, even for the low-dimensional

problems. Nevertheless, we obtain

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)=\text{Leading Term}$$
$$+O_P\left(\frac{s_{\boldsymbol{\beta}^*}^{3/4}(\log(p\vee n))^{3/4}}{n^{1/4}}+\frac{(s_{\boldsymbol{\beta}^*}+s_j)^2 s_j s_{\boldsymbol{\beta}^*}\sqrt{\log(p\vee n)}}{n^{1/2}}+\frac{s_{\boldsymbol{\beta}^*}^{7/4}(\log(p\vee n))^{5/4}}{n^{3/4}}\right).$$

Moreover, the right hand side stays the same for the case of $v_i = q_i = 1$ and the Mallow's and the Hill –Ryan's weights. For the Schweppe's weight schemes we obtain

$$\sqrt{n}\left(\breve{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)=\text{Leading Term}$$
$$+O_P\left(\frac{s_{\boldsymbol{\beta}^*}^{7/4}(\log(p\vee n))^{3/4}}{n^{1/4}}+\frac{(s_{\boldsymbol{\beta}^*}+s_j)^2 s_j s_{\boldsymbol{\beta}^*}^3\sqrt{\log(p\vee n)}}{n^{1/2}}+\frac{s_{\boldsymbol{\beta}^*}^{11/4}(\log(p\vee n))^{5/4}}{n^{3/4}}\right).$$

## References


[1] A. Javanmard and A. Montanari, *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression*, Journal of Machine Learning Research, **15** (2014), 2869–2909.

[2] C.W. Coakley and T.P. Hettamansperger, *A Bounded Influence, High Breakdown, Efficient Regression Estimator* Journal of the American Statistical Association, **88**(423) (1993), 872–880

[3] A. Golan, G. Judge and J. Perloff, *Estimation and inference with censored and ordered multinomial response data*, Journal of Econometrics, **79** (1) (1997), 23–51.

[4] C. Leng and X. Tong, *A quantile regression estimator for censored data*, Bernoulli, **19** (1) (2013), 344–361.

[5] P. Müller and S. van de Geer, *Censored linear model in high dimensions*, TEST, **25**(1) (2016), 75–92.

[6] W.K. Newey and J.L. Powell, *Efficient estimation of linear and type I censored regression modets under conditional quantile restrictions* Economic Theory **6** (1990), 295–317.

[7] J. Neyman, *Outline of a theory of statistical estimation based on the classical theory of probability*, Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences, **236** (767) (1937), 333–380

[8] J.L. Powell, *Least absolute deviations estimation for the censored regression model*, Journal of Econometrics, **25** (3) (1984), 303–325.

[9] J.L. Powell, *Censored regression quantiles*, Journal of Econometrics, **32** (1986), 143–155.

[10] J.L. Powell, *Symmetrically trimmed Ieast squares estimation for Tobit models*, Econometrica, **54** (1986), 1435–1460.

[11] Z. Ren, T. Sun, C.-H. Zhang and H.H. Zhou, *Asymptotic normality and optimalities in estimation of large Gaussian graphical models*, The Annals of Statistics, **43** (2015), no. 3, 991–1026.

[12] W. Song, *Distribution-free test in Tobit mean regression model*, Journal of Statistical Planning and Inference, **141** (8) (2011), 2891–2901.

[13] T. Sun and C. Zhang, *Scaled sparse linear regression*, Biometrika, **99** (2012), 879–898

[14] J. Tobin, *Estimation of relationships for limited dependent variables*, Econometrica, **26** (1958), 24–36.

[15] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics, **42**(3) (2014), 1166–1202

[16] A. Zellner, *Bayesian method of moments/instrumental variable (BMOM/IV). Analysis of Mean and Regression Models*, Lee, J.C. Zellner, A., Johnson, W.O. (Eds.), Modelling and Prediction: Essays in Honor of Seymour Geisser, Springer, New York, 1996.

[17] C.H. Zhang and S.S. Zhang, *Confidence intervals for low dimensional parameters in high-dimensional linear models*, Journal of Royal Statistical Society. Series B. Statistical Methodology, **76** (1) (2014) 217–242.
[18] Y. Zhao, B.M. Brown and Y-G. Wang, *Smoothed rank-based procedure for censored data*, Electronic Journal of Statistics, **8** (2) (2014), 2953–2974.
[19] X. Zhou and Z. Pan, *Two-step semiparametric estimation of the Type-3 Tobit model*, Statistics & Probability Letters, **105** (2015), 96–105.

## Stringing high-dimensional data into functional data: Application to functional Cox model

Jane-Ling Wang
(joint work with Kun Chen, Kehui Chen, Hans-Georg Müller, Qimeng Qu and Xiao Wang)

There is a close relation between high-dimensional data and functional data. For instance, densely observed functional data can be viewed as high-dimensional data endowed with a natural ordering. In this talk, we explore the opposite question whether one can find a proper ordering of high-dimensional data so they can be reordered and viewed as functional data. In [1], Chen et al. proposed stringing, a method that takes advantage of the high dimension by representing such data as discretized and noisy observations that originate from a hidden smooth stochastic process.

Assuming that the $p$-dimensional features, $X_1, \ldots, X_p$, result from scrambling the original ordering of the observations of the process $X(t)$ at time grid $0 = t_1 < \ldots < t_p = 1$, such that $X(t_j) = X_{\pi(j)}$, where $\pi(j)$ is a permutation function of $j$ and $j = 1, \ldots, p$. That is, $\{\pi(1), \ldots, \pi(p)\}$ is a permutation of $\{1, \ldots, p\}$. Stringing reorders the components of the high-dimensional vectors $(X_1, \ldots, X_p)$ in such a way that $(X(t_1), \ldots, X(t_p))$ is recovered followed by a reconstruction (e.g. through linear or quadratic interpolation) of the process $X(t)$, for $t \in [0, 1]$. Stringing thus transforms the high-dimensional vectors of observations $(X_1, \ldots, X_p)$ into functional data $X(t)$, for $t \in [0, 1]$. Established methods from functional data analysis can then be applied for further statistical analysis once an underlying stochastic process and the corresponding random trajectory $X(t)$ for each subject have been identified.

The reconstruction of the process $X$ from high-dimensional data is implemented with distance-based metric multidimensional scaling (MDS), mapping high-dimensional data to locations on a real interval, e.g. $[0, 1]$, such that features that are close in a suitable sample metric also are located close to each other on the interval. Specifically, MDS ([4, 3] ) aims at mapping $p$ objects to points $s_1, \ldots, s_p$, situated in a low-dimensional space $\mathbb{R}^m$, given distances (or proximities) $D_{jk}$ between any pair of objects $j$ and $k, 1 \leq j, k \leq p$. The configuration of the low-dimensional points is determined by minimizing a cost function, which measures how well a particular configuration in the low-dimensional space approximates the original distances. In our implementation of stringing, we choose $m = 1$, the stress

function by Kruskal [5] as our cost function. For the distance-based metric $d$, we use both the Euclidean distance and a transformed Pearson correlation to estimate $D_{jk}$.

We provide some theoretical support, showing that under certain assumptions, an underlying stochastic process can be constructed asymptotically, as the dimension $p$ of the data tends to infinity. The advantages of the stringing methodology is illustrated through two data sets: 1) the analysis of tree ring data, and 2) the prediction of survival time from high-dimensional gene expression data. In regression applications involving high-dimensional predictors, stringing compares favorably with existing methods. A byproduct of the survival data analysis in 2) is a new Cox model that accommodates functional covariates and takes the form:

$$h(t|X_i) = h_0(t) \exp\left[\int \beta(s) X_i(s)\, ds\right], \tag{1}$$

where for a baseline hazard function $h_0(t)$, the conditional hazard rate for a given functional covariate $Z_i$ is $h(t|Z_i)$.

In model (1), the entire covariate trajectory relates to the hazard function through the coefficient function $\beta$. This is very different from the Cox model with time-varying covariates which takes the form:

$$h(t|X_i) = h_0(t) \exp\left[\beta(t) X_i(t)\right]. \tag{2}$$

This model is concurrent in the sense that only the current covariate value $X_i(t)$ at time $t$ relates to the hazard function at time $t$. Statistical inference is much more challenging for the functional Cox model in (1), as it presents an ill-posed problem.

A penalized likelihood approach using reproducing kernel Hilbert space is proposed in Qu.et al. [2] to estimate the regression parameter. This involves a more general model that accommodates baseline covariates $Z$ as well as functional covariates $X(t)$. Specifically, let $Z$ be a $p$-dimensional vector and $X$ be a functional covariate, the model considered in Qu et al. (2016) takes the form:

$$h(t|Z_i, X_i(t)) = h_0(t) \exp\left[\theta Z_i + \int \beta(s) X_i(s)\, ds\right]. \tag{3}$$

Penalized partial likelihood was employed to estimate the regression vector parameter $\theta$ and the coefficient function $\beta$. The resulted estimates for $\theta$, which maximizes the penalized partial likelihood is shown to be $\sqrt{n}$-consistent and semiparametrically efficient, and the estimate for $\beta$ is shown to attain the optimal rate of convergence.

## References

[1] K. Chen, K. Chen, H.-G. Müller and J.-L. Wang, *Stringing high dimensional data for functional analysis*, Journal of the American Statistical Association **106** (2011), 275–284.

[2] S. Qu, X. Wang and J.-L. Wang, *Optimal estimation for the functional Cox model*, Accepted by the Annals of Statistics.

[3] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, London: Chapman & Hall, 2001.
[4] J. Ramsay, *Some statistical approaches to multidimensional scaling data*, Journal of the Royal Statistical Association Series A **145** (1982), 285–312.
[5] J. Kruskal, *Nonmetric multidimensional scaling: a numerical method*, Psychometrika **29** (1964), 115–129.

## Nonparametric first-order analysis of spatial and spatio-temporal point processes. Application to wildfire patterns

Wenceslao González-Manteiga
(joint work with María Isabel Borrajo, Isabel Fuentes-Santos and Jorge Mateu)

### 1. Introduction

Data irregularly distributed in the spatial or spatio-temporal domain arise in a wide variety of scientific contexts, including seismology, forestry, geography and epidemiology. Wildfire is the most ubiquitous natural disturbance in the world and represents a problem of considerable social and environmental importance; particularly, in Galicia (NW Spain) arson fires are the main cause of forest destruction. Knowing the spatial distribution of forest fires and whether this distribution varies over time would be a key factor for future development of fire prevention and fire fighting plans. A point process is a stochastic process that generates a random collection of events in some metric space. Spatial point processes generate events in a planar bounded region W, and spatio-temporal point processes generate events in a three dimensional volume $W \times T$ defined by a by a planar region and a temporal interval. The first-order intensity function of spatial and spatio-temporal point processes, which can be interpreted as the expected number of events per unit area or volume, characterizes the distribution of events in the observation domain. For this reason, estimating this function is one of the main issues in the analysis of any observed point pattern. Taking into account that parametric inference can provide unreliable estimates when the assumed models deviates form the true intensity, [1] introduced the kernel estimator of the density of event locations for point processes in R, which extension to the spatial and spatio-temporal frameworks is trivial. However the lack of consistence of this estimator has reduced its use to the exploratory analysis, in contrast with the wide number of inference techniques based on kernel smoothing developed in other areas of Statistics. In order to overcome this problem [2] defined the density of event locations and proved the consistency of its kernel estimator. In addition [3] and [4] introduced consistent kernel intensity estimators based on covariates. Estimating the joint distribution of spatial locations and times of occurrence of a spatio-temporal point process is challenging, and testing whether the spatio-temporal intensity function is separable a main issue. This report provide some advances on nonparametric inferences for the spatial and spatio-temporal first-order intensity function, addressing the two issues introduced above.

## 2. Inference for spatial point processes

2.1. **Consistent nonparametric estimators of the first-order intensity function.** Bandwidth selection, which is crucial in kernel smoothing has received little attention in the point process framework. In fact, although spatial point processes arise in the plane, scalar bandwidths have been used to estimate the first-order intensity function. Following the ideas of bivariate kernel density estimation [5] analyzed the consistency of the kernel estimator of the density of event locations with full matrix bandwidths. This work extended the smooth bootstrap proposed by [6] to estimate the MISE of the kernel density of event locations and developed a plug-in bandwidth selector based on the minimization of the bootstrap AMISE. We have also analyzed the performance of the kernel intensity estimators pro- posed by [3] and [4], which assume that the first-order intensity depends on some continuous covariates.

2.2. **Nonparametric comparison of first-order intensity functions.** Taking advantage of the relationship between the density of event locations of spatial point processes and the density of bivariate distributions, we extended to the point process framework the nonparametric test for comparison of multivariate data introduced by [7]. This test compares the density of event locations of two spatial point processes in order to test whether they have the same spatial distribution. We proposed a bootstrap algorithm to calibrate the null distribution of the test statistic and discussed the bandwidth selectors need to implement the test.

## 3. Inference for spatio-temporal point processes

The kernel estimator of the first-order-intensity and the nonparametric test allow us to characterize the spatial distribution of the wildfires registered in Galicia in the decade 1999-2008, compare the spatial distributions between causes, sizes or testing whether the spatial distribution of wildfire risk varies over years. However, as we also know the time of occurrence of each ignition point, we should use spatio-temporal point process analysis to a better understanding of wildfire behavior.

3.1. **Separability test for spatio-temporal point processes.** Modeling the joint distribution of spatial locations and times of occurrence in spatio-temporal point processes can be a difficult task, which complexity increases when the point process is marked or depends on covariates. For this reason, most of the current models assume that the spatio-temporal intensity function is separable, i.e. it can be expressed as the product of its spatial and temporal components. However, this assumption can be very restrictive and unrealistic in practice. This situation has motivated the development of nonparametric separability tests based on Monte Carlo simulations of the separable model [8, 9, 10, 11]. In a separable spatio-temporal point process the risk of observing an event at time t is spatially invariant, i.e. the ratio between the spatio-temporal and spatial intensity functions does not depend on the spatial locations. Taking into account this property we propose using a no effect test that checks whether the log-ratio function depends on the spatial locations in order to provide a new separability test. To implement the

test, we have developed a kernel estimator of the log-ratio function, using the similarity between spatio-temporal relative risk function and the ratio between the spatio-temporal and the spatial intensity functions, and a least-squares cross-validation bandwidth selection procedure for this kernel estimator.

## 4. Acknowledgements

We acknowledge the support of Project MTM2008-0310 from the Spanish Ministry of Science and Innovation, Projects MTM2010-14961, MTM2013-41383-P and MTM2013-43917-P from the Spanish Ministry of Economy and Competitivity, and grant P1-1B2012-52 from Bancaja Foundation.

## References

[1] P. Diggle, *A kernel method for smoothing point process data*, Journal of the Royal Statistical Society. Series C (Applied Statistics) **34**(2) (1985), 138–147.
[2] L. Cucala, *Espacements bidimensionnels et données entachées d'erreurs dans l'analyse des processus ponctuels spatiaux*, PhD thesis, Université des Sciences Sociales, Toulouse I (2006).
[3] Y. Guan, *On consistent nonparametric intensity estimation for inhomogeneous spatial point processes*, Journal of the American Statistical Association, **103**(483) (2008), 1238–1247.
[4] A. Baddeley, Y.M Chang, Y. Song, and R. Turner, *Nonparametric estimation of the dependence of a spatial point process on spatial covariates*, Stat Interface, **5**(2) (2012), 221–236.
[5] I. Fuentes-Santos, W. González-Manteiga, and J. Mateu, *Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial Poisson point processes*, Scandinavian Journal of Statistics (2015), DOI: 10.1111/sjos.12183.
[6] A. Cowling, P. Hall, and M.J. Phillips, *Bootstrap confidence regions for the intensity of a Poisson point process*, Journal of the American Statistical Association, **91**(436) (1996), 1516-1524.
[7] T. Duong, B. Goud, and K. Schauer, *Closed-form density-based framework for automatic detection of cellular morphology changes*, Proceedings of the National Academy of Sciences of the United States of America, **109**(22) (2012), 8382-8387.
[8] F.P. Schoenberg, *Testing separability in spatial-temporal marked point processes*, Biometrics, **60**(2) (2004), 471–481.
[9] R. Assuncáo and A. Maia, *A note on testing separability in spatial-temporal marked point processes*, Biometrics, **63**(1) (2007), 290–294.
[10] C.H. Chang and F.P. Schoenberg, *Testing separability in marked multidimensional point processes with covariates*, Annals of the Institute of Statistical Mathematics, **63**(6) (2011), 1103–1122.
[11] C. Díaz-Avalos, P. Juan, and J. Mateu, *Similarity measures of conditional intensity functions to test separability in multidimensional point processes*, Stochastic Environmental Research and Risk Assessment, **27**(5) (2013), 1193–1205.

## Testing the heteroscedastic error structure in quantile varying coefficient models

Anneleen Verhasselt

(joint work with Irène Gijbels and Mohammed Ibrahim)

We consider quantile regression in varying coefficient models.

Regression quantiles generalize naturally mean regression for Gaussian linear models, while substantially out-performing the least-squares estimator over a wide

class of non-Gaussian error distributions. They allow a wide range of applications, where several conditional quantiles are of interest instead of just the conditional mean. Quantile regression as a generalization of median regression was introduced by [4] in a linear regression setting.

Studying a fully nonparametric relationship between a response variable $Y$ and several covariates $(T, X_1, \ldots, X_p)$ is very difficult. Therefore, (to avoid the curse of dimensionality) varying coefficient models, introduced by [2], are considered. Varying coefficient models are an extension of classical linear regression models that allow the coefficients to depend on other variables. Assuming longitudinal data, the regression coefficients are allowed to vary with time.

In particular, we consider varying coefficient models with various structures for the variance of the errors (the variability function) in order to allow for heteroscedasticity:

$$Y(T) = \beta_0(T) + \beta_1(T)X^{(1)}(T) + ... + \beta_p(T)X^{(p)}(T) + V(\mathbf{X}(T), T)\epsilon(T)$$

where $\mathbf{X}(T) = (X^{(1)}(T), \ldots, X^{(p)}(T))$. We call $V(\mathbf{X}(T), T)$ the variability function. The longitudinal observations are $(Y(t_{ij}), X^{(1)}(t_{ij}), \ldots, X^{(p)}(t_{ij}), t_{ij})$ of $(Y(T), X^{(1)}(T), \ldots, X^{(p)}(T), T)$ and errors $\epsilon(t_{ij})$ of $\epsilon(T)$ for $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, where $t_{ij}$ is the $j$th measurement time for the $i$th subject, $N_i$ is the number of repeated measurements for the $i$th subject, $Y(t_{ij})$ and $(X^{(1)}(t_{ij}), \ldots, X^{(p)}(t_{ij}))$ are the observed outcome and covariates of the $i$th subject at time point $t_{ij}$.

The coefficient functions and the variability function are estimated with P-splines (as in [1]). The considered structures for the variability functions in [1] were linear in the parameters/covariates. However, we consider more general structures of the variability function, like power and exponential functions. Inspired by [3], we also provide a likelihood-ratio-type testing procedure to choose between two variability functions.

## References


[1] Y. Andriyana, I. Gijbels and A. Verhasselt, *P-splines quantile regression estimation in varying coefficient models*, Test **23** (2014), 153–194.

[2] T. Hastie and R. Tibshirani, *Varying-coefficient models*, Journal of the Royal Statistical Society. Series B **55** (1993), 757–796.

[3] M.-O. Kim, *Quantile regression with varying coefficients*, The Annals of Statistics **35** (2007), 92–108.

[4] R. Koenker and Jr. G. Bassett, *Regression quantiles*, Econometrica **46** (1978), 33–50.


## A High-dimensional Focused Information Criterion


Thomas Gueuning

(joint work with Gerda Claeskens)


In multivariate statistics, most of the variable selection methods (such as the AIC, the BIC and Mallow's $C_p$) aim at selecting one best model that is used to estimate all the quantities of interest related to the data. For instance, the same model is

often used for performing prediction on a number of new data points, for estimating the variance parameter and for estimating a quantile. Conversely, the focused information criterion (FIC) introduced by [1] selects the model that best estimates a particular quantity of interest (the focus) in terms of mean squared error (MSE). The FIC can select different models for different focuses and can produce estimators with small MSE. However, the current FIC literature is restricted to the low-dimensional case $p < n$. A first step towards a high-dimensional FIC has been made by [2] who developed a FIC for penalized estimators by smoothing the penalty function near zero, with the major restriction that the dimension of the parameter vector has to be fixed, so that the *small n - large p* case is asymptotically not covered. In the present work, we extend the focused information criterion to a diverging dimension of the parameter vector, covering the important $p > n$ case. This new FIC can thus be used for high-dimensional data. One condition is that we need the size of the true active set $s_n = o(n^{1/4})$. We now give some more details.

Let $Y_1, \ldots, Y_n$ be independent response variables with $Y_i$ having a density $f(y|X_i, \theta_0, \gamma_n)$, for $i = 1, \ldots, n$. The covariate vector $X_i$ is assumed to be of length larger than $n$. The parameter vector $\theta_0$ contains the protected variables, those we want to include in every possible model, and its dimension $p$ is not allowed to grow with $n$. The parameter vector $\gamma_n = \gamma_{0,n} + \delta_n/\sqrt{n}$ is of dimension $q_n$ growing with $n$ and contains parameters on which we want to perform variable selection. A simple example is the linear model $Y_i = \beta_0 + X_i\beta + \sigma\epsilon_i$ with $\epsilon_i \sim N(0,1)$ for which a natural choice is to take $\theta_0 = (\sigma, \beta_0)$ and $\gamma_n = \beta$. Variable selection is performed on $\gamma_n$ only.

Following the FIC philosophy, we are interested in estimating a particular quantity of interest $\mu_{\text{true}} = \mu(\theta_0, \gamma_n)$ called the focus. For a subset $S$ of $\{1, \ldots, q_n\}$, we consider an estimator $(\hat{\theta}_S, \hat{\gamma}_S)$ of the parameters $(\theta_0, \gamma_{n,S})$ and its corresponding estimator $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,n,S^c})$ of the focus. The idea is to estimate the limiting distribution $\Lambda_S$ of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ and to define the focused information criterion as $\text{FIC}(S) = \widehat{\text{E}[\Lambda_S]^2} + \widehat{\text{Var}[\Lambda_S]}$.

We distinguish two cases: (i) the case where the considered submodel is of low-dimension and (ii) the case where it is of high-dimension. In the former case, we use a maximum likelihood estimator in the submodel and we obtain an alternative low-dimensional FIC formula that can directly be applied. In the latter case we use a desparsified estimator that allows us to derive the MSE of the focus estimator.

Let us first consider the former case for which the maximum likelihood estimator (MLE) is available. We consider

$$(\hat{\theta}_S, \hat{\gamma}_S) = \text{argmax}_{(\theta,\gamma_S)\in\mathbb{R}^{p+|S|}} \sum_{i=1}^{n} \log\ f(Y_i|X_i, \theta, \gamma_S, \gamma_{0,n,S^c}) \tag{1}$$

and

$$\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,n,S^c}). \tag{2}$$

Then under mild regularity conditions, we have the following results

**Theorem 1.** *Assume that* $s_n = o(n^{1/4})$ *with* $s_n = |\{j : \delta_{n,j} \neq 0\}|$. *Consider* $S \subset \{1, \ldots, q_n\}$ *with* $p + |S| \leqslant n$. *For the maximum likelihood estimator (1) and the estimator (2) of the focus in model* $S$, *it holds that*

$$\sqrt{n}(\hat{\mu}_S - \mu_{true}) \doteq_d \Lambda_S$$

*with*

$$\Lambda_S = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^t \left( B_S \delta + \pi_S^{*t} J_S^{-1} \begin{pmatrix} U \\ V_S \end{pmatrix} \right)$$

*where the partial derivatives are evaluated at the null point* $(\theta_0, \gamma_0)$ *and where* $J_S$ *is the Fisher matrix in the submodel* $S$, $\begin{pmatrix} U \\ V_S \end{pmatrix} \sim N_{p+|S|}(0, J_S)$, $B_S = \pi_S^{*t} J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} - \begin{pmatrix} 0_{p \times q_n} \\ I_{q_n} \end{pmatrix}$ *and where* $\pi_S^*$ *is the* $(p+|S|) \times (p+q_n)$ *projection matrix corresponding to the submodel* $S$.

This leads to the following expression for the limiting mean squared error of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$:

$$\text{MSE}(S) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^t \left( B_S \delta \delta^T B_S^t + \pi_S^{*t} J_S^{-1} \pi_S^* \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix} \tag{3}$$

and we define $\text{FIC}(S) = \widehat{\text{MSE}(S)}$ as the FIC in the high-dimensional setting for a low-dimensional submodel. Note that if $p+q_n < n$, the formula (3) is equivalent to the one introduced in [1]. The equivalence can be shown using (with the notations of [3]) $Q^{-1} = J_{11} - J_{10} J_{00}^{-1} J_{01}$ and $G_S Q_S G_S^t = \pi_S^t Q_S \pi_S$.

Let us now consider a high-dimensional submodel $S$ such that $p + |S| > n$. In that case the maximum likelihood estimator is not available and we propose to use the desparsified estimator studied by [4]. From now, we restrict the developments to the linear model $Y_i = X_{\beta,i}^t \beta_0 + X_{\gamma,i}^t \gamma_n + \sigma \epsilon_i$ but extensions to generalized linear models are possible.

We consider the desparsified estimator

$$\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} + M_S \frac{1}{n\sigma^2} X_S^{*t} \left( Y - X_S^* \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} \right) \tag{4}$$

where $(\hat{\beta}_S^{\text{Lasso}}, \hat{\gamma}_S^{\text{Lasso}})$ is the Lasso estimator in the submodel $S$ and $M_S$ is a relaxed inverse of $J_S$ obtained by the nodewise regression technique, and we obtain the corresponding estimator $\hat{\mu}_S = \mu(\hat{\beta}_S^{\text{desp}}, \hat{\gamma}_S^{\text{desp}}, \gamma_{0,n,S^c})$ of the focus. Algebraic manipulations give

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{pmatrix} &= \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} + M_S \frac{1}{n\sigma^2} X_S^{*t} X_{\gamma,S^c} \delta_{S^c} + M_S \frac{1}{\sqrt{n}\sigma^2} X_S^{*t} \epsilon \\ &\quad - \sqrt{n} \left( I_{p+|S|} - M_S J_S \right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} - \beta_0 \\ \hat{\gamma}_S^{\text{Lasso}} - \frac{\delta_S}{\sqrt{n}} \end{pmatrix}. \end{aligned} \tag{5}$$

It can be shown that the last term of (5) is negligible if $S$ contains the true active set. For a general submodel $S$, we propose to use the approximations $\mathrm{E}\begin{bmatrix}\sqrt{n}(\hat{\beta}_S^{\mathrm{desp}}-\beta_0)\\ \sqrt{n}\hat{\gamma}_S^{\mathrm{desp}}\end{bmatrix} \approx \begin{pmatrix}0_p\\ \delta_S\end{pmatrix} + M_S\frac{1}{n\sigma^2}X_S^{*t}X_{\gamma,S^c}\delta_{S^c}$ and $\mathrm{Var}\begin{bmatrix}\sqrt{n}(\hat{\beta}_S^{\mathrm{desp}}-\beta_0)\\ \sqrt{n}\hat{\gamma}_S^{\mathrm{desp}}\end{bmatrix} \approx M_S J_S M_S^t$ which lead to the following definition of a high-dimensional FIC:

$$\mathrm{FIC}(S) = \widehat{\mathrm{MSE}(S)}$$

with

$$\mathrm{MSE}(S) = \begin{pmatrix}\frac{\partial\mu}{\partial\theta}\\ \frac{\partial\mu}{\partial\gamma}\end{pmatrix}^t \left(B_S'\delta\delta^T B_S'^t + \pi_S^{*t}M_S J_S M_S^t \pi_S^*\right)\begin{pmatrix}\frac{\partial\mu}{\partial\theta}\\ \frac{\partial\mu}{\partial\gamma}\end{pmatrix}$$

and

$$B_S' = \left(\pi_S^{*t}M_S\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix} - \begin{pmatrix}0_{p\times q_n}\\ I_{q_n}\end{pmatrix}\right)\left(I_q - \pi_S^t\pi_S\right).$$

This formula corresponds to (3) if $M_S = J_S^{-1}$. Numerical studies show satisfactory results.

## References


[1] G. Claeskens and N. L. Hjort, *The focused information criterion.* Journal of the American Statistical Association, **98** (2003), 900–916.

[2] G. Claeskens, *Focused estimation and model averaging with penalization methods: an overview*, Statistica Neerlandica, **66**(3) (2012), 272–287.

[3] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, New York, 2008.

[4] S. van de Geer, P. Bühlmann, Y. Ritov and R. Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics, **42**(3) (2014), 1166–1202.


## When are principal component scores a good tool for visualizing high-dimensional data?


Kristoffer H. Hellton

(joint work with Magne Thoresen)


Principal component analysis (PCA) is a popular method for visualizing and reducing the dimension of high-dimensional data. In particular in genetic applications, the initial data exploration is often done by visually investigating the first principal component (PC) scores.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be a $p \times n$ data matrix, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ are independent and identically distributed with $\mathrm{E}\,\mathbf{x}_i = \mathbf{0}$ and $\mathrm{var}\,\mathbf{x}_i = \Sigma$. Then the eigenvectors, $\mathbf{v}_1, \ldots, \mathbf{v}_p$, of the population covariance matrix $\Sigma$ define the population PCs as

$$\mathbf{s}_j^T = \mathbf{v}_j^T\mathbf{X} = [\mathbf{v}_j^T\mathbf{x}_1, \ldots, \mathbf{v}_j^T\mathbf{x}_n],$$

such that the PCs are linear and orthogonal combinations of variables expressing maximal variability. As the variance of each PC is given by the corresponding

eigenvalue, $\lambda_1 \geq \cdots \geq \lambda_p$, the standardized PC is defined as $\mathbf{z}_j^T = \mathbf{v}_j^T \mathbf{X}/\sqrt{\lambda_j}$. The sample PCs are based on the eigenvalues and -vectors of the sample covariance matrix $\hat{\Sigma}$, denoted by $d_1, \ldots, d_p$ and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p$, such that the sample scores and standardized sample scores are given

$$\hat{\mathbf{s}}_j^T = \hat{\mathbf{v}}_j^T \mathbf{X}, \qquad \hat{\mathbf{z}}_j^T = \frac{\hat{\mathbf{v}}_j^T \mathbf{X}}{\sqrt{d_j}}.$$

In the classical asymptotic setting ($p$ is fixed and $n \to \infty$) the sample eigenvalues and eigenvectors will be consistent estimators for the population eigenvalues and eigenvectors. In the high-dimensional case, however, Paul [1] established within the random matrix framework where $p/n \to \gamma > 0$ as $p, n \to \infty$, that the sample eigenvalues and eigenvectors are in fact inconsistent. The same holds within the High Dimension Low Sample Size (HDLSS) framework ([2, 3]), where $n$ is fixed and the $m$ first eigenvalues scale with the dimension, $\lambda_i = \sigma_i^2 p^\alpha$ as $p \to \infty$. Then for the special case of $\alpha = 1$, it is possible to show that relative positions of the PC scores, and thereby the visual information, can be consistent even if the eigenvectors are not [4]. This offers an explanation for the paradoxical situation where classical PCA works successfully in practice, despite being theoretically inconsistent. The key assumption of $\alpha = 1$ can be interpreted in terms of the generating mechanism behind the data.

Let $\mathbf{x}_i \sim N(0, \Sigma)$ be *i.i.d.* and let the $m$ first eigenvalues scale with $p$

$$\lambda_1 = \sigma_1^2 p, \quad \ldots, \quad \lambda_m = \sigma_m^2 p,$$

while the remaining eigenvalues are constant $\lambda_m = \cdots = \lambda_p = \tau^2$. Then the vector of the $m$ first sample scores of the $i$th observation converges in distribution, as $p \to \infty$, to

$$\begin{bmatrix} \hat{z}_{i1} \\ \vdots \\ \hat{z}_{im} \end{bmatrix} \xrightarrow{d} \underbrace{\begin{bmatrix} \sqrt{n/d_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{n/d_m} \end{bmatrix}}_{Scaling} \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_m \\ | & & | \end{bmatrix}^T}_{Rotation} \begin{bmatrix} \sigma_1 z_{i1} \\ \vdots \\ \sigma_m z_{im} \end{bmatrix},$$

for $i = 1, \ldots, n$, where $d_j$ and $\mathbf{u}_j$ are the $j$th eigenvalue and -vector of an $m \times m$ Wishart distributed matrix, $W \sim Wishart\left(n, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)\right)$. The consequence is that the vector of the first $m$ sample scores will be a *rotated* and *scaled* version of the population scores in $m$ dimensions, and the relative positions of the scores will remain the same.

For the purpose of visualizing data one would plot pairs of the $m$ first sample scores in two dimensions. Simulations based on normally distributed data show that for moderate sample sizes, a two-dimensional representation of the sample scores will also be a scaled and approximately rotated version of the population scores. Thus the visual information of the population scores is preserved by the sample scores. The behavior is demonstrated in Figure 1 showing two realizations of the sample scores compared to the population scores.

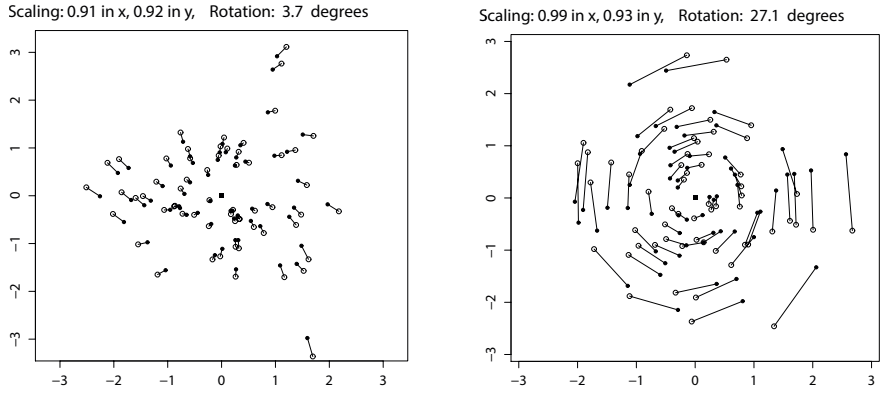


FIGURE 1. Plots of sample PC scores (black dots) compared to population PC scores (circles) showing scaling and rotation.

The assumption of linearly increasing eigenvalues can be motivated by viewing the principal components as latent factors. If a factor is pervasive, affecting most or a significant proportion of the observed variables, the corresponding population eigenvalue will scale linearly with the dimension asymptotically. Some applications where pervasive effects are reasonable include genetics markers (SNPs), where different populations and ethnicity act as latent factors, and stock returns with supply and demand shocks.

## REFERENCES

[1] D. Paul *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model*. Statistica Sinica **17**(4) (2007), 1617–1642.
[2] S. Jung and J.S. Marron, *PCA consistency in high dimension, low sample size context*, Annals of Statistics **37** (2009), 4104–4130.
[3] S. Jung, A. Sen and J.S. Marron *Boundary behavior in high dimension, low sample size asymptotics of PCA*. Journal of Multivariate Analysis **109**(3) (2012), 190–203.
[4] K.H. Hellton and M. Thoresen *Asymptotic distribution of principal component scores for pervasive, high-dimensional eigenvectors*. arXiv preprint arXiv:1401.2781 (2014).

## High-dimensional mixture regression models, application to functional data

EMILIE DEVIJVER

Owing to the increasing availability of high-dimensional datasets, regression models for multivariate response and high-dimensional predictors have become important tools.

In this extended abstract, we describe two procedures where a random target variable $\boldsymbol{Y} \in \mathbb{R}^q$ depends on explanatory variables within a cluster-specific regression model. Each cluster is represented by a parametric distribution, the entire dataset being modeled by a mixture of these distributions. The model assumes that each observation $i \in \{1, \ldots, n\}$ originates from one of $K$ disjoint classes and

that the data $(\boldsymbol{Y}_i, \boldsymbol{x}_i) \in \mathbb{R}^q \times \mathbb{R}^p$ are independent and identically distributed such that if $i$ belongs to class $k \in \{1, \ldots, K\}$, the target variable $\boldsymbol{Y}_i$ results from a regression model

$$\boldsymbol{Y}_i = \mathbf{B}_k \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

with an unknown matrix of coefficients $\mathbf{B}_k \in \mathbb{R}^{q \times p}$ and independent errors $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_q(0, \Sigma_k)$ with an unknown diagonal covariance matrix $\Sigma_k$ of size $q \times q$. Each observation $i \in \{1, \ldots, n\}$ has a probability $\pi_k$ to belong to the cluster $k \in \{1, \ldots, K\}$.

We work with high-dimensional data, in other words the number of parameters to estimate $K(qp + q + 1) - 1$ is larger than the number of observed target values $q \times n$. Two ways are considered in this paper, coefficients sparsity and ranks sparsity. The first approach consists in estimating the matrix $\mathbf{B}_k$ by a matrix with few nonzero coefficients. The well-known Lasso estimator, introduced by [7] for linear models, is the solution chosen here. In the second approach, we consider the rank sparsity in $\mathbf{B}_k$. This idea dates back to the 1950's and was initiated by [1] for the linear model. [6] introduced the term of reduced-rank regression for this class of models. For more recent works, we refer to [5] and to [3]. Nevertheless, the linear model is appropriate for homogeneous observations, which is a strong assumption. We extend here those methods to mixture regression models.

We propose here two procedures for clustering high-dimensional or functional data, where the high-dimensional or functional random target variable $\boldsymbol{Y} \in \mathbb{R}^q$ depends on high-dimensional or functional predictors $\boldsymbol{x} \in \mathbb{R}^p$ with a cluster-specific regression model.

We consider a finite mixture of Gaussian regression models. The two procedures we propose follow the same steps. Firstly, a penalized likelihood approach is considered to determine potential sets of relevant indices. Varying the regularization parameter, a data-driven collection of models is constructed where each model has a reasonable complexity. The second step of the procedures consists in refitting parameters by a less biased estimator, restricting the model on selected indices. Then, we select a model among the collection using the slope heuristic, which was introduced by [2]. The difference between the two procedures is in the refitting step. In the first procedure, later called Lasso-MLE procedure, the maximum likelihood estimator is used. The second procedure, called Lasso-Rank procedure, deals with low rank estimation. For each model in the collection, a subcollection of models with means estimated by various low rank matrices is constructed. It leads to sparsity for the coefficients and for the rank, and it considers the mean within its matrix structure.

## References


[1] T. W. Anderson. *Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions*, The Annals of Mathematical Statistics, **22**(3) (1951), 327–351.

[2] L. Birgé and P. Massart. *Minimal penalties for Gaussian model selection*, Probab. Theory Related Fields, **138**(1-2) (2007).

[3] F. Bunea, Y. She and M. Wegkamp. *Joint variable and rank selection for parsimonious estimation of high-dimensional matrices*, The Annals of Statistics, **40**(5) (2012), 2359–2388.

[4] F. Ferraty and P. Vieu. *Nonparametric functional data analysis : theory and practice*. Springer series in statistics. Springer, New York, 2006.
[5] C. Giraud. *Low rank multivariate regression*, Electronic Journal of Statistics, **5** (2011), 775–799.
[6] A. Izenman. *Reduced-rank regression for the multivariate linear model*, Journal of Multivariate Analysis, **5**(2) (1975), 248–264.
[7] R. Tibshirani. *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B., **58**(1) (1996), 267–288.

## Functional Data Depth

Stanislav Nagy
(joint work with Irène Gijbels, Daniel Hlubinka and Marek Omelka)

Data depth is a concept proposed in nonparametric statistics as a generalization of quantiles for multivariate data. Formally, for an arbitrary sample space $S$, depth is a mapping which to an object $s \in S$ and $P \in \mathcal{P}(S)$, $\mathcal{P}(S)$ standing for the set of all probability distributions on $S$, assigns a number

$$D\colon S \times \mathcal{P}(S) \to [0,1]\colon\ (s,P) \mapsto D(s;P)$$

characterizing how "central" $s$ is with respect to $P$. Higher values of $D$ indicate centrality, values close to zero can be interpreted as potential outlyingness of $s$. Using depth we obtain an ordering of the sample points relative to $P$ in the centre-outward sense. It allows construction of multivariate L-estimators, nonparametric testing procedures, or simple visualisation tools for high-dimensional observations.

The first depth in $S = \mathbb{R}^d$, $d \geq 1$ was considered by Tukey [8] who defined the *halfspace depth* of $x \in \mathbb{R}^d$ with respect to $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ as

$$D_H(x;P) = \inf_{H \in \mathcal{H}(x)} P(H), \tag{1}$$

for $\mathcal{H}(x)$ the set of all closed halfspaces in $\mathbb{R}^d$ containing $x$. For a general treatment of depth functions in $\mathbb{R}^d$ see Zuo and Serfling [9].

The potential of the introduction of depth to functional (and general infinite-dimensional) data was first explored by Fraiman and Muniz [2]. In that contribution, the authors consider $\mathcal{C}$ — the space of continuous functions over $[0,1]$, and propose a simple depth of $x \in \mathcal{C}$ with respect to $P \in \mathcal{P}(\mathcal{C})$ taking the form

$$D_{FM}(x;P) = \int_0^1 D_H(x(t);P_t)\,. \tag{2}$$

Here, $P_t \in \mathcal{P}(\mathbb{R})$ is the marginal distribution of $P$ corresponding to $t \in [0,1]$, and $D_H$ is the depth (1) for $d = 1$. The resulting *integrated depth* for functions is later elaborated further by, among others, Cuevas and Fraiman [1].

Another major approach towards the computation of depth in $\mathcal{C}$ arises when one replaces the integral in (2) by an infimum. This way we obtain the *infimal depth* described by Mosler [5]

$$D_M(x;P) = \inf_{t \in [0,1]} D_H(x(t);P_t)\,. \tag{3}$$

Many other depths for functional data can be found in the literature. Nevertheless, the vast majority of them can be represented either as an integral, or an infimum, of a set of finite-dimensional depths of certain low-dimensional projections of $x$ and $P$, just as in (2) and (3). For instance, the important band depth [4] follows the general idea of depths of infimal type (3), whereas its modified counterpart is of integrated type (2).

In the present contribution we deal with theoretical properties of both integrated and infimal depths for functions, with the main emphasis placed on their sample version consistency. In particular, we are interested in conditions under which the uniform consistency

$$\lim_{n\to\infty} \sup_{x\in\mathcal{C}} |D(x; P_n) - D(x; P)| = 0 \qquad \text{almost surely} \tag{4}$$

holds true for a depth $D$, where $P_n \in \mathcal{P}(\mathcal{C})$ is the empirical measure of a random sample of $n$ functions from $P$.

For the depth $D_M$ from (3) and, by extension, for all infimal depths we demonstrate that (4) cannot be true for these for all $P$. Such observations allow us to identify the difficulties with the family of infimal depths for random curves, and eventually also to derive sufficient conditions on $P$ for (4) to hold for them. For a detailed exposition, counterexamples, and extensions see Gijbels and Nagy [3].

For integrated depths we first resolve the crucial technical problem of the existence of the integral in the expression (2), and some related quantities. This is achieved by rigorous verification of the measurability of the corresponding integrand functions — a nontrivial task that has been widely ignored in the literature on the topic. These basic results enable the formulation of an elementary proof of universal weak, and later also universal strong consistency (meaning that (4) is true for any $P \in \mathcal{P}(\mathcal{C})$) of all integrated depths for functions. For this we require only mild conditions to be imposed on the finite-dimensional depth used in the definition of integrated depths (known to be satisfied for $D_H$). For the proofs, and a comprehensive study of all integrated depths see Nagy et al. [7].

Having resolved the consistency issues connected with many depths for functions, we focus on the case when the random sample curves $X_1, \ldots, X_n$ from $P$ are not observed continuously over their domain $[0, 1]$, but rather only discretely. This means that for the function $X_i$, $i = 1, \ldots, n$, the experimenter cannot observe all its functional values $\{X_i(t) \colon t \in [0, 1]\}$, but knows only a finite number $m_i \in \mathbb{N}$ of these $\{X_i(T_{j,i}) : T_{j,i} \in [0, 1], j = 1, \ldots, m_i\}$. The observation points $T_{j,i}$ and their number $m_i$ may be fixed, or random. In this setup, we propose to reconstruct the unobservable curves based on the points that one can observe using a simple interpolation technique. The interpolated curves are then plugged into the depth computation. This approach is followed in Nagy et al. [6], where under suitable conditions it is shown that the consistency results described above remain in order also for such discretely observed functional data.

## References

[1] A. Cuevas and R. Fraiman. *On depth measures and dual statistics. A methodology for dealing with general data*, Journal of Multivariate Analysis, **100**(4) (2009), 753–766.
[2] R. Fraiman and G. Muniz. *Trimmed means for functional data*, Test, **10**(2) (2001), 419–440.
[3] I. Gijbels and S. Nagy. *Consistency of non-integrated depths for functional data*, Journal of Multivariate Analysis, **140** (2015), 259–282.
[4] S. López-Pintado and J. Romo. *On the concept of depth for functional data*, Journal of the American Statistical Association, **104**(486) (2009), 718–734.
[5] K. Mosler. *Depth statistics*, In Robustness and complex data structures, pages 17–34. Springer, Heidelberg (2013).
[6] S. Nagy, I. Gijbels, and D. Hlubinka. *Weak convergence of discretely observed functional data with applications*, Journal of Multivariate Analysis (2015), To Appear.
[7] S. Nagy, I. Gijbels, M. Omelka, and D. Hlubinka. *Integrated depth for functional data: Statistical properties and consistency*, ESAIM Probab. Stat. (2016) To Appear.
[8] J. W. Tukey. *Mathematics and the picturing of data.* Proc. Int. Congr. Math., Vancouver 1974, **2** (1975), 523–531.
[9] Y. Zuo and R. Serfling. *General notions of statistical depth function*, The Annals of Statistics, **28**(2) (2000), 461–482.


## Factorisable Sparse Tail Event Curves with Expectiles


Wolfgang K. Härdle

(joint work with Chen Huang and Shih-Kang Chao)


Data are observed more and more in form of curves, thus prompting a joint modelling to find out common patterns and also individual variations. Real data curve modelling occurs e.g. in neuroeconomics, wind speed analysis, demographics among many other disciplines.

Functional data analysis studies variation of random objects in a high dimensional contact and provides insight into main factors, typically extracted as principal components via a Karhunen-Loève decomposition. However, in a variety of applications one is more interested in the tail behavior rather than the variations around the mean. Thus the analysis of curve variation is around a tail event curve (TEC) rather than around a mean curve as in functional PCA. TECs may be identified through tail probabilities or more general through functions based on conditional tail events. Modeling such Tail Event Curves (TEC) requires to deviate from Hubert $\ell_2$ geometry and to introduce asymmetric norms or check functions. For example, quantile regression is a widely used method can be exploited to grasp the whole information on the conditional distribution and especially the tail structure, which plays crucial roles in risk management. Concerning multivariate quantile regression, many previous works study in this direction under different frameworks. But none of them worked in high-dimensional case. [2] introduced factorisable sparse tail event curves (FASTEC) method to implement high-dimensional multivariate quantile regression.

fMRI risk perception analysis requires to study the shape (e.g., amplitude, delay, and duration) of the estimated hemodynamic response function (HRF) to particular tasks answered by every individual. More noteworthily, extreme behaviors

of the response function may reveal unobserved neuronal activation information. Therefore, we need a global measure which can capture the tail moments and be more sensitive to the outliers. Expectiles can be a better choice than quantiles in consideration of extremes although it is not robust. This fact motivates us to build an expectile based FASTEC model.

Denote $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{n\times m}$ as the multivariate curves we want to jointly model, where $n$ is the length of observations and $m$ is the number of curves. $\{\boldsymbol{X}_i\}_{i=1}^n \in \mathbb{R}^p$ are the covariates with dimension $p$, e.g., B-spline basis. Both $p$ and $m$ are allowed to tend to infinity with the sample size $n$ (but no quicker than $n$).

Let $e_j(\tau|\boldsymbol{X}_i)$ be the conditional expectile function of $Y_{ij}$ given $\boldsymbol{X}_i$, for $i = 1,\ldots,n$ and $j = 1,\ldots,m$ with $\tau \in (0,1]$, and approximate it by a linear factor model,

$$Y_{ij} = e_j(\tau|\boldsymbol{X}_i) + u_{ij}$$

$$e_j(\tau|\boldsymbol{X}_i) = \sum_{k=1}^{r} \psi_{j,k}(\tau) f_k^\tau(\boldsymbol{X}_i), \tag{1}$$

where $f_k^\tau(\boldsymbol{X}_i)$ is the $k$th factor, $r$ is the number of factors (much less than $p$), $\psi_{j,k}(\tau)$ are the factor loadings. Furthermore, factors are constructed by linear combination of covariates $\boldsymbol{X}_i$

$$f_k^\tau(\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\varphi}_k(\tau). \tag{2}$$

Substituting (2) into (1) yields

$$e_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\gamma}_j(\tau), \tag{3}$$

with $\boldsymbol{\gamma}_j(\tau) = \left(\sum_{k=1}^r \psi_{j,k}(\tau)\varphi_{k,1}(\tau), \ldots, \sum_{k=1}^r \psi_{j,k}(\tau)\varphi_{k,p}(\tau)\right)^\top$ as the unknown coefficient vector. In multivariate case, what needs to be estimated becomes a $p \times m$ coefficient matrix $\boldsymbol{\Gamma}$, where $\boldsymbol{\gamma}_j(\tau)$ in (3) is the $j$th column of $\boldsymbol{\Gamma}$.

With increasing dimension of both explanatory and response variables one faces the difficulty of estimating a very high dimensional coefficient matrix. A natural way to reduce the burden of this estimation task is to introduce a penalty term. [4] proposed a penalization approach with nuclear norm, the sum of the singular values of the coefficient matrix as the penalty. Numerically the estimator can be easily obtained since it involves a convex optimization. Moreover, compare with previous traditional works such as reduced rank approach, the number of factors does not need to be predetermined. Dimension reduction and coefficient estimation can be done simultaneously.

To be more precise it is proposed to estimate the coefficient matrix $\boldsymbol{\Gamma}$ by solving:

$$\widehat{\boldsymbol{\Gamma}}_\lambda(\tau) = \arg\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{p\times m}} F(\boldsymbol{\Gamma}), \tag{4}$$

$$F(\boldsymbol{\Gamma}) = (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau\left(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j}\right) + \lambda\|\boldsymbol{\Gamma}\|_*, \tag{5}$$

$$\rho_\tau(u) = |\tau - \mathbf{1}\{u < 0\}|\,|u|^2. \tag{6}$$

Nuclear norm $\|\mathbf{\Gamma}\|_*$ is defined by $\sum_{l=1}^{\min(p,m)} \sigma_l(\mathbf{\Gamma})$ given the singular values of $\mathbf{\Gamma}$: $\sigma_1(\mathbf{\Gamma}) \geq \sigma_2(\mathbf{\Gamma}) \geq \ldots \geq \sigma_{\min(p,m)}(\mathbf{\Gamma})$. The convexity of the nuclear norm results in a convex optimization problem that can be solved via various of efficient methods. The number of nonzero singular values of $\mathbf{\Gamma}$ is identified as $r$. A high dimension $p \times m$ is reduced to $r \times \max(p, m)$ by regularization, when $\mathbf{\Gamma}$ is sparse. After obtaining the $\widehat{\mathbf{\Gamma}}_\lambda(\tau)$ from (4), singular value decomposition (SVD) can be employed to estimate the factors and normalized factor loadings respectively.

Moreover, the loss function for expectile regression has a smooth convex function form. Combining with the nuclear norm penalty, we can use Fast Iterative Shrinkage-Thresholding Algorithm proposed by [1] to solve the optimization directly. Without smoothing the asymmetric absolute check function, the convergence rate in the iterative procedure is quicker than in quantile regression case. Based on the unified framework for high-dimensional $M$-estimators with decomposable regularizers provided by [3], the finite sample oracle properties of the estimator associated expectile loss and nuclear norm regularizer are studied formally in this paper.

As an empirical illustration, our model is applied on fMRI data to see if individual's risk perception can be recovered by brain activities. Results show that main factors can reflect the common patterns of curves. Factor loadings over different tail levels can help to find out the most risk-seeking and averse behaviours. Taking tail risks into consideration, individual's risk attitudes can be predicted more precisely, especially the extremes.

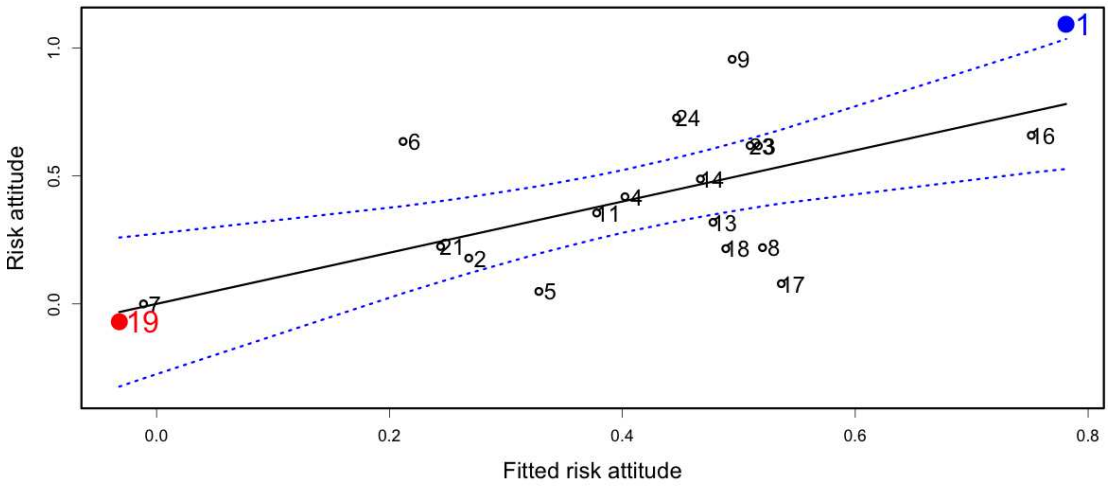


FIGURE 1. Horizontal axis denotes the fitted risk attitudes by the first factor loadings estimated from the brain data when $\tau = 0.1$, vertical axis denotes the risk attitudes parameters based on their choices. #1 and #19 are the most risk-averse and risk-seeking people respectively.

FASTEC_with_Expectiles

References

[1] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, Topology **2** (2009), 183–202.
[2] S.-K. Chao, W. K. Härdle and M. Yuan, *Factorisable sparse tail event curves*, SFB 649 Discussion Paper 2015-034 (2015).
[3] S. N. Negahban, P. Ravikumar, M. J. Wainwright and B. Yu, *A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers*, Statistical Science **27** (2012), 538–557.
[4] M. Yuan, A. Ekici, Z. Lu and R. Monteiro, *Dimension reduction and coefficient estimation in multivariate linear regression*, Journal of the Royal Statistical Society: Series B **69** (2007), 329–346.

## Statistical Blind Source Separation - with Applications in Cancer Genetics

Axel Munk
(joint work with Merle Behr and Chris Holmes)

**The SBSSR model**.
We are studying a particular kind of blind source separation problem embedded in a change-point regression setting. In blind source separation problems one observes a mixture of source functions, and aims to recover the original sources from the available observations. The blindness refers to the fact that neither the sources nor the mixing is known. We consider single linear mixtures of step functions with a known finite alphabet in a Statistical Blind Source Separation Regression (SBSSR) model.

More precisely, for a given finite alphabet $\mathfrak{A} \subset \mathbb{R}$, a given number of source components $m \geq 2$, unknown step functions $f = (f^1, \ldots, f^m)^\top$ each taking values in the known alphabet, $\mathrm{imag}(f^i) \subset \mathfrak{A}$, and unknown probability mixing weights $\omega = (\omega_1, \ldots, \omega_m)^\top \in \mathbb{R}_+^m$ with $\sum_{i=1}^m \omega_i = 1$, one observes from the mixture

$$Y_j = \omega^\top f(x_j) + \epsilon_j = \sum_{i=1}^m \omega_i f^i(x_j) + \epsilon_j, \quad j = 1, \ldots, n, \tag{1}$$

where $\epsilon$ is normal noise with mean zero. We assume that $\omega^\top f$ in (1) is sampled equidistantly at $x_j := (j-1)/n$, $j = 1, \ldots, n$ and that all step functions $f^i$ are defined on the domain $[0, 1)$. Extensions to more general domains $\subset \mathbb{R}$ and sampling designs are straightforward under suitable assumptions. The aim in model (1) is to estimate $\omega$ and $f$ from the observations $Y = (Y_1, \ldots, Y_n)$ and to construct honest confidence statements for all quantities.

**Identifiability**.
We stress that already in the noiseless case, i.e., $\epsilon \equiv 0$ in (1), it is far from obvious under which criteria the weights $\omega$ and the sources $f$ are identifiable. We characterize the identifiability issue as a combinatorial problem and derive simple sufficient and necessary identifiability criteria which, to the best of our knowledge, has been elusive. On the one hand, these conditions ensure discriminability of

different mixture values, which is a necessary conditions on $\omega$ to guarantee identifiability of $f$, and, on the other hand, they ensure a sufficient variability of $f$, which is necessary to guarantee identifiability of $\omega$.

Moreover, we discuss how likely it is for the derived identifiability criteria to be satisfied when the mixing weights are drawn from the uniform distribution and when the underlying sources are discrete Markov processes. We show that the mixture becomes identifiable exponentially fast, which reveals identifiability not to be an issue in most practical situations. See [1] for more details on the identifiability issue in model (1).

**The SESAME estimation methodology**.
In the regression setting we propose a new methodology, called SESAME (SEparateS finite Alphabet MixturEs), which yields uniform confidence sets and optimal estimation rates (up to log-factors) for all parameters in model (1) under very weak identifiability conditions [2].

First, SESAME provides honest confidence regions $\mathcal{C}_{1-\alpha}(Y)$ for the mixing weights $\omega$ which are characterized by the acceptance region of a certain multiscale test [3] with level $\alpha$. Then, it estimates $\hat{\omega} \in \mathcal{C}_{1-\alpha}(Y)$, where now the level $\alpha$ can be seen as a tuning parameter for which we propose data driven selection methods. For $\hat{\omega}$ and $\mathcal{C}_{1-\alpha}(Y)$ we obtain almost optimal estimation rates and diameter $\ln(n)/\sqrt{n}$, respectively. Second, SESAME estimates the sources $f$ as a constrained maximum likelihood estimator, where the constraint comes from the same multiscale statistic as for $\mathcal{C}_{1-\alpha}(Y)$ but with a possibly different level $\beta$. This yields asymptotically honest multivariate confidence bands $\mathcal{H}(\beta)$ for the sources $f$. For the resulting estimate $\hat{f}$ we derive exact recovery, i.e., the number of change-points of $f$ and its function values are estimated exactly and its change-point locations with the minimax rate $1/n$ up to a log square term with probability converging to one at a superpolynomial rate.

SESAME's estimates and confidence statements can be computed efficiently using dynamic programming and certain pruning steps.

**Applications**.
Model (1) arises in many different applications, for instance in digital communication with mixtures of multi-level PAM signals [4]. Our motivation, however, comes from an application in cancer genetics. We use SESAME to analyze genetic sequencing data in order to estimate clonal proportions in a tumor, and the corresponding copy number variations [5].

**Acknowledgement**.
This work has been initiated through intensive discussions at the MFO workshop 'Frontiers in Nonparametric Statistics' in 2012.

## References

[1] M. Behr and A. Munk. *Identifiability for blind source separation of multiple finite alphabet linear mixtures.* arXiv preprint. arXiv:1505.05272.

[2] M. Behr, C. Holmes, and A. Munk. *Multiscale blind source separation.* preprint 2015.

[3] K. Frick, A. Munk, and H. Sieling. *Multiscale change point inference.* Journal of the Royal Statistical Society: Series B (Statistical Methodology). (2014) **76** 495 – 580.
[4] J. G. Proakis. *Digital communications.* Communications and Signal Processing. (1995) McGraw-Hill
[5] B. Liu, C.D. Morrison, C.S. Johnson, D.L. Trump, M. Qin, J.C. Conroy, J. Wang, and S. Liu. *Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges.* Oncotarget. (2013) **4** 1868.

## Statistical inverse problems: Motivations and data-driven Bayesian estimation

Jan Johannes
(joint work with Anna Simoni and Ruldolf Schenk)

Statistical ill-posed inverse problems are becoming increasingly important in a diverse range of disciplines, including geophysics, astronomy, medicine and economics. Roughly speaking, in all of these applications the observable signal $g = Tf$ is a transformation of the functional parameter of interest $f$ under a linear operator $T$. Statistical inference on $f$ based on an estimation of $g$ which usually necessitates an inversion of $T$ is thus called an inverse problem. Moreover, by ill-posed we mean that the transformation $T$ is not stable, i.e., $T$ has not a continuous inverse. In this presentation special attention is given to three models and their extensions, namely non-parametric instrumental regression under endogeneity, functional linear regression and density deconvolution, each of them leading naturally to a statistical ill-posed inverse problem. In those applications, however, both the signal $g$ and the inherent transformation $T$ are not known in practise, although they can be estimated from the data. Consequently, a statistical inference has to take into account that a random noise is present in both the estimated signal and the estimated operator.

Typical questions in this context are the non-parametric estimation of the functional parameter $f$. It is well-known that in terms of its risk the attainable accuracy of an estimation procedure is essentially determined by the conditions imposed on $f$ and the operator $T$ which are, for example, expressed in the form $f \in \mathcal{F}$ and $T \in \mathcal{T}$ for suitable chosen classes $\mathcal{F}$ and $\mathcal{T}$. Minimax-optimality of an estimator is then usually shown by establishing both an upper and a lower bound of the maximal risk over given classes $\mathcal{F}$ and $\mathcal{T}$.

In many cases the proposed estimation procedures rely on the choice of at least one tuning parameter, which in turn, crucially influences the attainable accuracy of the constructed estimator. Its optimal choice, however, follows often from a classical squared-bias-variance trade-off and relies on an *a-priori* knowledge about the classes $\mathcal{F}$ and $\mathcal{G}$, which is usually inaccessible in practise. This motivates its data-driven choice in the context of non-parametric statistics since its very beginning in the fifties of the last century. Among the most prominent approaches stand without doubts model selection, Stein's unbiased risk estimation, Lepski's method or combinations of the aforementioned strategies.

On the other hand side, it seems natural to adopt a Bayesian point of view where the dimension parameter can be endowed with a prior. As the theory for a general inverse problem – with a possibly unknown or noisy operator – is technically highly involved, the focus of this presentation is on an indirect Gaussian sequence space model which is well known to be equivalent to an indirect Gaussian regression.

Let $\ell_2$ be the Hilbert space of square summable real valued sequences endowed with the usual inner product $\langle\cdot,\cdot\rangle_{\ell_2}$ and associated norm $\|\cdot\|_{\ell_2}$. In an indirect Gaussian sequence space model (iGSSM) the aim is to recover a parameter sequence $\theta = \big(\theta_j\big)_{j\geq 1} \in \ell_2$ from a transformed version $(\lambda_j\theta_j)_{j\geq 1}$ that is blurred by a Gaussian white noise. Precisely, an observable sequence of random variables $(\mathrm{Y})_{j\geq 1}$, Y for short, obeys an indirect Gaussian sequence space model, if

$$\mathrm{Y}_j = \lambda_j\theta_j + \sqrt{\epsilon}\xi_j, \qquad j \in \mathbb{N}, \tag{1}$$

where $\{\xi_j\}_{j\geq 1}$ are unobservable error terms, which are independent and standard normally distributed, and $0 < \epsilon < 1$ is a noise level known in advance. The sequence $\lambda = \big(\lambda_j\big)_{j\geq 1}$ represents the operator that transforms the signal $\theta$. In the particular case of a constant sequence $\lambda$ the sequence space model is called direct while it is called an indirect sequence space model if the sequence $\lambda$ tends to zero.

Adopting a Bayesian approach the parameter sequence of interest $\theta = (\theta_j)_{j\geq 1}$ itself is a realisation of a random variable $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_j)_{j\geq 1}$ and the observable random variable $\mathrm{Y} = (\mathrm{Y}_j)_{j\geq 1}$ satisfies

$$\mathrm{Y}_j = \lambda_j\,\boldsymbol{\vartheta}_j + \sqrt{\epsilon}\xi_j, \quad j \in \mathbb{N} \tag{2}$$

with independent and standard normally distributed error terms $\{\xi_j\}_{j\geq 1}$ and known noise level $0 < \epsilon < 1$. Moreover, we assume that random parameters $\{\boldsymbol{\vartheta}_j\}_{j\geq 1}$ and the error terms $\{\xi_j\}_{j\geq 1}$ are independent. Consequently, (2) and a specification of the prior distribution $P_{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$ determine completely the joint distribution of Y and $\boldsymbol{\vartheta}$.

In this presentation we consider a sieve prior family $\{P_{\boldsymbol{\vartheta}^m}\}_m$ where the prior distribution $P_{\boldsymbol{\vartheta}^m}$ of the random parameter sequence $\boldsymbol{\vartheta}^m = (\boldsymbol{\vartheta}_j^m)_{j\geq 1}$ is Gaussian and degenerated for all $j > m$. More precisely, the first $m$ coordinates $\{\boldsymbol{\vartheta}_j^m\}_{j=1}^m$ are independent and normally distributed random variables while the remaining coordinates $\{\boldsymbol{\vartheta}_j^m\}_{j>m}$ are degenerated at a point. Assuming an observation $\mathrm{Y} = (\mathrm{Y}_j)_{j\geq 1}$ satisfying $\mathrm{Y}_j = \lambda_j\,\boldsymbol{\vartheta}_j^m + \sqrt{\epsilon}\xi_j$, we denote by $P_{\boldsymbol{\vartheta}^m\,|\,\mathrm{Y}}$ the corresponding posterior distribution of $\boldsymbol{\vartheta}^m$ given Y. Note that the dimension parameter $m$ plays the role of a tuning parameter. As usual, its choice depends on the noise level $\epsilon$ so that we write $m_\epsilon$ and, in general, $m_\epsilon \to \infty$ as $\epsilon \to 0$. Given a choice $m_\epsilon$ of the dimension parameter, which in turn determines a prior sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon}}\}_{m_\epsilon}$ in dependence of the noise level $\epsilon$, our objective is the study of frequentist properties of the associated posterior sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon}\,|\,\mathrm{Y}}\}_{m_\epsilon}$. To be more precise, let $\theta^\circ$ be the realization of the random parameter $\boldsymbol{\vartheta}$ associated with the data-generating distribution and denote by $\mathbb{E}_{\theta^\circ}$ the corresponding expectation. A quantity $\Phi_\epsilon$ which is up to a constant a lower and an upper bound of the concentration of the

posterior sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon}\,|\,\mathrm{Y}}\}_{m_\epsilon}$, i.e.,

$$(3)\quad \lim_{\epsilon\to 0} \mathbb{E}_{\theta^\circ} P_{\boldsymbol{\vartheta}^{m_\epsilon}\,|\,\mathrm{Y}}((K)^{-1}\,\Phi_\epsilon \leqslant \|\boldsymbol{\vartheta}^{m_\epsilon} - \theta^\circ\|_{\ell_2}^2 \leqslant K\,\Phi_\epsilon) = 1 \quad \text{with } 1 \leqslant K < \infty,$$

is called exact posterior concentration (see, e.g., [3] or [2] for a broader discussion of the concept of posterior concentration). We shall emphasise that the derivation of the posterior concentration relies strongly on tail bounds for non-central $\chi^2$ distributions established in [1]. Moreover, if $\Phi_\epsilon \to 0$ as $\epsilon \to 0$ then the lower and upper bound given in (3) establish posterior consistency and $\Phi_\epsilon$ is called exact posterior concentration rate. Obviously, the exact rate depends on the prior sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon}}\}_{m_\epsilon}$, the choice of the dimension parameter $m_\epsilon$ and on the unknown parameter $\theta^\circ$.

In the spirit of a frequentist oracle approach, given a parameter $\theta^\circ$, by characterising an oracle-choice $m_\epsilon^\circ$ of the dimension parameter, we derive in this paper a prior sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon^\circ}}\}_{m_\epsilon^\circ}$ with smallest possible exact posterior concentration rate $\Phi_\epsilon^\circ$ which we call, respectively, an oracle prior sub-family and an oracle posterior concentration rate. On the other hand side, following a minimax approach, [4], for example, derive the minimax rate of convergence $\Phi_\epsilon^\star$ of the maximal mean integrated squared error (MISE) over a given class $\Theta_{\mathfrak{a}}$ of parameters (introduced below). By determining first a minimax-choice $m_\epsilon^\star$ of the dimension parameter we construct a sub-family $\{P_{\boldsymbol{\vartheta}^{m_\epsilon^\star}}\}_{m_\epsilon^\star}$ of prior distributions with exact posterior concentration rate $\Phi_\epsilon^\star$ uniformly over $\Theta_{\mathfrak{a}}$ which does not depend on the true parameter $\theta^\circ$ but only on the set of possible parameters $\Theta_{\mathfrak{a}}$. We shall emphasize, that the prior specifications we propose in [5] lead to exact posterior concentration rates that are optimal in an oracle or minimax sense over certain classes of parameters not only in the direct model but also in the more general indirect model. However, both oracle and minimax sieve prior are unfeasible in practise since they rely on the knowledge of either $\theta^\circ$ itself or its smoothness.

Our main contribution in [5] is the construction of a hierarchical prior $P_{\boldsymbol{\vartheta}^{\mathrm{M}}}$ that is adaptive. Meaning that, given a parameter $\theta^\circ \in \ell_2$ or a classes $\Theta_{\mathfrak{a}} \subset \ell_2$ of parameters, the posterior distribution $P_{\boldsymbol{\vartheta}^{\mathrm{M}}\,|\,\mathrm{Y}}$ contracts, respectively, at the oracle rate or the minimax rate over $\Theta_{\mathfrak{a}}$ while the hierarchical prior $P_{\boldsymbol{\vartheta}^{\mathrm{M}}}$ does not rely neither on the knowledge of $\theta^\circ$ nor on the class $\Theta_{\mathfrak{a}}$. Let us briefly elaborate on the hierarchical structure of the prior which induces an additional prior on the tuning parameter $m$, i.e., $m$ itself is a realisation of a random variable M. We construct a prior for $M$ such that the marginal posterior for $\boldsymbol{\vartheta}^{\mathrm{M}}$ (obtained by integrating out M with respect to its posterior) contracts exactly at the oracle concentration rate. This is possible for every $\theta^\circ$ whose components differ from the components of the prior mean infinitely many times. In addition, for every $\theta^\circ$ in the class $\Theta_{\mathfrak{a}}$ we show that the posterior distribution $P_{\boldsymbol{\vartheta}^{\mathrm{M}}\,|\,\mathrm{Y}}$ contracts at least at the minimax rate $\Phi_\epsilon^\star$ and that the corresponding Bayes estimate is minimax-optimal. Thereby, the proposed Bayesian procedure is *minimax adaptive* over the class $\Theta_{\mathfrak{a}}$. Interestingly, in the particular case of a diffuse prior each component of the data-driven Bayes estimator is shrunk proportional to the associated values of a penalised contrast

criterion which motivates a data-driven estimator, for example, in a deconvolution model.

## References


[1] L. Birgé, *An alternative point of view on lepski's method.* In Monogr., editor, *State of the art in probability and statistics*, **36** of *IMS Lecture Notes* (2001), pages 113–133. (Leiden 1999).

[2] I. Castillo, *Lower bounds for posterior rates with gaussian process priors.* Electronic Journal of Statistics, **2** (2008), 1281–1299.

[3] S. Ghosal, J. K. Ghosh and A. W. Van Der Vaart, *Convergence rates of posterior distributions.* The Annals of Statistics, **28**(2) (2000) 500–531.

[4] J. Johannes and M. Schwarz, *Adaptive gaussian inverse regression with partially unknown operator.* Communications in Statistics - Theory and Methods, **42**(7) (2013), 1343–1362.

[5] J. Johannes, A. Simoni and R. Schenk, *Adaptive Bayesian estimation in indirect Gaussian sequence space models.*, Discussion paper, (2015), http://arxiv.org/abs/1502.00184


## Dating structural breaks in functional data without dimension reduction


Alexander Aue

(joint work with Gregory Rice and Ozan Sonmez)


This talk is on functional data analysis and structural breaks analysis for dependent observations. Functional data analysis (FDA) has seen an upsurge in research contributions in the past decade. These are documented, for example, in the comprehensive books by Ramsay and Silverman [15] and Ferraty and Vieu [12]. Research concerned with structural breaks has a longstanding tradition in both the statistics and econometrics communities. Two recent reviews by Aue and Horváth [8] and Horváth and Rice [14] highlight newer developments, the first with a particular focus on time series.

Early work in functional structural break analysis dealt primarily with random samples of independent curves, the question of interest being whether all curves have a common mean function or whether there are two or more segments of the data that are homogeneous within but heterogeneous without. For example, Berkes et al. [10] developed statistical methodology to test the null hypothesis of no structural break against the alternative of a (single) mean function change assuming that the error terms are independent and identically distributed curves. Aue et al. [6] quantified the large-sample behavior of a break date estimator under a similar set of assumptions. The work in these two papers was generalized by Aston and Kirch ([3, 4]) to include functional time series into the modeling framework. The methods developed in these papers were applied to temperature curves [10] and functional magnetic resonance imaging [4]. Other papers dealing with functional time series include Antoniadis and Sapatinas [2] and Antoniadis et al. [1], who predicted daily electricity demand curves, Aue et al. ([5, 7]), who studied daily particulate matter curves, and Besse et al. (2000), who dealt with functional climate variations.

Most of the procedures in the aforementioned papers are based on dimension reduction techniques, for example, using the widely popular functional principal components analysis (fPCA), by which the functional variation in the data is projected onto the directions of a small number of principal curves, and multivariate techniques are then applied on the resulting sequence of score vectors. This is also the case in functional structural break detection, in which after an initial fPCA step multivariate structural break theory is utilized. Despite the fact that functional data are, at least in principle, infinite dimensional, the state of the art in FDA remains to start the analysis with an initial dimension reduction procedure.

Dimension reduction approaches, however, automatically incur a loss of information, namely all information about the functional data that is orthogonal to the basis onto which it is projected. This weakness is easily illustrated in the context of detecting and dating structural breaks in the mean function: if the function representing the mean break is orthogonal to the basis used for dimension reduction, there cannot be a consistent test or estimator for the break date in that basis.

The main purpose of this talk it to introduce methodology for the dating of structural breaks in functional data without the application of dimension reduction techniques, an idea that was touched upon in Horváth et al. [13] in the context of stationarity tests for functional time series. Here, a fully functional estimator for the break date is proposed.

The theory developed for this case illuminates a number of potential advantages of the fully functional estimator. When the direction of the break is orthogonal to the leading principal components of the data, the estimation of the mean break is asymptotically improved over fPCA based techniques. In addition, the assumptions required for the fully functional theory are weaker than the ones used in Aue et al. [6] and Aston and Kirch [3, 4], as convergence of the eigenvalues of the empirical covariance operator to the eigenvalues of the population covariance operator do not have to be accounted for. These assumptions are typically formulated as finiteness of fourth moment conditions. The relaxation obtained here may be particularly useful for applications to intra-day financial data such as one-minute log-returns on Microsoft stock. The talk is based on the manuscript Aue et al. [9].

## References


[1] A. Antoniadis, E. Paparoditis & T. Sapatinas. *A functional wavelet-kernel approach for time series prediction*, Journal of the Royal Statistical Society, Series B **68** (2006), 837–857.

[2] A. Antoniadis & T. Sapatinas. *Wavelet methods for continuous time prediction using Hilbert-valued autoregressive processes*, Journal of Multivariate Analysis **87** (2003), 133–158.

[3] J.A.D. Aston & C. Kirch. *Detecting and estimating changes in dependent functional data* Journal of Multivariate Analysis **109** (2012a), 204–220.

[4] J.A.D. Aston & C. Kirch. *Evaluating stationarity via change-point alternatives with applications to FMRI data*, The Annals of Applied Statistics **6** (2012b), 1906–1948.

[5] A. Aue, D. Dubart Nourinho & S. Hörmann. *On the prediction of stationary functional time series*, Journal of the American Statistical Association **110** (2015), 378–392.

[6] A. Aue, R. Gabrys, L. Horváth & P. Kokoszka. *Estimation of a change-point in the mean function of functional data*, Journal of Multivariate Analysis **100** (2009), 2254–2269.

[7] A. Aue, S. Hörmann, L. Horváth & M. Hušková. *Dependent functional linear models with applications to monitoring structural change*, Statistica Sinica **24** (2014), 1043–1073.
[8] A. Aue & L. Horváth. *Structural breaks in time series*, Journal of Time Series Analysis **34** (2013), 1–16.
[9] A. Aue , G. Rice & O. Sönmez. *Dating structural breaks in functional data without dimension reduction*, Preprint available at `arXiv:1511.04020`.
[10] I. Berkes, R. Gabrys, L. Horváth & P. Kokoszka. *Detecting changes in the mean of functional observations*, Journal of the Royal Statistical Society, Series B **71** (2009), 927–946.
[11] P. Besse, H. Cardot & D. Stephenson. *Autoregressive forecasting of some functional climatic variations*, Scandinavian Journal of Statistics **27** (2000), 673–687.
[12] F. Ferraty & P. Vieu (2010). *Nonparametric Functional Data Analysis*. Springer-Verlag, New York.
[13] L. Horváth, P. Kokoszka & G. Rice. *Testing stationarity of functional time series*, Journal of Econometrics **179** (2014), 66–82.
[14] L. Horváth & G. Rice. *Extensions of some classical methods in change point analysis*, Test **23** (2014), 219–255.
[15] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer Series in Statistics, 2005

## The functional discrete Fourier transform

Siegfried Hörmann
(joint work with Clément Cerovecki)

Functional data often arise by segmenting a continuous time process into natural units, such as days. Then a certain degree of dependence between the observations is well expected and, consequently, a thorough statistical investigation requires time series methodology. More precisely, we consider a time series $(X_t\colon t \in \mathbb{Z})$ with realizations in some function space. Then every observation $X_t$ is a random curve $(X_t(\tau)\colon \tau \in \mathcal{U})$ with some continuous domain $\mathcal{U}$ and we call the discrete time process $(X_t\colon t \in \mathbb{Z})$ a *functional time series (FTS)*. During recent years functional time series (FTS) analysis has seen an upsurge in the scientific community and diverse related practical and theoretical problems have been addressed.

Some of the latest publications are related to frequency domain topics for FTS. We refer to [3], [4], [1] and [2]. In their seminal article on frequency domain methodology for FTS, [3] have studied, among others, the limiting behavior of the discrete Fourier transform of some FTS:

$$S_n(\theta) = \sum_{t=1}^{n} X_t e^{-\mathrm{i}t\theta}, \quad \theta \in (-\pi, \pi].$$

This object is of interest to statisticians since it is closely related to the *periodogram* which can, for example, be used to detect some underlying periodic behavior of the time series. Applying such a test for periodicity requires knowledge of the distribution of $S_n(\theta)$. But unless $(X_t\colon t \in \mathbb{Z})$ is a Gaussian process, the exact distribution is infeasible and then we need to rely on asymptotics. Moreover we notice that with $\theta = 0$ this framework also contains the regular partial sums process, which is without any doubt a crucial building block in many statistical procedures.

For real valued processes asymptotic normality for $S_n(\theta)$ has been obtained under several dependence conditions. A recent contribution is [5], which covers a variety of special cases, including strong mixing sequences. The latter article also contains a more detailed literature survey. For functional data the afore mentioned paper of [3] shows that under higher order moment and cumulant assumptions $\frac{1}{\sqrt{n}}S_n(\theta)$ converges to a (complex) Gaussian random element whose covariance operator is given by

$$2\pi\mathcal{F}_\theta := \sum_{h\in\mathbb{Z}} C_h e^{-\mathrm{i}h\theta},$$

where $C_h$ is the lag $h$ covariance operator of the stationary functional time series. The operator $\mathcal{F}_\theta$, which can be shown to be self-adjoint and non-negative definite, is called the *spectral density operator*.

The first main results of this article shows the weak convergence of $S_n(\theta)/\sqrt{n}$ for *purely non-deterministic* processes. More precisely, letting $\mathcal{G}_t = \sigma(X_t, X_{t-1}, \ldots)$ the $\sigma$-algebra generated by $(X_s\colon s \leqslant t)$ and $\mathcal{G}_{-\infty} = \bigcap_{t\geq 0}\mathcal{G}_{-t}$ we impose the following assumption.

**Assumption 1.** *The process $(X_t\colon t\in\mathbb{Z})$ is stationary and ergodic with values in a separable Hilbert space $H$ and satisfies $E[X_0|\mathcal{G}_{-\infty}] = 0$ a.s.*

Below $\mathcal{CN}_H(\mu,\Gamma)$ denotes a complex Gaussian element in $H$ with mean $\mu$, covariance $\Gamma$ and with a relation operator which is zero. We use $\|\cdot\|$ for the norm of $H$. We write $X\in L^2_H(\Omega)$ if $E\|X\|^2<\infty$.

**Theorem 2.** *Let $(X_t\colon t\in\mathbb{Z})$ is process with values in $L^2_H(\Omega)$ which satisfies Assumption 1. Then for almost every $\theta\in(-\pi,\pi]$ there exists a linear operator $\mathcal{F}_\theta$, which is self-adjoint and non-negative definite such that*

$$\frac{1}{\sqrt{n}}S_n(\theta) \xrightarrow{d} \mathcal{CN}_H(0,\pi\mathcal{F}_\theta).$$

*Moreover we have that*

- **(I)** : $\frac{1}{n}\mathrm{Var}\big(S_n(\theta)\big)$ *converges in weak operator topology to* $2\pi\,\mathcal{F}_\theta$*;*
- **(II)** : $\frac{1}{n}\mathbb{E}\|S_n(\theta)\|^2 = \frac{1}{n}\mathrm{tr}\big(\mathrm{Var}\big(S_n(\theta)\big)\big) \to 2\pi\,\mathrm{tr}\big(\mathcal{F}_\theta\big) < \infty$*;*
- **(III)**: $C_h = \int_{-\pi}^{\pi}\mathcal{F}_\theta\, e^{\mathrm{i}h\theta}d\theta, \qquad \forall h\in\mathbb{Z}$*;*
- **(IV)**: *for almost all $\theta,\theta'$ the components of $\frac{1}{\sqrt{n}}(S_n(\theta),S_n(\theta'))$ are asymptotically jointly Gaussian and independent.*

For practical reasons it is useful to know for which frequencies Theorem 2 holds. For example, $\theta=0$ is an important special case, but we cannot say if this frequency is part of the exceptional null set or not. Requiring Assumption 2 below allows us to establish the same result for some fixed frequency $\theta_0$. To formulate this assumption we introduce the projection operator $\mathcal{P}_k := E[\,\cdot\,|\mathcal{G}_k] - E[\,\cdot\,|\mathcal{G}_{k-1}]$, $k\in\mathbb{Z}$.

**Assumption 2.** *The process $(X_t\colon t\in\mathbb{Z})$ is stationary and ergodic and for some $\theta_0\in(-\pi,\pi]$ the following properties hold:*

**(A1):** $\sum_{t=0}^{n} \mathcal{P}_0(X_t)e^{-\mathrm{i}t\theta_0}$ *is a Cauchy sequence in* $L_H^2(\Omega)$*;*
**(A2):** $E\big\|E[S_n(\theta_0)|\mathcal{G}_0]\big\|^2 = o(n)$.

This is our second main result.

**Theorem 3.** *Let* $(X_t : t \in \mathbb{Z})$ *be a sequence in* $L_H^2(\Omega)$ *which satisfies Assumption 2 for some given* $\theta_0 \in (-\pi, \pi]$*. Then the CLT and the conclusions* **(I)**–**(III)** *of Theorem 2 hold for frequency* $\theta_0$*. If Assumption 2 holds in addition for some* $\theta_0' \neq \theta_0$*, then conclusion* **(IV)** *of Theorem 2 holds with* $(\theta, \theta') = (\theta_0, \theta_0')$*.*

Although Assumption 2 looks technical it is relatively easy to verify in a variety of models, including functional ARMA, linear processes and $L^2$-$m$-approximable processes.

## References


[1] S. Hörmann, L. Kidziński and M. Hallin, *Dynamic Functional Principal Component*, Journal of the Royal Statistical Society: Series B **77** (2015), 319–348.
[2] S. Hörmann, L. Kidziński and P. Kokoszka, *Estimation in functional lagged regression*, Journal of Time Series Analysis **36** (2015), 541–561.
[3] V.M. Panaretos and S. Tavakoli, *Fourier analysis of stationonary time series in function spaces*, The Annals of Statistics, **41** (2013), 568–603.
[4] V.M. Panaretos and S. Tavakoli, *Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series*, Stochastic Processes and their Applications, **123** (2013), 2779–2807.
[5] M. Peligrad and W.B. Wu, *Central limit theorem for Fourier transforms of stationary processes*, The Annals of Probability, **38** (2010), 2009–2022.


## Principal Components Analysis for Time Series


Qiwei Yao
(joint work with Jinyuan Chang and Bin Guo)


Let $\mathbf{y}_t$ be observable $p \times 1$ weakly stationary time series. We seek for a linear transformation

$$\mathbf{x}_t = \mathbf{A}\mathbf{y}_t$$

such that the transformed series $\mathbf{x}_t$ consists of $q\,(> 1)$ both contemporaneously and serially uncorrelated subseries. Hence all the autocovariance matrices of $\mathbf{x}_t$ are of the same block-diagonal structure with $q$ blocks. Denote the segmentation of $\mathbf{x}_t$ by

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{x}_t^{(1)} \\ \vdots \\ \mathbf{x}_t^{(q)} \end{pmatrix}, \quad \text{where } \mathrm{Cov}(\mathbf{x}_t^{(i)}, \mathbf{x}_s^{(j)}) = \mathbf{0} \text{ for all } t, s \text{ and } i \neq j. \tag{1}$$

Therefore $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(q)}$ can be modelled or forecasted separately as far as their linear dynamic structure is concerned.

The proposed method can be viewed as an extension of the standard PCA for multiple time series, therefore, is abbreviated as TS-PCA. However the segmented

subseries are not guaranteed to exist as those subseries must not correlate with each other across all times. This is a marked difference from the standard PCA. The real data examples indicate that it is often reasonable to assume that the segmentation exists. Furthermore, when the assumption is invalid, the proposed method provides some approximate segmentations which ignore some weak though significant correlations, and those weak correlations are of little practical use for modelling and forecasting. Thus the proposed method can be used as an initial step in analysing multiple time series, which often transforms a multi-dimensional problem into several lower-dimensional problems. Furthermore the results obtained for the transformed subseries can be easily transformed back to the original multiple time series. Illustration with real data examples indicates clearly the advantages in post-sample forecasting from using the proposed TS-PCA. The *R*-package `PCA4TS`, available from CRAN project, implements the proposed methodology.

## Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference

Ingrid Van Keilegom

(joint work with Francesco Bravo and Juan Carlos Escanciano)

In both parametric and certain nonparametric statistical models, the empirical likelihood ratio satisfies a nonparametric version of Wilks' theorem. For many semiparametric models, however, the commonly used two-step (plug-in) empirical likelihood ratio (see [2]) is not asymptotically distribution free, that is, Wilks' phenomenon breaks down. In this paper we suggest a general approach to restore Wilks' phenomenon in two-step semiparametric empirical likelihood inferences. The main insight consists in using as the moment function in the estimating equation the influence function of the plug-in sample moment. The proposed method is general, and leads to distribution-free inference, namely to a $\chi^2$-limit. Hence, Wilks' phenomenon is valid. Moreover, the proposed procedure is less sensitive to the first step estimator than alternative bootstrap methods. Several examples are studied in more detail, and the high level conditions under which the general theory is valid, are verified for these examples. They include empirical likelihood based inference for : (1) the mean of interval censored data; (2) the average treatment effect in observational studies; (3) estimating equations with missing data; and (4) censored quantile regression. A simulation study illustrates the generality of the procedure and demonstrates its good finite sample performance compared to competing procedures based on asymptotic normality, bootstrap or two-step (plug-in) empirical likelihood. In a second step the proposed procedure can be generalized to empirical likelihood based goodness-of-fit tests for semiparametric models that are distribution free.

## References

[1] F. Bravo, J.C. Escanciano and I. Van Keilegom, *Wilks' phenomenon in two-step semiparametric empirical likelihood inference*, (2015) (submitted).

[2] N.L. Hjort, I.W. McKeague and I. Van Keilegom, *Extending the scope of empirical likelihood*, The Annals of Statistics, **37** (2009), 568–603.


## Spectral analysis of high-dimensional sample covariance matrices in the missing-at-random scenario


Angelika Rohde

(joint work with Kamil Jurczak)


We study asymptotic spectral properties of high-dimensional sample covariance matrices with missing observations. Let

$$Y = (Y_1, ..., Y_n) \in \mathbb{R}^{d\times n},\ Y_k = (Y_{1k}, ..., Y_{dk})^* \in \mathbb{R}^d,\ k = 1, ..., n,$$

be a sample of independent identically distributed (iid) random vectors with covariance matrix

$$T = \mathbb{E}\big((Y_1 - \mathbb{E}Y_1) \otimes (Y_1 - \mathbb{E}Y_1)\big).$$

In examples as described above, we do not observe the whole random vector $Y_k$ but some of its components. This missingness is represented by a random matrix $\varepsilon \in \mathbb{R}^{d\times n}$ with entries

$$\varepsilon_{ik} = \begin{cases} 1 & \text{if } Y_{ik} \text{ is observed} \\ 0 & \text{if } Y_{ik} \text{ is missing.} \end{cases}$$

Under the assumption that the matrices $Y$ and $\varepsilon$ are independent, the estimator

$$\hat{T}_{ij} = \frac{1}{N_{ij}} \sum_{k\in\mathcal{N}_{ij}} \big(Y_{ik} - \bar{Y}_i\big)\big(Y_{jk} - \bar{Y}_j\big)$$

is the analogue of the sample covariance and hence the natural estimator for $T_{ij}$, where

$$\mathcal{N}_{ij} = \Big\{k \in \{1, \dots, n\} :\ \varepsilon_{ik}\varepsilon_{jk} = 1\Big\},\quad N_{ij} = 1 \vee \#\mathcal{N}_{ij} \tag{1}$$

and

$$\bar{Y}_i = \frac{1}{N_{ii}} \sum_{k\in\mathcal{N}_{ii}} Y_{ik}.$$

Subsequently, $\hat{T} = (\hat{T}_{ij}) \in \mathbb{R}^{d\times d}$ is referred to as sample covariance matrix with missing observations. If $\mathbb{E}Y_k = 0$ is known in advance one typically uses the estimator

$$\hat{\Sigma} = \big(\hat{\Sigma}_{ij}\big) \in \mathbb{R}^{d\times d},\quad \hat{\Sigma}_{ij} = \frac{1}{N_{ij}} \sum_{k\in\mathcal{N}_{ij}} Y_{ik}Y_{jk}.$$

In what follows we write $\hat{\Xi}$ for $\hat{T}$ and $\hat{\Sigma}$ if a statement holds for both estimators.

## 1. Assumptions

Let $(X(i,k))_{i,k\in\mathbb{N}}$ be a double array of iid centered random variables with unit variance. The left upper $d \times n$ submatrix is denoted by $X_{d,n}$. Then the random vectors $Y_{1,d,n}, \ldots, Y_{n,d,n} \in \mathbb{R}^d$ are the columns of the matrix

$$Y_{d,n} - \mathbb{E}Y_{d,n} = T_{d,n}^{1/2} X_{d,n}.$$

with

$$T_{d,n} = \operatorname{diag}(T_{11,d,n}, \ldots, T_{dd,d,n}) \in \mathbb{R}^{d\times d}.$$

This structure on the population covariance matrix is the simplest one which allows to visualize the effects of missing observations on the spectrum of the sample covariance matrix. In this article we investigate asymptotic spectral properties of $\hat{\Xi}$ under the classical missing (completely) at random (MAR) setting. $(\varepsilon_{d,n})_{d,n}$ is a triangular array of random matrices $\varepsilon_{d,n} \in \mathbb{R}^{d\times n}$ independent of $(X(i,k))_{i,k\in\mathbb{N}}$, where the entries $\varepsilon_{ik,d,n}$ are independent Bernoulli variables with observation probabilities

$$\mathbb{P}(\varepsilon_{ik,d,n} = 1) = p_{i,d,n},\ i = 1, \ldots, d,\ k = 1, \ldots n.$$

The dependence of the set $\mathcal{N}_{ij}$ and the number $N_{ij}$ in (1) on the sequence $(\varepsilon_{d,n})$ is indicated by an additional subscript $d, n$. Throughout this report we impose that the family of spectral measures of the population covariance matrices $(T_{d,n})$ as well as the family of empirical distributions

$$(\mu^{w_{d,n}})_{d,n}, \quad \text{with } \mu^{w_{d,n}} = \frac{1}{d}\sum_{i=1}^{d} \delta_{w_{i,d,n}} \text{ and } w_{d,n} = \left(p_{1,d,n}^{-1}, ..., p_{d,d,n}^{-1}\right),$$

are tight. Asymptotic statements refer to $d \to \infty$ while $n = n(d)$ satisfies $\limsup_{d\to\infty} (d/n) < \infty$. The sequence of sample covariance matrices with missing observations is denoted by $(\hat{\Xi}_{d,n})_{d,n}$, the corresponding sequence of spectral measures by $(\mu_{d,n})_{d,n}$ and their Stieltjes transforms by $(m_{d,n})_{d,n}$.

## 2. Results

Define

$$S_{d,n} = \operatorname{diag}\left(\frac{1-p_{1,d,n}}{p_{1,d,n}} T_{11,d,n}, \ldots, \frac{1-p_{d,d,n}}{p_{d,d,n}} T_{dd,d,n}\right)$$

$$\text{and } R_{d,n} = \operatorname{diag}\left(\frac{1}{p_{1,d,n}} T_{11,d,n}, \ldots, \frac{1}{p_{d,d,n}} T_{dd,d,n}\right).$$

**Theorem 2.1.** *Suppose that the assumptions stated in Section 1 hold, and*

$$\sup_d \|R_{d,n}\|_{S_\infty} < \infty.$$

*Then for any $z \in \mathbb{C}^+$, we have $|m_{d,n}(z) - m^\circ_{d,n}(z)| \to 0$ a.s., where $m^\circ_{d,n}(z)$ satisfies*

$$m^\circ_{d,n}(z) \;=\; \frac{1}{d}\operatorname{tr}\left\{\left(\frac{1}{1+\frac{d}{n}e^\circ_{d,n}(z)} R_{d,n} - S_{d,n} - zI_{d\times d}\right)^{-1}\right\}$$

*and* $e^{\circ}_{d,n}$ *is the (unique) solution of the fixed point equation*

$$e^{\circ}_{d,n}(z) \;=\; \frac{1}{d}\mathrm{tr}\left\{ R_{d,n}\left( \frac{1}{1+\frac{d}{n}e^{\circ}_{d,n}(z)} R_{d,n} - S_{d,n} - zI_{d\times d}\right)^{-1}\right\}.$$

*Moreover,* $m^{\circ}_{d,n}$ *is the Stieltjes transform of a probability measure* $\mu^{\circ}_{d,n}$ *on* $\mathbb{R}$ *and*

$$\mu^{\circ}_{d,n} - \mu_{d,n} \Longrightarrow 0 \quad a.s.$$

It is well-known that the Stieltjes transform of the Marčenko-Pastur law with parameters $\left(y, \sigma^2/p_0\right)$ is the unique solution to

$$s(z) = \left( \frac{\sigma^2}{p_0} \cdot \frac{1}{1+\frac{\sigma^2}{p_0} y s(z)} - z \right)^{-1}$$

from $\mathbb{C}^+ \to \mathbb{C}^+$. In the special case $T_{d,n} = \sigma^2 I_{d\times d}$ and $p_{d,n} = (p_0, \ldots, p_0) \in (0,1)^d$, we have

$$m^{\circ}_{d,n}\left( z - \sigma^2 \frac{1-p_0}{p_0} \right) = \left( \frac{\sigma^2}{p_0} \frac{1}{1+\frac{d}{n}\frac{\sigma^2}{p_0} m^{\circ}_{d,n}\left( z - \sigma^2 \frac{1-p_0}{p_0}\right)} - z \right)^{-1}.$$

Hence, $\mu^{\circ}_{d,n}$ is the Marčenko-Pastur law $\mu^{MP}_{\frac{d}{n},\frac{\sigma^2}{p_0}}$ shifted by $\sigma^2\frac{1-p_0}{p_0}$ to the left.

**Corollary 2.2.** *Grant the conditions of Theorem 2.1. If* $p_{i,d,n} = p_0 > 0$ *for* $i = 1, \ldots, d$ *and* $d, n \in \mathbb{N}$ *and* $T_{d,n} = \sigma^2 I_{d\times d}$, $\sigma^2 > 0$, *we obtain*

$$\mu_{d,n} \Longrightarrow \mu^{MP}_{y,\frac{\sigma^2}{p_0}} \star \delta_{-\frac{1-p_0}{p_0}\sigma^2} \quad a.s.$$

*as* $d \to \infty$ *and* $d/n \to y > 0$.

For $\hat{\Sigma}_{d,n}$ we even determine the a.s. limit of the extremal eigenvalues.

**Theorem 2.3.** *Grant the conditions of Corollary 2.2 let additionally* $\mathbb{E}X_{11}^4 < \infty$ *and* $\varepsilon_{d,n} \in \mathbb{R}^{d\times n}$ *be the upper left corner of a double array* $(\varepsilon(i,k))_{i,k\in\mathbb{N}}$ *of iid Bernoulli variables with parameter* $p_0$. *Assume that* $\mathbb{E}Y_{d,n} = 0$. *Then, if* $0 < y < 1$,

$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) = \frac{\sigma^2}{p_0}\left(1-\sqrt{y}\right)^2 - \frac{1-p_0}{p_0}\sigma^2 \quad a.s., \quad \text{and}$$

$$\lim_{d\to\infty} \lambda_{\max}\left(\hat{\Sigma}_{d,n}\right) = \frac{\sigma^2}{p_0}\left(1+\sqrt{y}\right)^2 - \frac{1-p_0}{p_0}\sigma^2 \quad a.s.$$

The characterization of positive definiteness in the null case under the missing at random scenario is an immediate corollary of Theorem 2.3.

**Corollary 2.4.** *Under the condition of Theorem 2.3,*

$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) < 0 \quad a.s. \text{ if } \quad p_0 < 1 - (1-\sqrt{y})^2, \quad \text{and}$$

$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) > 0 \quad a.s. \text{ if } \quad p_0 > 1 - (1-\sqrt{y})^2.$$

## Methodology for nonparametric deconvolution when the error distribution is unknown

Aurore Delaigle
(joint work with Peter Hall)

In the nonparametric deconvolution problem we seek to estimate the distribution $F_X$ of a random variable $X$, but only observe independent and identically distributed (i.i.d.) data $W_1, \ldots, W_n$ on $W = X + U$, where $U$ denotes a measurement error independent of $X$. This classical error problem has received considerable attention since the late 80s. For several decades, research in the area was performed under the assumption that the distribution $F_U$ of the error $U$ was known. See for example Carroll and Hall [4], Stefanski and Carroll [14] and Fan [7].

The case of unknown $F_U$ has also been considered in the literature, where a common approach is to assume the availability of additional data that make it feasible to estimate $F_U$, for example a sample of replicated contaminated data (Li and Vuong [10], Lin and Carroll [11], Hall and Ma [8], Delaigle et al. [5] and Stefanski and McIntyre [15]), or a direct sample from the error distribution (Diggle and Hall, [6], and Neumann [13]). There is also interest in estimating $F_X$ when $F_U$ is unknown and no additional data are available, although work in this case consists largely of theoretical results in settings where a particular parametric model for $F_U$ is available, not general methods that can be applied broadly and enjoy particularly good performance. See Butucea and Matias [2], Meister [12], Butucea et al. [3] and Kneip et al. [9].

Taking a completely new viewpoint, in this work we argue that $F_X$ can in many cases be estimated consistently without knowing even the shape of $F_U$, and without extra data. All we require of $F_U$ is that it have basic properties of symmetry. This level of generality is unusual in errors-in-variables problems. To achieve it we use mathematical models for $F_X$ that are different from those that conventionally are assumed when techniques are devised, or theory is developed, or simulation studies are designed, in deconvolution problems. In particular, we suppose that $F_X$ is drawn from a "random universe," and for example is not symmetric.

We propose a new approach to inference in deconvolution problems, having two novel, distinct components. First, we use a minimum variance method to pinpoint the basic distribution that has been sampled, with noise, from a random universe. Secondly, we use discrete rather than continuous distributions as the basis for our methodology. In particular, our initial estimator of $F_X$ is discrete, and we suggest running a smoother through it to make it continuous. The performance of the resulting method is remarkably good, often equalling, or even exceeding, techniques that use additional data to estimate the distribution of measurement error.

This is a long abstract of the paper by [1], which was presented at the Oberwolfach workshop.

## References

[1] A. Delaigle, P. Hall, *Methodology for nonparametric deconvolution when the error distribution is unknown*, Journal of Royal Statistical Society. Series B. **78** (2016), 231–252.
[2] C. Butucea, C. Matias, *Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model*, Bernoulli **1**1 (2005), 309–340.
[3] C. Butucea, C. Matias, P. Pouet, *Adaptivity in convolution models with partially known noise distribution*, Electronic Journal of Statistics **2** (2008) 897–915.
[4] R.J. Carroll, P. Hall, *Optimal rates of convergence for deconvolving a density*, Journal of the American Statistical Association **83** (1988), 1184–1186.
[5] A. Delaigle, P. Hall, A. Meister, *On deconvolution with repeated measurements*, The Annals of Statistics **36** (2008), 665–685
[6] P. Diggle, P. Hall, *A Fourier approach to nonparametric deconvolution of a density estimate*, Journal of Royal Statistical Society. Series B. **55** (1993), 523–531.
[7] J. Fan, *On the optimal rates of convergence for nonparametric deconvolution problems*, The Annals of Statistics **19** (1991), 1257–1272.
[8] P. Hall, Y. Ma, *Semiparametric estimators of functional measurement error models with unknown error*, Journal of Royal Statistical Society. Series B. **69** (2007), 429–446.
[9] A. Kneip, L. Simar, I. Van Keilegom, *Boundary estimation in the presence of measurement error with unknown variance*, J. Econometrics **184** (2015), 379–393.
[10] T. Li, Q. Vuong, *Nonparametric estimation of the measurement error model using multiple indicators*, Journal of Multivariate Analysis **65** (1998), 139–165.
[11] X. Lin, R.J. Carroll, *Semiparametric estimation in general repeated measures problems*, Journal of Royal Statistical Society. Series B. **68** (2006), 69–88.
[12] A. Meister, *Density estimation with normal measurement error with unknown variance*, Statistica Sinica **16** (2006), 195–211.
[13] M.H. Neumann, *On the effect of estimating the error density in nonparametric deconvolution*, J. Nonparametric Statist. **7** (1997), 307–330.
[14] L.A. Stefanski, R.J. Carroll, *Deconvoluting kernel density estimators*, Statistics **21** (1990), 169–184.
[15] L.A. Stefanski, J. McIntyre, *Density estimation with replicate heteroscedastic measurements*, Ann Inst Stat Math. **63** (2011), 81–99.


## Functional Canonical Correlations

Tailen Hsing

We introduce a general notion of canonical correlation that extends the classical multivariate concept to include functional variables. The approach is based on the singular value decomposition of a particular linear operator defined on reproducing kernel Hilbert spaces corresponding to the functional variables $X$ and $Y$. In this context, canonical correlations and variables are limits of finite-dimensional subproblems thereby providing a seamless transition between Hotellings original development and infinite dimensional settings. As a proof of concept, we present some preliminary asymptotic theory for the inference problem. We first show that it is possible to estimate the canonical correlations and the corresponding variables consistently using a data-driven approach in a general setting. We also derive the rate of convergence for the special case where the eigenvalues of the marginal covariance operators decay in polynomial rates.

## References

[1] R. Eubank and T. Hsing, *Canonical correlation for stochastic processes*, Stochastic Processes and their Applications **118** (2007),1634–1661.

[2] T. Hsing and R. Eubank, *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (2015), Wiley.

*Reporter: Emilie Devijver*

# Participants

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

**Annika Betken**
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

**Dr. Jelena Bradic**
Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
UNITED STATES

**Prof. Dr. Gerda Claeskens**
OR & Business Statistics
Katholieke Universiteit Leuven
Naamsestraat 69
3000 Leuven
BELGIUM

**Prof. Dr. Aurore Delaigle**
Department of Mathematics
The University of Melbourne
Parkville, Victoria 3052
AUSTRALIA

**Prof. Dr. Holger Dette**
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

**Dr. Emilie Devijver**
Departement Wiskunde
Faculteit der Wetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200B
3001 Leuven
BELGIUM

**Prof. Dr. Irène Gijbels**
Department of Mathematics and
Center for Statistics
Katholieke Universiteit Leuven
W. de Croylaan 54
3001 Leuven (Heverlee)
BELGIUM

**Prof. Dr. Wenceslao Gonzalez-Manteiga**
Depto. de Estadistica e Investigacion Operativa
Facultad de Matematicas
Universidad de Santiago de Compostela
Rua Lope Gomez de Marzoa s/n
15782 Santiago de Compostela
SPAIN

**Thomas Gueuning**
ORSTAT
Katholieke Universiteit Leuven
Naamsestraat 69
3000 Leuven
BELGIUM

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

**Dr. Kristoffer Hellton**
Department of Mathematics
University of Oslo
Blindern
P.O. Box 1053
0316 Oslo
NORWAY

**Prof. Dr. Siegfried Hörmann**
Département de Mathématiques
Université Libre de Bruxelles
CP 210 Campus Plaine
Bd. du Triomphe
1050 Bruxelles
BELGIUM

**Prof. Dr. Tailen Hsing**
Department of Statistics
University of Michigan
460 West Hall
1085 South University
Ann Arbor MI 48109-1107
UNITED STATES

**Prof. Dr. Paul L. J. Janssen**
Universiteit Hasselt
Campus Diepenbeek
Building D
Agoralaan
3590 Diepenbeek
BELGIUM

**Prof. Dr. Jan Johannes**
Institut für Angewandte Mathematik
Ruprecht-Karls-Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

**Prof. Dr. Alois R. Kneip**
Institut für Gesellschafts-
und Wirtschaftswissenschaften
Universität Bonn
Adenauerallee 24-42
53113 Bonn
GERMANY

**Prof. Dr. Alexander Meister**
Fachbereich Mathematik
Universität Rostock
18051 Rostock
GERMANY

**Prof. Dr. Hans-Georg Mueller**
Division of Statistics
University of California
469 Kerr Hall
Davis, CA 95616-8705
UNITED STATES

**Prof. Dr. Axel Munk**
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

**Stanislav Nagy**
Statistics Section
Katholieke Universiteit Leuven
Celestijnenlaan 200B
3001 Leuven
BELGIUM

**Prof. Dr. Angelika Rohde**
Fakultät für Mathematik
Ruhr-Universität Bochum
Lehrstuhl für Stochastik
44780 Bochum
GERMANY

**Prof. Dr. Ingrid Van Keilegom**
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays 20
1348 Louvain-la-Neuve
BELGIUM

**Dr. Anneleen Verhasselt**
Center for Statistics
Hasselt University
Building D
Agoralaan
3590 Diepenbeek
BELGIUM

**Prof. Dr. Jane-Ling Wang**
Division of Statistics
University of California
469 Kerr Hall
Davis, CA 95616-8705
UNITED STATES

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