Mathematisches Forschungsinstitut Oberwolfach

Report No. 32/2016

DOI: 10.4171/OWR/2016/32

Statistics for Shape and Geometric Features

Organised by
Dragi Anevski, Lund
Christopher Genovese, Pittsburgh
Geurt Jongbloed, Delft
Wolfgang Polonik, Davis

3 July – 9 July 2016

Abstract. The constant emergence of novel technologies result in novel data generating devices and mechanisms that lead to a prevalence of highly complex data. To analyze such data, novel statistical methodologies need to be developed. This workshop addressed challenges that arise in the theoretical analyses of procedures in which *geometry, shape and topology* play central roles. The theoretical ideas involved here intersect deeply with a wide variety of fields, including mathematical statistics, probability theory, computational topology, and computational and differential geometry. The workshop brought together scholars with different perspectives, with the goal of facilitating cross-pollination to spur the development of new ideas, new analytical approaches, and new methods in geometric and shape statistics.

Mathematics Subject Classification (2010): 62xx, Secondary: 62Gxx, 62Hxx, 62Pxx, 60Bxx, 60Dxx, 60Fxx.

Introduction by the Organisers

The half-workshop *Statistics for Shape and Geometric Features*, organized by Dragi Anevski (Lund), Geurt Jongbloed (Delft), Christopher Genovese (Pittsburgh) and Wolfgang Polonik (Davis), was held July, 3rd – July 9th, 2016. This meeting was well attended by 26 participants with diverse geographic, demographic and disciplinary representation. For several of the participants it was the first time they attended an Oberwolfach Workshop, and they were deeply impressed by the workshop and the immensely stimulating atmosphere at the Forschungsinstitut.

The workshop consisted of presentations of the participants and discussions between them, either in groups or individually. The presentations earlier in the weeks

were intended to build a common platform for the participants, who came to the workshop with different backgrounds. These presentations were addressing principal component analysis for non-Euclidean data, inference for geometric objects, inference under shape constraints (log-concavity), and topological data analysis, respectively. Later in the week, the group discussed several emerging problems and ideas, including estimation in graphs under monotonicity constraints, algorithmic approaches for non-standard big data using the divide-and-conquer paradigm, and the estimation of flow lines. Some presentations also addressed applications of geometric/shape ideas to cutting-edge scientific problems, such as improving microscopy based image analysis, or the analysis of the filamentary structure of the world wide web. The PhD students attending the workshop also had an opportunity to present their dissertation research.

In summary, the workshop brought together scholars with related, but different statistical backgrounds that included shape constrained inference, topological data analysis, inference for geometric objects, and shape analysis. Corresponding major statistical problem areas include clustering and mode finding, identification and characterization of low-dimensional structures (e.g., embedded manifolds), as well as asymptotic distribution theory for estimators under various shape constraints. Another interesting aspect of the workshop was provided by a second half-workshop on Learning Theory and Approximation that was running in parallel. There was a lively interaction between the participants of the two workshops with complementary themes.

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 Jon A. Wellner in the "Simons Visiting Professors" program at the MFO.

Workshop: Statistics for Shape and Geometric Features

Table of Contents

Abstracts

Object Oriented Data Analysis

J. S. Marron

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the currently fashionable special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Challenges in modern medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as manifolds (where an approximating tangent plane can be fit to allow the application of standard methods), or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. For example, the former provides connection with the *statistics on manifolds* theme of this meeting. The latter has a connection to topology, since persistent homology has proven to provide a very useful approach. The notion of Object Oriented Data Analysis also impacts data analysis, through providing a useful terminology for interdisciplinary discussion of the many choices needed in many modern complex data analyses.

Rates of convergence for robust geometric inference

Bertrand Michel
(joint work with Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Pascal Massart, Alessandro Rinaldo, Larry Wasserman)

The last decades have seen an explosion in the amount of available data in almost all domains of science, industry, economy and even everyday life. These data, often coming as point clouds embedded in Euclidean spaces, usually lie close to some lower dimensional geometric structures (e.g. manifold, stratified space,...) reflecting properties of the system from which they have been generated. Inferring the topological and geometric features of such multivariate data has recently attracted a lot of interest in both statistical and computational topology communities.

Considering point cloud data as independent observations of some common probability distribution P in $\mathbb{R}^d$, many statistical methods have been proposed to infer the geometric features of the support of P such as principal curves and surfaces [14], multiscale geometric analysis [1], density-based approaches [13] or support estimation, to name a few. Although they come with statistical guarantees these methods usually do not provide geometric guarantees on the estimated features.

On another hand, with the emergence of Topological Data Analysis (TDA) [5], purely geometric methods have been proposed to infer the geometry of compact subsets of $\mathbb{R}^d$. These methods aims at recovering precise geometric information of

a given shape – see, e.g. [10, 16, 7]. Although these methods come with strong topological and geometric guarantees they usually rely on sampling assumptions that do not apply in statistical settings. In particular, these methods can be very sensitive to outliers. Indeed, they generally rely on the study of the sublevel sets of distance functions to compact sets. In practice only a sample drawn on, or close, to a geometric shape is known and thus only a distance to the data can be computed. The sup norm between the distance to the data and the distance to the underlying shape being exactly the Hausdorff distance between the data and the shape, we see that the statistical analysis of standards TDA methods boils down to the problem of support estimation in Hausdorff metric. This last problem has been the subject of much study in statistics, see for instance [12, 11, 18]. Being strongly dependent of the estimation of the support in Hausdorff metric, it is now clear why standard TDA methods may be very sensitive to outliers.

To provide a more robust approach of TDA, a notion of distance function to a measure (DTM) in $\mathbb{R}^d$ has been introduced by [8] as a robust alternative to the classical distance to compact sets. Given a probability distribution P in $\mathbb{R}^d$ and a real parameter $0 \leq u \leq 1$, [8] generalize the notion of distance to the support of P by the function

$$\delta_{P,u} : x \in \mathbb{R}^d \mapsto \inf\{t > 0\,;\, P(\bar{B}(x,t)) \geq u\}$$

where $\bar{B}(x,t)$ is the closed Euclidean ball of center x and radius t. For $u = 0$, this function coincides with the usual distance function to the support of P. For higher values of u, it is larger than the usual distance function since a portion of mass u has to be included in the ball centered on x. To avoid issues due to discontinuities of the map $P \to \delta_{P,u}$, the distance to measure (DTM) function with parameter $m \in [0,1]$ and power $r \geq 1$ is defined by

$$d_{P,m,r}(x) : x \in R^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^r(x)du\right)^{1/r}.$$

It was shown in [8] that the DTM shares many properties with classical distance functions that make it well-adapted for geometric inference purposes. First, it is stable with respect to perturbations of P in the Wasserstein metric . This property implies that the DTM associated to close distributions in the Wasserstein metric have close sublevel sets. Moreover, when $r = 2$, the function $d^2_{P,m,2}$ is semiconcave ensuring strong regularity properties on the geometry of its sublevel sets. Using these properties, [8] show that, under general assumptions, if $\tilde{P}$ is a probability distribution approximating P, then the sublevel sets of $d_{\tilde{P},m,2}$ provide a topologically correct approximation of the support of P. The introduction of DTM has motivated further works and applications in various directions such as topological data analysis [3], GPS traces analysis [6], density estimation [2], deconvolution [4] or clustering [9] just to name a few. However no strong statistical analysis of the DTM has not been proposed so far.

In practice, the measure P is usually only known through a finite set of observations $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ sampled from P, raising the question of the approximation of the DTM. A natural idea to estimate the DTM from $\mathbb{X}_n$ is to

plug the empirical measure P_n instead of P in the definition of the DTM. This "plug-in strategy" corresponds to computing the distance to the empirical measure (DTEM). It can be applied with other estimators of the measure P, for instance in [4] it was proposed to plug a deconvolved measure into the DTM.

For $m = \frac{k}{n}$, the DTEM satisfies

$$d^r_{P_n,k/n,r}(x) := \frac{1}{k}\sum_{j=1}^{k} \|x - \mathbb{X}_n\|^r_{(j)} ,$$

where $\|x - \mathbb{X}_n\|_{(j)}$ denotes the distance between x and its j-th neighbor in $\{X_1, \ldots, X_n\}$. This quantity can be easily computed in practice since it only requires the distances between x and the sample points.

In this talk, I present recent results about the deviations and the rate of convergence of $\Delta_{n,m,r}(x) := d^r_{P_n,m,r}(x) - d^r_{P,m,r}(x)$. Our results rely on a local analysis of the empirical process to compute tight deviation bounds of $\Delta_{n,\frac{k}{n},r}(x)$. More precisely, we use a sharp control of a supremum defined on the uniform empirical process. Such local analysis has been successfully applied in the literature about non asymptotic statistics, for instance [15] obtain fast rates of convergence in classification. We show that the rate of convergence of $\Delta_{n,\frac{k}{n},r}(x)$ directly depends on the regularity at zero of the quantile function of the push forward probability measure of P by the function $\|x - \cdot\|^r$.

I in this talk, I will also present some results about the functional convergence of $\Delta_{n,\frac{kn}{n},r}(x)$.

References

[1] E Arias-Castro, D. Donoho and X. Huo, *Adaptive multiscale detection of filamentary structures in a background of uniform random points*, The Annals of Statistics **34**(2006), 326–349.

[2] G. Biau,F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodriguez, *A weighted k-nearest neighbor density estimate for geometric inference*, Electronic Journal of Statistics**5**(2011), 204–237.

[3] M. Buchet, F. Chazal, T. K Dey, F. Fan, S. Oudot and Y. Wang *Topological analysis of scalar fields with outliers*, Proc. Sympos. on Computational Geometry 2015.

[4] C. Caillerie, F. Chazal, J. Dedecker and B. Michel, *Deconvolution for the Wasserstein metric and geometric inference*, Electron. J. Stat. **5**(2011), 1394–1423.

[5] G. Carlsson, *Topology and data*, Bulletin of the American Mathematical Society **46**(2009), 255–308.

[6] F. Chazal, D. Chen, L. Guibas, X. Jiang, and C. Sommer, *Data-driven trajectory smoothing*, In Proc. ACM SIGSPATIAL GIS 2011.

[7] F. Chazal, D. Cohen-Steiner, A. Lieutier and B. Thibert, *Stability of Curvature Measures*, Computer Graphics Forum (2009),1485–1496.

[8] F. Chazal, D. Cohen-Steiner and Q. Mérigot, *Geometric inference for probability measures*, Foundations of Computational Mathematics **11**(2011),733–751.

[9] F. Chazal, L. Guibas, S. Oudot and P. Skraba, *Persistence-based clustering in riemannian manifolds*, Journal of the ACM **41**(2013).

[10] F. Chazal and A. Lieutier, *Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees*, Computational Geometry **40**(2008), 156–170.

[11] A. Cuevas and A. Rodríguez-Casal, *On boundary estimation*, Advances in Applied Probability (2004), 340–354.

[12] L. Devroye and G. L. Wise, *Detection of abnormal behavior via nonparametric estimation of the support*, SIAM Journal on Applied Mathematics **38**(1980), 480–488.
[13] C. Genovese, M. Perone-Pacifico, I. Verdinelli and L. Wasserman, *On the path density of a gradient field*, The Annals of Statistics **37**(2009), 3236–3271.
[14] T Hastie and W Stuetzle, *Principal curves*, J. Amer. Statist. Assoc., **84**(1989), 502–516.
[15] E. Mammen and A. Tsybakov, *Smooth discrimination analysis*, The Annals of Statistics **27**(1999), 1808–1829.
[16] P. Niyogi, S. Smale and S. Weinberger, *Finding the homology of submanifolds with high confidence from random samples*, Discrete & Computational Geometry **39**(2008), 419–441.
[17] G. Shorack and J.A. Wellner, Empirical processes with applications to statistics, SIAM (2009)
[18] A. Singh, C. Scott and R. Nowak, *Adaptive hausdorff estimation of density level sets*, The Annals of Statistics, **37**(2009), 2760–2782.

Inference for Geometric Features

Christopher R. Genovese

The geometric features of a function are often of, direct or indirect, interest as the target of inference in many scientific problems. These features derived from the "shape" of a function and the geometry of its graph. Examples include local modes, ridges, level sets, conditional local modes, and a variety of derived complexes (e.g., Morse-Smale). In this talk, I describe a theoretical framework and a set of methods for infering the geometric features of densities (and other smooth functions). This includes both computing estimators and practically useful confidence sets for these features. I also show how these methods can be combined to address some larger statistical problems, including clustering, regression, and high-dimensional visualization. My collaborators in these lines of research (as reflected in the a sequence of sixteen papers over the last seven years) are Larry Wasserman (Carnegie Mellon), Isa Verdinelli (Carnegie Mellon, University of Rome), Marco Perone Pacifico (University of Rome), and Yen-Chi Chen (University of Washington).

This work reflects a broader trend in the field. As data sets become larger, higher dimensional, and more complex, there is increasing interest in making inferences about complex or aggregate objects, such as trees, networks, fields, elements of structured spaces, and discrete collections of smooth objects. Such objects raise a number of practical and theoretical challenges, including quantifying accuracy, deriving meaningful confidence sets, spatial matching, and visualization of result. Moreover, as described here, some of these problems are *statistically hard* in the sense that rates of convergence are logarithmic. Finally, complex and high-dimensional data sets often interesting low-dimensional structure that can be difficult to identify or (as a complex/aggregate object) difficult to estimate/learn.

A motivating example for the value of geometric features, especially for densities derived from point-cloud data is the problem of estimating filamentary structures in the distribution of mass across the universe. Cosmologists have discovered that mass is not distributed uniformly but rather looks like a "cosmic web." The structure of this web (a dense network of ridges) has cosmological significance in

that it is informative about conditions in the early universe. We also have shown that the estimated object can shed light on several other astronomical problems.

To understand how difficult it is to estimate various features, I focus on the simpler theoretical problem of estimating data drawn from a smooth (in the sense of bounded reach) manifold with random scatter. I derive the minimax rates of convergence for this problem, which varies strongly with noise model. Under the realistic model of smooth (e.g., Gaussian) noise in the embedding space, this problem is statistical hard in the sense described above.

What then to do if we are to avoid logarithmic rates? I discuss the idea of a *surrogate* – an object that captures the essential features of the true object (e.g., the features most interesting for a particular problem) and can be estimated with a polynomial rate of convergence. This necessarily entails some loss of information, but the benefit is that we can make useful inferences about some aspects of the true object.

I then describe a surrogate framework for estimating manifolds that becomes a general approach to estimating ridges (loosely, the zeroes of a project gradient) that are topologically and geometrically good surrogates for the underlying manifold. Building on the Subspace Constrained Mean-Shift algorithm of Ozertem and Erdogmus (2011), I develop an estimator for these ridge sets, which exhibits good performance in Hausdorff distance. (Optimality of the resulting rates has not yet been established as the rates do involve the dimension of the embedding space.) I show a variety of simulated and real examples to illustrate this method. Note that the method can be used not only to estimate the objects themselves but to compute bootstrap confidence sets as well.

I then illustrate a new method built on this framework for finding and separating ridges by dimension. I describe a ridge-based method for soft clustering in high dimensions, a technique based on these ideas for modal regression, a method for estimating the Morse-Smale complex of a smooth (Morse) function, and a method for testing for density modes. I briefly show the results of a promising new method for high-dimensional clustering that is still in development.

Inference for the mode of a log-concave density: a likelihood ratio test and confidence intervals

Jon A. Wellner
(joint work with Charles Doss)

Wellner discussed a likelihood ratio test for the mode of a log-concave density. The new test is based on comparison of the log-likelihoods corresponding to the unconstrained maximum likelihood estimator of a log-concave density and the constrained maximum likelihood estimator where the constraint is that the mode of the density is fixed, say at m. The constrained estimators have many properties in common with the unconstrained estimators discussed by Pal, Woodroofe, and Meyer (2007), Dümbgen and Rufibach (2009), and Balabdaoui, Rufibach and Wellner (2010), but they differ from the unconstrained estimator under the null

hypothesis on $n^{-1/5}$ neighborhoods of the mode m. Using joint limiting properties of the unconstrained and constrained estimators under the null hypothesis (and strict curvature of $\log f = \varphi$ at the mode), we show that the likelihood ratio statistic is asymptotically pivotal: that is, it converges in distribution to a limiting distribution which is free of nuisance parameters, thus playing the role of the χ^2_1 distribution in classical parametric statistical problems. By inverting this family of tests we obtain new (likelihood ratio based) confidence intervals for the mode of a log-concave density f. These new intervals do not depend on any smoothing parameters. We study the new confidence intervals via Monte Carlo studies and illustrate them with several real data sets. The new confidence intervals seem to have several advantages over existing procedures.

This talk was based on joint work with Charles Doss.

References

[1] F. Balabdaoui, K. Rufibach, and Jon A. Wellner, *Limit distribution theory for maximum likelihood estimation of a log-concave density*, Annals of Statistics **37**, 1299–1331.

[2] L. Dümbgen and K. Rufibach, *Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency*, Bernoulli **15** (2009), 40–68.

[3] J.K. Pal, M.B. Woodroofe, and M.C. Meyer, *Estimating a Polya frequency function*, *Complex Datasets and Inverse Problems: Tomography, Networks, and Beyond*, IMS Lecture Notes-Monograph Series, **54** (2007), 239–249.

Nanostatistics

Axel Munk

(joint work with Timo Aspelmeier)

Conventional light microscopes have been used for centuries for the study of small length scales down to about 250nm. Images from such a microscope are typically blurred and noisy and the measurement error can often be well approximated by Gaussian or Poisson noise which is due to local and temporal aggregation of photon emitting fluorophores.

Recording, recovery and enhancement of such images has been the focus of a multitude of deconvolution techniques in imaging during the past. However, conventional microscopes have an intrinsic physical limit of resolution which remained unchallenged for a century but which, with the advent of modern superresolution fluorescence microscopy techniques, was broken for the first time in the 1990s. These achievements have been awarded the Nobel Prize in Chemistry, 2014.

Since then, superresolution fluorescence microscopy has become an indispensible tool for studying structure and dynamics of living organisms. Current experimental advances go to the physical limits of imaging where discrete quantum effects are predominant. Consequently, superresolution fluorescence microscopy is inherently of a non-Gaussian statistical nature and we argue that recent technological

progress also challenges the Poisson assumption, which has been standing for a long time.

Hence it becomes necessary to analyze and exploit the discrete physical mechanisms of fluorescent molecules and light and their distributions in time and space in order to achieve the highest resolution possible. In this talk we present an overview of some physical principles underlying modern fluorescence microscopy techniques from a statistical modeling and analysis perspective.

Several issues of our own work will be discussed in more detail. This includes variational multiscale methods for confocal and stimulated emission depletion (STED) microscopy [2, 3, 4], drift correction for single marker switching (SMS) microscopy [8] as well as sparse estimation and background removal for superresolution by polarisation angle demodulation (SPoD) [5, 6, 7]. We illustrate that such methods benefit from advances in large scale computing, e.g. from recent tools from convex optimization. We argue that in the future, even higher resolutions will require more detailed models delving into sub-Poissonian statistics. To this end we develop a prototypical hidden Markov model for fluorophore dynamics [1] and use it to address recent challenges in quantitative biology to count the number of proteins at a certain spot. This task is tackled for STED microscopy by a diffrent approach based on quantum antibunching [9].

References

[1] Aspelmeier, T., Egner, A., Munk, A. (2015). Modern statistical challenges in high resolution fluorescence microscopy. *Annual Review of Statistics and its Application* 2, 163-202.

[2] Frick, K., Marnitz, P., Munk, A. (2012). Statistical Multiresolution Estimation in Imaging: Fundamental Concepts and Algorithmic Framework. *Electr. Journ. Statist.* 6, 231-268.

[3] Frick,K. Marnitz,P. Munk,A. (2013). Statistical multiresolution estimation for variational imaging: With an application in Poisson-biophotonics. *Journ. Math. Imaging Vision* 46, 370-387.

[4] Grasmair, M., Li, H., Munk, A. (2015). Variational multiscale nonparametric regression: smooth functions. arXiv: 1512.01068, Submitted.

[5] Hafi, N., Grunwald, M., van den Heuvel, L.S., Aspelmeier, T., Chen, J.-H., Zagrebelsky, M., Schütte, O.M., Steinem, C., Korte, M., Munk, A., Walla, P.J. (2014). Fluorescence nanoscopy by polarization modulation and polarization angle narrowing. *Nature Methods* 11, doi: 10.1038/nmeth.2919.

[6] Hafi, N., Grunwald, M., van den Heuvel, L.S., Aspelmeier, T., Chen, J.-H., Zagrebelsky, M., Schütte, O.M., Steinem, C., Korte, M., Munk, A., Walla, P.J. (2016). Reply to: Polarization modulation adds little additional information to super-resolution fluorescence microscopy. *Nature Methods* 13 (1), 8-9.

[7] Hafi, N., Grunwald, M., van den Heuvel, L.S., Aspelmeier, T., Chen, J.-H., Zagrebelsky, M., Schütte, O.M., Steinem, C., Korte, M., Munk, A., Walla, P.J. (2016). Corrigendum. *Nature Methods* 13 (1), 101.

[8] Hartmann, A., Huckemann, S., Dannemann, J., Laitenberger, O., Geisler, C., Egner, A., Munk, A. (2016). Drift estimation in sparse sequential dynamic imaging: With application to nanoscale fluorescence microscopy. *Journ. Royal Statist. Soc., Ser. B.*, 68, 563-587.

[9] Ta, H., Keller, J., Haltmeier, M. Saka, S.K., Schmied, J., Opazo, F., Tinnefeld, P., Munk, A., Hell, S.W. (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nature Communications* 6, doi: 10.1038/ncomms8977.

Asymptotics and optimal bandwidth selection for level set estimation

J.E. Chacón
(joint work with W. Polonik)

Given a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and $c \in (\inf f, \sup f)$, denote

$$L \equiv L(c) \equiv L_f(c) = \{x \in \mathbb{R}^d\colon f(x) \geq c\}$$

the (super)level set of f at level c. Level sets of density functions and regression functions have many applications in Statistics. Here we particularly focus on density level sets, which are useful for probability distribution representation [2, 5], nonparametric clustering [4] and topological data analysis [1]. Further examples of applications of level sets can be found in the introduction of [6].

As a first goal, the use of level sets for distribution representation is reviewed, noting its pros and cons, and providing a new class of (density) regions for distribution representation.

An alternative parametrization of density level sets is through its probability content: for $\alpha \in (0,1)$ use $L(c_\alpha)$ with

$$c_\alpha = \sup\big\{c > 0\colon P\big(X \in L(c)\big) \geq 1-\alpha \text{ for } X \sim f\big\}.$$

In this form, density level sets are also known as highest density regions (HDRs). It can be shown that, under regularity conditions, $P\big(L(c_\alpha)\big) = 1-\alpha$ so that $L(c_\alpha)$ can be interpreted as the most probable region containing at least $1-\alpha$ probability mass.

In the univariate case, highest density regions with $\alpha = 0.5$ can be compared to boxplots for distribution representation. Notice that, instead of most probable regions, boxplots are concerned with central regions. Hence, for symmetric unimodal distributions, HDRs and boxplots give the same distribution representation. For bimodal distributions, however, they provide different answers: whereas boxplots are in direct correspondence with the median and quartiles, HDRs are related to the modes and their domains of attraction, that is, to the density derivative. In the bivariate case, a similar comparison can be made between HDRs and bagplots [7].

Nevertheless, since different distribution features frequently appear at different values of the level c_α, fixing $\alpha = 0.5$ may fail to reveal the whole modal structure of the probability distribution. To solve this issue, here we introduce a further possibility for distribution representation: concave density regions (CDR). CDRs are defined as the regions of the space where the density function is concave. When the density is smooth enough, the boundary of a CDR is easily identified as the set of points where the determinant of the Hessian matrix vanishes. The main gain in using CDRs against HDRs is that all density clusters are visible in the CDR representation, as the number of connected components of the CDR equals the number of density modes. Also, CDRs do not depend on a level parameter. This is both an advantage, because the level does not have to be chosen, and a disadvantage, because the control of the probability content of CDRs is lost.

A second, different goal related to density level set estimation is to pose an open problem that highlights the fact that the current asymptotic results on the topic are somehow incomplete. The usual plug-in-type strategy for estimating $L = \{x \in \mathbb{R}^d \colon f(x) \geq c\}$ is to use $\widehat{L}_h = \{x \in \mathbb{R}^d \colon \widehat{f}_h(x) \geq c\}$, where

$$\widehat{f}_h(x) = (nh^d)^{-1} \sum_{i=1}^{n} K\big((x - X_i)/h\big)$$

is a kernel density estimator with bandwidth h.

A common way to evaluate the performance of $\widehat{L}_h$ as an estimator of L is to use the loss function induced by the distance in μ-measure, for a given measure μ. This is defined as $d_\mu\big(\widehat{L}_h, L\big) = \mu\big(\widehat{L}_h \triangle L\big)$, where $A \triangle B = (A \backslash B) \cup (B \backslash A)$ denotes the symmetric difference of any two sets A, B. Most used measures are those that can be written as $\mu_g(A) = \int_A g d\lambda$ with $g(x) = |f(x) - c|^p$ for some $p \geq 0$ or $g(x) = f(x)$, with λ standing for the Lebesgue measure in $\mathbb{R}^d$. The corresponding risk function is the mean distance in measure $\mathrm{MDM}(h) = \mathbb{E}\big[d_\mu\big(\widehat{L}_h, L\big)\big]$.

In a beautiful paper, [3] showed that it is possible to decompose $\mathrm{MDM}(h) = I_1(h) + I_2(h)$, where $I_1(h)$ is a term related to the variance of the kernel density estimator and $I_2(h)$ is a term related to the bias. Moreover, under some usual smoothness conditions it is proved that there exists an explicit constant $A \equiv A(f, K, c)$ such that $(nh^d)^{1/2} I_1(h) \to A$ as $n \to \infty$. If in addition $nh^{d+4} \to 0$ is assumed, then it can be shown that $(nh^d)^{1/2} I_2(h) \to 0$ as $n \to \infty$.

This additional condition on the bandwidth ensures that the squared bias is asymptotically negligible as compared to the variance, but note that this is not normally the case in practice when, for instance, optimal bandwidths are sought for. We conjecture that, without the additional condition on the sequence of bandwidths, there should be possible to find a constant $B \equiv B(f, K, c)$ such that $h^{-2} I_2(h) \to B$ as $n \to \infty$. The explicit identification of this constant B would allow to derive tools for asymptotically optimal bandwidth selection for density level set estimation.

References

[1] O. Bobrowski, S. Mukherjee and J.E. Taylor, *Topological consistency via kernel estimation*, Bernoulli (2015), to appear.

[2] A.W. Bowman and P. Foster, *Density based exploration of bivariate data*, Statistics and Computing **3** (1993), 171–177.

[3] B. Cadre, *Kernel estimation of density level sets*, Journal of Multivariate Analysis **97** (2006), 999–1023.

[4] J.E. Chacón, *A population background for nonparametric density-based clustering*, Statistical Science **30** (2015), 518–532.

[5] R.J. Hyndman, *Computing and graphing highest density regions*, American Statistician **50** (1996), 120–126.

[6] D.M. Mason and W. Polonik, *Asymptotic normality of plug-in level set estimates*, Annals of Applied Probability **19** (2009), 1108–1142.

[7] P.J. Rousseeuw, I. Ruts and J.W. Tukey, *The bagplot: a bivariate boxplot*, American Statistician **53** (1999), 382–387.

Current status linear regression

Kim Hendrickx
(joint work with Piet Groeneboom)

Investigating the relationship between a response variable Y and one or more explanatory variables is a key activity in statistics. Often encountered in regression analysis however, are situations where a part of the data is not completely observed due to some sort of censoring. We focus on modeling a linear relationship when the response variable is subject to interval censoring type I, i.e. instead of observing the response Y, one only observes whether or not $Y \leq T$ for some random censoring variable T, independent of Y. This type of censoring is often referred to as the current status model. Let $(X_i, T_i, \Delta_i), i = 1, \ldots, n$ be independent and identically distributed observations from $(X, T, \Delta) = (X, T, 1_{\{Y \leq T\}})$. We assume that Y is modeled as

$$Y = \beta_0' X + \varepsilon, \tag{1}$$

where β_0 is a k-dimensional regression parameter and ε is an unobserved random error, independent of (X, T) with unknown distribution function F_0. We assume that the distribution of (X, T) does not depend on (β_0, F_0) which implies that the relevant part of the log likelihood for estimating (β_0, F_0) is given by,

$$\begin{aligned} l_n(\beta, F) &= \sum_{i=1}^{n} \left[\Delta_i \log F(T_i - \beta' X_i) + (1 - \Delta_i) \log\{1 - F(T_i - \beta' X_i)\}\right] \\ &= \int \left[\delta \log F(t - \beta' x) + (1 - \delta) \log\{1 - F(t - \beta' x)\}\right] d\mathbb{P}_n(t, x, \delta), \end{aligned} \tag{2}$$

where $\mathbb{P}_n$ is the empirical distribution of the (T_i, X_i, Δ_i).

The profile maximum likelihood estimator (MLE) of β_0 was proved to be consistent by [Cosslett, 1983] but nothing seems to be known about its asymptotic distribution, apart from its consistency and upper bounds for its rate of convergence. Since the log likelihood as a function of β, obtained by maximizing the log likelihood with respect to the distribution function F for fixed β and substituting this maximizer back into the likelihood, is not a smooth function of β, it is unclear whether the MLE of β_0 is $\sqrt{n}$-consistent. [Murphy et al., 1999] derived an $n^{1/3}$-rate for the MLE under the condition that the support of the density of $T - \beta' X$ is strictly contained in the support of F_0, for all β. For a derivation of the efficient information $\tilde{\ell}^2_{\beta_0, F_0}$ given by,

$$\begin{aligned} \tilde{\ell}_{\beta, F}(t, x, \delta) = \{E(X | T - \beta' X = t - \beta' x) - x\}\, f(t - \beta' x) \\ \cdot \left\{ \frac{\delta}{F(t - \beta' x)} - \frac{1 - \delta}{1 - F(t - \beta' x)} \right\}, \end{aligned}$$

we refer to [Cosslett, 1987] for the binary choice model, and to [Huang and Wellner, 1993] and [Murphy et al., 1999] for the current status regression model.

Approaches to $\sqrt{n}$-consistent and efficient estimation of the regression parameters were considered by [Klein and Spady, 1993], [Murphy et al., 1999], [Li and Zhang, 1998], [Shen, 2000] and [Cosslett, 2007] among others.

We define a truncated score function

$$\psi_n^{(\epsilon)}(\beta, F) = \int_{F(t-\beta' x)\in[\epsilon,1-\epsilon]} [\phi(t,x,\delta)\{F(t-\beta' x)-\delta\}]\, d\mathbb{P}_n(t,x,\delta), \tag{3}$$

where $\epsilon \in (0, 1/2)$ is a truncation parameter and ϕ is some weight function. In this research, we consider estimates of β_0, obtained by solving the score equation

$$\psi_n^{(\epsilon)}(\beta, \hat{F}) = 0,$$

for some estimate $\hat{F}$ of F. Truncation is used to avoid theoretical and numerical difficulties. If one starts with the *efficient* score equation or an estimate thereof, the solution sometimes suggested in the literature, is to add a constant c_n, tending to zero as $n \to \infty$, to the factor $F(t-\beta' x)\{1-F(t-\beta' x)\}$ which inevitably will appear in the denominator. This is done in, e.g. [Li and Zhang, 1998]; similar ideas involving a sequence (c_n) are used in [Klein and Spady, 1993] and [Cosslett, 2007]. Picking a suitable sequence is more tricky, though, than just using the simple device in (3). It is perhaps somewhat remarkable that we can, instead of letting $\epsilon \downarrow 0$, fix $\epsilon > 0$ and still have consistency of our estimators; on the other hand, the estimate proposed by [Murphy et al., 1999] is also identified via a subset of the support of the distribution F_0, since their assumptions imply that F_0 stays strictly away from 0 and 1 on the support of the density $f_{T-\beta X}$ (see above). The drawback of this assumption is that we have no information about the whole distribution F_0.

We consider three different weight functions ϕ in (3). The first score function yields a simple estimate of β_0 based on the MLE $\hat{F}_{n,\beta}$ for the distribution function F in (2), without requiring any smoothing technique, i.e. we consider

$$\psi_{1,n}^{(\epsilon)}(\beta) = \int_{\hat{F}_{n,\beta}(t-\beta' x)\in[\epsilon,1-\epsilon]} x\big\{\hat{F}_{n,\beta}(t-\beta' x)-\delta\big\}\, d\mathbb{P}_n(t,x,\delta), \tag{4}$$

where $\hat{F}_{n,\beta}$ is the MLE based on the order statistics of the values $T_i - \beta' X_i, i = 1, \ldots, n$.

The second score function adds a smooth density estimate $f_{n,h}$ to the first score function resulting in an efficient estimation algorithm for β_0 involving the MLE $\hat{F}_{n,\beta}$. This second score function is defined by,

$$\begin{aligned}\psi_{2,nh}^{(\epsilon)}(\beta) = \int_{\hat{F}_{n,\beta}(t-\beta' x)\in[\epsilon,1-\epsilon]} & \frac{x f_{nh,\beta}(t-\beta' x)}{\hat{F}_{n,\beta}(t-\beta' x)\{1-\hat{F}_{n,\beta}(t-\beta' x)\}} \\ & \cdot \{\hat{F}_{n,\beta}(t-\beta' x)-\delta\}\, d\mathbb{P}_n(t,x,\delta),\end{aligned} \tag{5}$$

where $f_{nh,\beta}$ is a density estimate defined by,

$$f_{nh,\beta}(t-\beta' x) = \int K_h(t-\beta' x - w)\, d\hat{F}_{n,\beta}(w).$$

with K_h being a second order kernel function with bandwidth h (see [Groeneboom et al., 2010]).

The last approach uses a kernel estimate for the distribution function F and therefore no longer involves the MLE $\hat{F}_{n,\beta}$. More precisely, we define the plug-in estimate

$$F_{nh,\beta}(t-\beta' x) = \frac{\int \delta K_h(t-\beta' x - u + \beta' y)\, d\mathbb{P}_n(u,y,\delta)}{\int K_h(t-\beta' x - u + \beta' y)\, d\mathbb{G}_n(u,y)}, \tag{6}$$

where $\mathbb{G}_n$ is the empirical distribution function of the pairs (T_i, X_i). An asymptotic representation of the partial derivatives w.r.t. β_k $(k = 1, \ldots, d)$ of the profile log likelihood, defined in (2), is then given by the third score function defined by,

$$\psi^{(\epsilon)}_{3,nh}(\beta) = \int_{F_{nh,\beta}(t-\beta' x)\in[\epsilon,1-\epsilon]} \frac{\partial_\beta F_{nh,\beta}(t-\beta' x)}{F_{nh,\beta}(t-\beta' x)\{1-F_{nh,\beta}(t-\beta' x)\}} \cdot \{F_{nh,\beta}(t-\beta' x) - \delta\}\, d\mathbb{P}_n(t,x,\delta), \tag{7}$$

The estimate based on the first score function (4) is proved to be a $\sqrt{n}-$consistent but inefficient estimate of the regression parameter. Note that this estimate does not involve any smoothing techniques in the expression of the score function (4) and is only based on the piecewise constant MLE of F. In the second score function (5) we incorporate an estimate of the density, based on the MLE, which results in an extension of the first estimate that is an efficient estimate of the regression parameter. The last estimate, based on the third score function (7), has the same asymptotic distribution as the second estimate but does no longer involve the MLE $\hat{F}_{n,\beta}$ in its derivation. Both second and third estimates are $\sqrt{n}$-consistent and asymptotically normal with an asymptotic variance that is arbitrarily (determined by the truncation device) close to the information lower bound.

References

[Cosslett, 1987] Cosslett, S. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica*, 55(3):559–585.

[Cosslett, 1983] Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, 51(3):765–782.

[Cosslett, 2007] Cosslett, S. R. (2007). Efficient estimation of semiparametric models by smoothed maximum likelihood. *Internat. Econom. Rev.*, 48(4):1245–1272.

[Groeneboom et al., 2010] Groeneboom, P., Jongbloed, G., and Witte, B. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *Ann. Statist.*, 38:352–387.

[Huang and Wellner, 1993] Huang, J. and Wellner, J. (1993). Regression models with interval censoring. *Proceedings of the Kolmogorov Seminar, Euler Mathematics Institute. St. Petersburg, Russia.*

[Klein and Spady, 1993] Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421.

[Li and Zhang, 1998] Li, G. and Zhang, C.-H. (1998). Linear regression with interval censored data. *Ann. Statist.*, 26(4):1306–1327.

[Murphy et al., 1999] Murphy, S. A., van der Vaart, A. W., and Wellner, J. A. (1999). Current status regression. *Math. Methods Statist.*, 8(3):407–425.

[Shen, 2000] Shen, X. (2000). Linear regression with current status data. *J. Amer. Statist. Assoc.*, 95(451):842–852.

Nonparametric estimation of surface flow lines

Armin Schwartzman
(joint work with Alison Wu)

1. Problem setup

A flow line is an integral curve of the vector field defined by the gradient of a scalar potential. Specifically, if $F : \mathbb{R}^N \to \mathbb{R}$ is a differentiable scalar potential with gradient ∇F, then a flow line $\mathbf{x} : \mathbb{R} \to \mathbb{R}^N$ is a parametrized curve such that

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla F[\mathbf{x}(t)], \tag{1}$$

the negative sign simply specifying a flow from higher to lower potential. If the potential is a topographical function representing elevation of the Earth's surface in a given geographical region, the flow line starting from a fixed initial location approximates how water flows down from that location when pulled by gravity.

The motivation for this work is not the flow of water but that of ice. Mountain glaciers flow down mountain valleys at high altitudes and their shapes often follow water flow lines. Given the widely reported shrinking of mountain glaciers worldwide [1, 2], there is interest in studying the length of a glacier along its flow line in order to estimate its retreat, or advance, over time. A digital elevation model (DEM) describing the surface of the Earth surrounding the location of a mountain glacier may be obtained from Google Earth (Figure 1a).

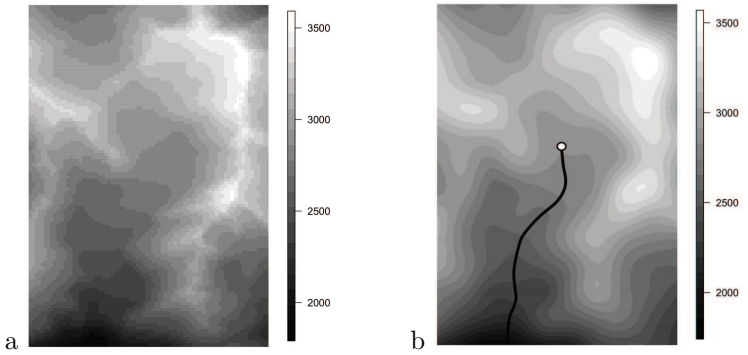


Figure 1. (a) DEM for Rhone Glacier in the Swiss Alps (elevation in meters). (b) Smoothed DEM using bivariate splines; solid line is the estimated flow line by gradient descent starting from the location marked by the circle.

The problem of finding flow lines of smooth surfaces has been well studied in differential geometry [3]. This setting typically requires the surface to be smooth and known over a fixed domain. Instead, DEMs are measured with error and are digitally quantized to a fixed rounding precision, resulting in discontinuous

step functions (Figure 1a). From this point of view, a feasible approach is to first smooth the DEM by a suitable nonparametric estimator, e.g. using bivariate splines. Then estimates of integral curves can be found numerically. If $\widehat{\nabla F}$ represents the estimated gradient, obtained by differencing of the smoothed surface, and $\mathbf{u} = \widehat{\nabla F}/|\widehat{\nabla F}|$ is the gradient field normalized to unit length, then a discrete implementation of (1) yields the iterative solution

$$\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}(k-1) - s \cdot \mathbf{u}[\hat{\mathbf{x}}(k)], \qquad k = 0, 1, \ldots \tag{2}$$

from an initial location $\hat{\mathbf{x}}(0)$, where s is a user-defined parameter (Figure 1b).

Because the regions analyzed may be large, containing in the order of 10^5 pixels, applying a smoothing operator to the entire DEM may be computationally time consuming and difficult to scale when processing DEMs for many mountain glaciers. For this reason, the goal is to devise a way to estimate the flow line directly from the unsmoothed surface.

2. Gradient descent by local linear regression

The iterative procedure (2) is in essence a gradient descent algorithm, not unlike gradient descent algorithms used to solve minimization problems [4], except that its objective is to estimate a path, rather than just arrive at the minimum. As such, it is a greedy algorithm that only depends on the data in the immediate vicinity of the current flow line point $\hat{\mathbf{x}}(k)$; effort spent to smooth the DEM far from the flow line does not contribute to the local estimation of the flow line.

Based on this observation, the proposed idea is to smooth the DEM locally at each iteration of the algorithm. For a fine enough pixel grid, a local linear approximation may be sufficient. We call this gradient descent by local linear regression. At each iteration k:

(1) A plane is fitted by weighted least squares to the observed data with weights given by a kernel function centered the current flow line point $\hat{\mathbf{x}}(k)$.
(2) The gradient ∇F at $\hat{\mathbf{x}}(k)$ is estimated as the gradient of the fitted plane.
(3) The flow line is updated via equation (2).

The algorithm above is closely related to two other algorithms known in statistics and computer science. First, the linear fit in Step 1 above is a multivariate version of the nonparametric estimation method of local linear regression [5, 6]. As in univariate local linear regression, a smooth solution with fast computation can be obtained using a smooth kernel function with finite support, such as the multivariate Epanechnikov kernel.

Second, the advance in the direction of the gradient is akin to the mean shift algorithm for finding modes of multivariate probability densities [7, 8, 9]. At each iteration, the mean shift algorithm moves in the direction of the density gradient, converging to a local maximum of the density along the steepest ascent path. The flow line algorithm can thus be seen as an extension of the mean shift algorithm to regression surfaces instead of densities.

3. Open questions

The gradient descent by local linear regression algorithm is fast and appears to recover the true flow line in artificial simulations. However, the questions of consistency and selection of the kernel bandwidth selection remain open.

Studies of consistency of the mean shift algorithm have focused on convergence to the mode of the density, rather than the path taken to get there. Here, we wish to show that the algorithm can consistently estimate the flow line as pixel resolution increases, similar to [10]. A suitable definition of error between the estimated curve $\hat{\mathbf{x}}$ and the true integral curve $\mathbf{x}$ is the L_2 distance defined via

$$d^2(\hat{\mathbf{x}}, \mathbf{x}) = \int_0^1 \left[\hat{\mathbf{x}}(t) - \mathbf{x}(t)\right]^2 \, dt, \tag{3}$$

where both curves are parametrized so that both begin at $t = 0$, end at $t = 1$, and traverse the length of the curve at a constant speed. This distance function increases both with the departure of the curves from each other and with their discrepancy in length.

The main difficulty in handling (3) is that the target integral curve cannot be written explicitly but is only characterized by the fact that it satisfies (1). It also makes difficult to devise data-dependent bandwidth selection procedures. In local linear regression, the method of cross-validation is based on the ability to estimate the error by comparing the estimated function to the observed values. Because the integral curve is not directly observed but is only a functional of the observed surface, it is unclear how to estimate the error (3) from the data in a way similar to cross-validation.

References

[1] J. Oerlemans, *Holocene glacier fluctuations: is the current rate of retreat exceptional?*, Annals of Glaciology, International Glaciological Society **31** (2000), 39–44.

[2] M.S. Pelto, *Alpine Glaciers and Ice Sheets [in State of the Climate in 2015]*, Bulletin of the American Meteorological Society **97**(8) (2016), S23–S24.

[3] S. Lang, *Differential and Riemannian manifolds*, Springer-Verlag New York (1995).

[4] J. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, Springer Publishing (2005).

[5] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer (2001).

[6] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics (2006).

[7] K. Fukunaga and L.D. Hostetler, *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE T. Inform. Theory **21** (1975), 32–40.

[8] D. Comaniciu and P. Meer, *Mean Shift: A Robust Approach Toward Feature Space Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5) (2002), 603–619.

[9] J. Chacon and T. Duong, *Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting*, Electronic Journal of Statistics **7** (2013), 499–532.

[10] V. Koltchinskii, L. Sakhanenko and S. Cai, *Integral cruves of noisy vector fields and statistical problems in diffusion tensor imaging: nonparametric kernel estimation and hypotheses testing*, The Annals of Statistics **35**(4) (2007), 1576–1607.

Divide and Conquer in Non-Standard Problems and the Super-efficiency Phenomenon

Bodhisattva Sen
(joint work with Cecile Durot, Moulinath Banerjee)

We study how the divide and conquer principle - partition the available data into subsamples, compute an estimate from each subsample and combine these appropriately to form the final estimator - works in non-standard problems where rates of convergence are typically slower than $\sqrt{n}$ and limit distributions are non-Gaussian, with a special emphasis on the least squares estimator of a monotone regression function. We find that the pooled estimator, obtained by averaging non-standard estimates across the mutually exclusive subsamples, outperforms the non-standard estimator based on the entire sample in the sense of pointwise inference. We also show that, under appropriate conditions, if the number of subsamples is allowed to increase at appropriate rates, the pooled estimator is asymptotically normally distributed with a variance that is empirically estimable from the subsample-level estimates. Further, in the context of monotone function estimation we show that this gain in pointwise endciency comes at a price - the pooled estimator's performance, in a uniform sense (maximal risk) over a class of models worsens as the number of subsamples increases, leading to a version of the super-endciency phenomenon. In the process, we develop analytical results for the order of the bias in isotonic regression for the first time in the literature, which are of independent interest.

Distributional Limits for Wasserstein Distance on Discrete Spaces

Max Sommerfeld
(joint work with Axel Munk)

The empirical Wasserstein distance between distributions is an attractive tool for statistical applications but suffers from two major obstacles: First, inference is hindered by the lack of distributional limits for spaces other than the real line. Second, the computational cost is prohibitive even for moderately sized problems. We argue that both obstacles can be overcome in the setting of finite spaces. To this end, for probability measures supported on finitely many points, we derive the asymptotic distribution of the Wasserstein distance of empirical distributions as the optimal value of a linear program with random objective function. As a consequence statistical inference for sample based Wasserstein distances becomes doable in large generality. We introduce the concept of directional Hadamard differentiability in this context. To approximate the limiting distribution, we discuss bootstrapping schemes accounting for the non-linear derivative of the Wasserstein distance and explore modifications that reduce the computational burden.

Nevertheless, when problem sizes become large, exact computation of the Wasserstein distance as well as the bootstrap become computationally infeasible. To

facilitate inference (e.g. testing) in these situations, we lower bound the Wasserstein distance and stochastically upper bound the limiting distribution using tree metrics and give efficient algorithms to compute these bounds.

Adaptation in log-concave density estimation

Richard J. Samworth
(joint work with Arlene Kyoung Hee Kim and Aditya Guntuboyina)

Shape constraints have recently found successful application in many different statistical problems, including convex regression [Seijo and Sen (2011)], generalised additive models [Chen and Samworth (2016)], independent component analysis [Samworth and Yuan (2012)] and many others. This has led to intensive efforts to understand the theoretical properties of shape-constrained estimators. In some cases, it is now known that they can achieve minimax optimal rates of convergence; see, e.g., [Birgé (1987)] for the Grenander estimator, [Baraud and Birgé (2016)] for ρ-estimators and [Han and Wellner (2016)] for convex regression estimators. However, the fact that these estimators are tuning-free raises the prospect that they might adapt to certain data generating mechanisms in the sense of attaining a faster rate of convergence than that predicted by the 'worst-case' minimax theory.

In this work we explore this adaptation phenomenon in the context of log-concave density estimation. Say a density f on $\mathbb{R}$ is log-concave if $f = e^\phi$ for some concave $\phi : \mathbb{R} \to [-\infty, \infty)$. Write $\mathcal{F}$ for the set of all log-concave densities. Very recently, [Kim and Samworth (2016)] proved that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_0 \in \mathcal{F}$, then

$$\inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}} \mathbb{E}_{f_0} d^2_{\mathrm{H}}(\tilde{f}_n, f_0) \asymp n^{-4/5},$$

and the log-concave maximum likelihood estimator (MLE) $\hat{f}_n$ based on $X_1, \ldots, X_n$ attains this minimax optimal rate. Here, the infimum is over all estimators $\tilde{f}_n$ of f_0, and $d^2_{\mathrm{H}}(f, g) := \int_{-\infty}^{\infty} (f^{1/2} - g^{1/2})^2$ denotes the squared Hellinger distance. Let

$$d_{\mathrm{TV}}(f, g) := \frac{1}{2} \int_{-\infty}^{\infty} |f - g|, \quad \text{and} \quad d^2_{\mathrm{KL}}(f, g) := \int_{-\infty}^{\infty} f \log \frac{f}{g},$$

denote the total variation distance and Kullback–Leibler divergence respectively, and recall the standard inequalities $d^2_{\mathrm{TV}}(f, g) \le d^2_{\mathrm{H}}(f, g) \le d^2_{\mathrm{KL}}(f, g)$. We will also be interested in another notion of divergence: by an application of Remark 2.3 of [Dümbgen et al. (2011)] to the function $x \mapsto \log \frac{f_0(x)}{\hat{f}_n(x)}$,

$$d^2_{\mathrm{KL}}(\hat{f}_n, f_0) \le \frac{1}{n} \sum_{i=1}^{n} \log \frac{\hat{f}_n(X_i)}{f_0(X_i)} =: d^2_X(\hat{f}_n, f_0).$$

If $\log f_0$ is composed of a relatively small number of affine pieces, then we might expect $\hat{f}_n$ to converge to f_0 at an especially fast rate. To this end, for $k \in \mathbb{N}$, say $f \in \mathcal{F}$ belongs to $\mathcal{F}^k$ if $\log f$ is *k-affine* in the sense that there exist intervals $I_1, \ldots, I_k$ such that f is supported on $I_1 \cup \ldots \cup I_k$, and $\log f$ is affine on each I_j.

1. Rates for densities that are close to log-affine on their support

Let $\mathcal{T}_0 := \{(s_1, s_2) \in \mathbb{R}^2 : s_1 < s_2\}$ and

$$\mathcal{T} := (\mathbb{R} \times \mathcal{T}_0) \bigcup \big((0,\infty) \times \{-\infty\} \times \mathbb{R}\big) \bigcup \big((-\infty, 0) \times \mathbb{R} \times \{\infty\}\big).$$

Now, for $(\alpha, s_1, s_2) \in \mathcal{T}$, let

$$f_{\alpha,s_1,s_2}(x) := \begin{cases} \frac{1}{s_2-s_1} \mathbb{1}_{\{x\in[s_1,s_2]\}} & \text{if } \alpha = 0 \\ \frac{\alpha}{e^{\alpha s_2} - e^{\alpha s_1}} e^{\alpha x} \mathbb{1}_{\{x\in[s_1,s_2]\}} & \text{if } \alpha \neq 0, \end{cases}$$

so $\mathcal{F}^1 = \{f_{\alpha,s_1,s_2} : (\alpha, s_1, s_2) \in \mathcal{T}\}$. Define $\rho : \mathbb{R} \to [0,1]$ by

$$\rho(x) := \begin{cases} \frac{2x-2-e^{-x}(x^2-2)}{2-e^{-x}(x^2+2x+2)} & \text{for } x \neq 0 \\ 2 & \text{for } x = 0, \end{cases} \tag{1}$$

so ρ is continuous and increasing with $\rho(x) \leq \max\{\rho(2), \rho(x)\} \leq \max(3, 2x)$.

Theorem 1. *Let f_0 be any density on $\mathbb{R}$, let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_0$ for some $n \geq 5$, and let $\hat{f}_n$ denote the corresponding log-concave MLE. Fix any $f_{\alpha,s_1,s_2} \in \mathcal{F}^1$, write $\kappa^* := \alpha(s_2 - s_1)$, let $d_{\mathrm{TV}} := d_{\mathrm{TV}}(f_{\alpha,s_1,s_2}, f_0)$ and let $d_{\mathrm{KS}}^{(n)} := \|F^n_{\alpha,s_1,s_2} - F_0^n\|_\infty + \|(1-F_{\alpha,s_1,s_2})^n - (1-F_0)^n\|_\infty$, where F_{α,s_1,s_2} and F_0 are the distribution functions corresponding to f_{α,s_1,s_2} and f_0 respectively. Then*

$$\mathbb{E}_{f_0} d_{\mathrm{TV}}(\hat{f}_n, f_0) \leq \inf_{f_{\alpha,s_1,s_2} \in \mathcal{F}^1} \left\{ \frac{c_n}{n^{1/2}} + (1 + c_n) d_{\mathrm{TV}} + d_{\mathrm{KS}}^{(n)} \right\}, \tag{2}$$

where $c_n = c_n(f_{\alpha,s_1,s_2}) := \min\{2\rho(|\kappa^*|), 6 \log n\}$.

If $f_0 = f_{\alpha,s_1,s_2} \in \mathcal{F}^1$, then $d_{\mathrm{TV}} = d_{\mathrm{KS}}^{(n)} = 0$, so provided $|\kappa^*| = |\alpha|(s_2 - s_1)$ is not too large, the first term in the minimum in the definition of c_n guarantees that $\hat{f}_n$ attains the parametric rate of convergence. In particular, if $f_0 \in \mathcal{F}^1$ is a uniform density on a compact interval, then we may take $\alpha = 0 = \kappa^*$, and find that

$$\mathbb{E}_{f_0} d_{\mathrm{TV}}(\hat{f}_n, f_0) \leq \frac{4}{n^{1/2}}.$$

If f_0 is any density such that $\inf_{f_{\alpha,s_1,s_2} \in \mathcal{F}^1}(d_{\mathrm{TV}} + d_{\mathrm{KS}}^{(n)}) = o(n^{-2/5} \log^{-1} n)$, then the rate provided by (2) is faster than that given by the worst-case minimax theory. The proof of Theorem 1 is crucially based on the following analogue of the classical Marshall's inequality for decreasing density estimation [Marshall (1970)].

Lemma 1. *Let $n \geq 2$, let $X_1, \ldots, X_n$ be real numbers that are not all equal, with empirical distribution function $\mathbb{F}_n$, and let $\hat{f}_n$ denote the corresponding log-concave MLE. Let $X_{(1)} := \min_i X_i$ and $X_{(n)} := \max_i X_i$. Let f_0 be a density such that $f_0(x) = e^{\alpha_0 x} h_0(x)$ for $x \in [X_{(1)}, X_{(n)}]$, where $\alpha_0 \in \mathbb{R}$ and $h_0 : [X_{(1)}, X_{(n)}] \to \mathbb{R}$ is concave, and let $\kappa := \alpha_0(X_{(n)} - X_{(1)})$. Writing F_0 and $\hat{F}_n$ for the distribution functions corresponding to f_0 and $\hat{f}_n$ respectively, we have*

$$\|\hat{F}_n - F_0\|_\infty \leq \rho(|\kappa|) \|\mathbb{F}_n - F_0\|_\infty. \tag{3}$$

The original Marshall's inequality applies to the integrated Grenander estimator when F_0 is concave; in that case, $\rho(|\kappa|)$ can be replaced with 1. [Dümbgen et al. (2007)] proved a similar result for the integrated version of the least squares estimator of a convex density on $[0,\infty)$; there, a multiplicative constant 2 is needed. When f_0 is concave on the convex hull of the data, we can take $\alpha_0 = 0 = \kappa$, and the multiplicative constant in Lemma 1 can also be taken to be 2.

2. Rates for densities whose logarithms are close to k-affine

Our main result on adaptation over $\mathcal{F}^k$ is the following sharp oracle inequality:

Theorem 2. *There exists a universal constant $C > 0$ such that for every $n \geq 2$ and every $f_0 \in \mathcal{F}$, we have*

$$\mathbb{E}_{f_0} d^2_{\mathrm{KL}}(\hat{f}_n, f_0) \leq \inf_{k\in\mathbb{N}} \left\{ \frac{Ck}{n} \log^{5/4} n + \inf_{f_k \in \mathcal{F}^k} d^2_{\mathrm{KL}}(f_0, f_k) \right\}. \tag{4}$$

A consequence of Theorem 2 is that if $\log f_0$ is close to k-affine for some k in that $\inf_{f_k\in\mathcal{F}^k} d^2_{\mathrm{KL}}(f_0, f_k) = O\big(\frac{k}{n}\log^{5/4} n\big)$, then $\hat{f}_n$ converges to f_0 at rate $O\big(\frac{k}{n}\log^{5/4} n\big)$, which is almost the parametric rate when k is small. In particular, provided $k = o(n^{1/5}\log^{-5/4} n)$, the rate provided by Theorem 2 is faster than the minimax rate over all log-concave densities of $O(n^{-4/5})$.

Acknowledgements: This research is supported by an EPSRC Early Career Fellowship and a grant from the Leverhulme Trust. The full paper is available as [Kim, Guntuboyina and Samworth (2016)].

References

[Baraud and Birgé (2016)] Baraud, Y. and Birgé, L. (2016) Rates of convergence of rho-estimators for sets of densities satisfying shape constraints. *Stoch. Proc. Appl., to appear.*

[Birgé (1987)] Birgé, L. (1987) Estimating a density under order restrictions: Nonasymptotic minimax risk. *Ann. Statist.*, **15**, 995–1012.

[Chen and Samworth (2016)] Chen, Y. and Samworth, R. J. (2016) Generalised additive and index models with shape constraints. *J. Roy. Statist. Soc., Ser. B*, **78**, 729–754.

[Dümbgen et al. (2007)] Dümbgen, L., K. Rufibach, J. A. Wellner (2007) Marshall's lemma for convex density estimation. In *Asymptotics: Particles, Processes and Inverse Problems*, pp. 101–107. Institute of Mathematical Statistics.

[Dümbgen et al. (2011)] Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011) Approximation by log-concave distributions with applications to regression. *Ann. Statist.*, **39**, 702–730.

[Han and Wellner (2016)] Han, Q. and Wellner, J. A. (2016) Multivariate convex regression: global risk bounds and adaptation. `http://arxiv.org/abs/1601.06844`.

[Kim, Guntuboyina and Samworth (2016)] Kim, A. K. H., Guntuboyina, A. and Samworth, R. J. (2016) Adaptation in log-concave density estimation. `http://arxiv.org/abs/1609.00861`.

[Kim and Samworth (2016)] Kim, A. K. H. and Samworth, R. J. (2016) Global rates of convergence in log-concave density estimation. *Ann. Statist., to appear.*

[Marshall (1970)] Marshall, A. W. (1970) Discussion of Barlow and van Zwet's paper. In *Nonparametric Techniques in Statistical Inference*. Proceedings of the First International Symposium on Nonparametric Techniques held at Indiana University, June, 1969 (M. L. Puri, ed.) pp. 174–176. Cambridge University Press, Cambridge.

[Samworth and Yuan (2012)] Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.

[Seijo and Sen (2011)] Seijo, E. and Sen, B. (2011) Nonparametric least squares estimation of a multivariate convex regression. *Ann. Statist.*, **39**, 1633–1657.

Elastic Shape Analysis and Shape-Constrained Density Estimation

Anuj Srivastava

1. Shape Analysis Functions and Curves

This abstract has two goals: (1) provide an overview of the *elastic framework* for analyzing shapes of Euclidean curves [1], and (2) an application of these ideas in shape-constrained density estimation.

For shape analysis, we are given a collection of curves $\beta_1, \ldots, \beta_n : [0,1] \to \mathbb{R}^d$, we want to be able to: (1) quantify pairwise differences in their shapes, (2) summarize their shapes treating them as samples from a population, and (3) develop efficient probability models to capture their shapes. The challenge comes from the fact that shape is a property that is invariant to certain transformations – rotation, translation, global scaling, and even parameterizations of curves. Additionally, shape analysis requires solving the difficult problem of registration – the optimal matching of points across curves. Most current methods perform shape analysis in two steps – first, they use some technique to register all the curves of interest, and then, they perform shape analysis of these registered curves. The problem with this approach is that registration part often unrelated to the shape analysis part. Elastic shape analysis is unified comprehensive framework where both registration and shape analysis are performed jointly.

If we use the standard Hilbert structure, resulting from the $\mathbb{L}^2$ norm, for aligning and comparing curves, then then the solution has several limitations. To understand this, let Γ be the group of all boundary-preserving diffeomorphisms from $[0,1]$ to itself. Elements of Γ control registration of curves since for any $t \in [0,1]$, the points $\beta_i(t)$ and $\beta_j(\gamma(t))$ are considered registered. Let the objective function for alignment be: $\inf_{\gamma\in\Gamma} \left(\|\beta_i - \beta_j \circ \gamma\|^2 + \lambda R(\gamma)\right)$, where R is a roughness penalty on γ. Not only is this solution not symmetric in β_i and β_j, but, more importantly, it is not a metric and, hence, cannot contribute in ensuing statistical analysis.

A better solution is to derive a distance d_c that is invariant to the action of Γ in the following way: $d_c(\beta_i, \beta_j) = d_c(\beta_i \circ \gamma, \beta_j \circ \gamma)$ for all $\gamma \in \Gamma$. One such distance can be obtained using a Riemannian elastic metric [2] that satisfies this invariance property. Although the metric is invariant, it turns out to be too complicated to use directly in practice. A way out comes from using a different mathematical representation of functions, as follows. Let F denote the set of absolutely continuous $\mathbb{R}^d$-valued functions on $[0,1]$, and for any $f \in F$ define its square-root velocity function (SRVF) ([2]) to be

$$q : [0,1] \to \mathbb{R}^d, \quad q(t) = \frac{\dot{\beta}(t)}{\sqrt{|\dot{\beta}(t)|}} \,. \tag{1}$$

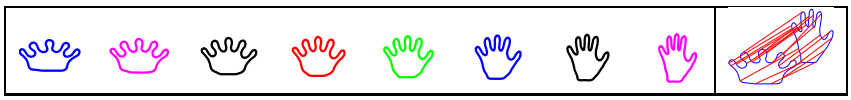


FIGURE 1. A geodesic path between a *crown* shape into a *hand* shape. Right panel shows the optimal registration across curves.

In case of $d = 1$, this expression simplifies to $\text{sign}(\dot{\beta}(t))\sqrt{|\dot{\beta}(t)|}$. It can be shown that if β is absolutely continuous, then q is square-integrable. In fact, the mapping $\beta \mapsto (\beta(0), q)$ is a bijection. The inverse map is given by: $\beta(t) = \beta(0) + \int_0^t |q(s)|q(s)ds$. Furthermore, it can be shown that an elastic Riemannian metric on F becomes the $\mathbb{L}^2$ metric under the change of variables from β to q [2]. Thus, one can use the $\mathbb{L}^2$ norm between SRVFs to register and analyze shapes of curves. Note that if the SRVF of an $\beta \in F$ is $q \in \mathbb{L}^2$, then the SRVF of $(\beta \circ \gamma)$, for any $\gamma \in \Gamma$, is given by $(q, \gamma)(t) = q(\gamma(t))\sqrt{\dot{\gamma}(t)}$.

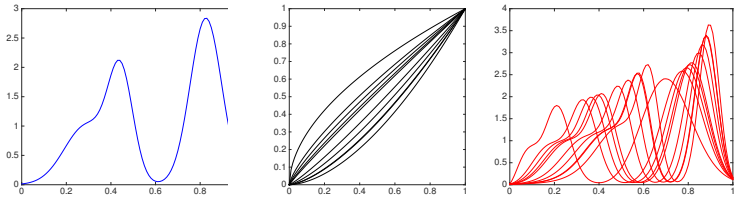
To remove the scale variability, we assume that all curves are of unit length; this implies that for the corresponding SRVF $\|q\| = 1$ or $q \in \mathbb{S}_\infty$. To unify all curves that have the same shape, one defines an equivalence relation with an equivalence class given by: $[q] = \{O(q, \gamma) | \gamma \in \Gamma, O \in SO(d)\} \subset \mathbb{S}_\infty$. Each equivalence class represents a shape uniquely, and the set of such equivalence classes is defined to be the shape space $S = \mathbb{S}_\infty/(SO(d) \times \Gamma)$. With this setup, we can define the elastic shape metric between any two curves β_1 and β_2 to be:

$$d_s([q_1], [q_2]) = \inf_{O \in SO(d), \gamma \in \Gamma} \left(\cos^{-1}\langle q_1, O(q_2, \gamma)\rangle\right) \ . \tag{2}$$

This distance is called the *shape metric* and provides a way for simultaneous registration of points across the two curves. Since d_s is a proper metric, it can be used for computing sample means, sample covariances, and even some kind of principal component analysis of sample curves. Since this framework allows a simultaneous registration of curves, while analyzing their shapes, it is labeled *elastic*. Fig. 1 shows an example of computing elastic geodesic between two similar, yet quite different shapes under this elastic metric. It also shows the optimal correspondence between points across the curves.

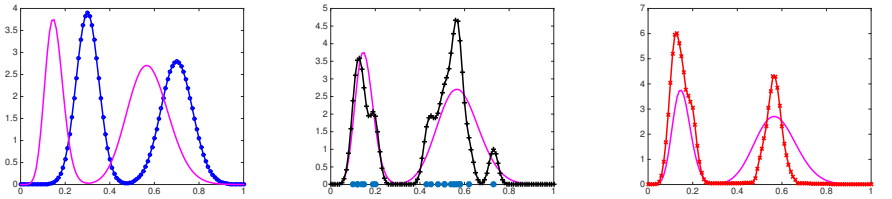
2. Shape-Constrained Density Estimation

Now we outline an application of this framework on a topic that is more central to the theme of this workshop – *shape-constrained density estimation*. This time let F be the set of all non-negative, real-valued, absolutely-continuous functions on $[0, 1]$, and $F_0 \subset F$ be the set of probability density functions (pdfs). While the past work in this area has sought to impose more generic constraints – unimodality, log-concavity, etc, we wish to impose a much stronger shape constraint on the estimated pdf. Furthermore, we wish to do it in a nonparametric way using elastic shape analysis.



$f \in F_0$ $\{\gamma_i \in \Gamma\}$ $\{(f \circ \gamma_i)\}$

FIGURE 2. Examples of pdfs with the same shape.



f_0 (solid), f_1 (marked) $\{X_i \sim f_1\}$, f_1 (solid) , $\hat{f}_k^{(h)}$ (marked) f_1 (solid), $\hat{f}_s$ (marked)

FIGURE 3. An example of shape-constrained density estimation.

The first task, therefore, is to specify the meaning of shape in this context. The shape here will be specified by the counts and heights of peaks and valleys in a function. Therefore, we will consider the time-warping of f as a shape-preserving operation. That is, for any $f \in F$ and $\gamma \in \Gamma$, the composition $f \circ \gamma$ is said to have the same shape as f. (Note that $(f \circ \gamma)$ may not integrate to one, even if $f \in F_0$, but one can easily rescale it to make it an element of F_0.) This equivalence relation partitions F into equivalence classes that are orbits under Γ: $[f] = \{(f \circ \gamma) | \gamma \in \Gamma\}$. Fig. 2 shows examples elements of an equivalence class. Here we take a pdf f (left panel) and apply a number of time-warps γ_is (middle panel) , and the show the resulting (scaled to be pdfs) functions in the right panel. From our perspective, they all have the same shape. Under this setting, let $[f_0]$ denote a shape class of interest. Let $f_1 \in [f_0]$ and let $X_i \sim f_1$ for $i = 1, 2, \ldots, n$ be independent samples. Our goal is to estimate f_1 given $\{X_i\}$ and we pose a maximum-likelihood estimator as: $\hat{f}_s = \text{argmax}_{f \in [f_0]} \sum_{i=1}^{n} \log(f(X_i))$. We approximate this solution using $\hat{f}_s = f_0 \circ \hat{\gamma}$, where $\hat{\gamma} = \text{argmin}_{\gamma \in \Gamma} \| q_k^{(h)} - (q_0, \gamma) \|^2$. Here $q_k^{(h)}$, q_0 are the SRVFs of $\hat{f}_k^{(h)}$ (kernel estimate at a bandwidth h) and f_0, respectively. The estimate $\hat{f}_s$ is found to be stable with respect to h when it is very small. Fig. 3 shows an example of this density estimation using simulated data.

References

[1] A. Srivastava and E. Klassen. *Functional and Shape Data Analysis*. Springer, New York, 2016, expected.

[2] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1415–1428, 2011.

Multiscale feature extraction with applications to classification

Wolfgang Polonik

(joint work with Gabriel Chandler)

1. Introduction

We present a method for feature extraction based on geometric information with the goal of classification. Our method exhibits multiscale characteristics, and has relations to various other methodology known from the literature, including the shorth-plot (see [6]). First we present a brief outline of our proposed methodology. For simplicity consider a binary classification problem, and suppose we have available a training set (X_i, Y_i), $i = 1, \ldots, n$, with $Y_i \in \{0, 1\}$ indicating the class label.

For each pair (X_i, X_j) in d-dimensional Euclidean space, we construct a real-valued function $\widehat{q}_{ij}(\alpha), 0 < \alpha < 1$, which is a quantile function of a certain depth distribution (see below for details). For each given X_i, we then consider a pair of functions $(\widehat{q}_i^{(s)}(\alpha), q_i^{(d)}(\alpha))$ that are formed by averaging $\widehat{q}_{ij}(\alpha)$ over all j with $Y_j = Y_i$ (*s*ame label) and $Y_j \neq Y_i$ (*d*ifferent label), respectively. As a result we have n pairs of functions based on the training set on which we base classification via an FDA methodology to be described below. The following figure shows these functions for the wine data set available at the UC Irvine Machine Learning Repository. One can see different features for functions in different classes (same and between).

2. Constructions of the depth quantile functions $\widehat{q}_{ij}(\alpha)$

For a given pair of data $(X_i, X_j) \in \mathbb{R}^d \times \mathbb{R}^d$ consider the line $\ell_{ij} = \{s \in \mathbb{R}^d : s = \gamma X_i + (1-\gamma)X_j, \gamma \in \mathbb{R}\}$ and the midpoint $m_{ij} = \frac{1}{2}(X_i + X_j)$. For $s \in \ell_{ij}$ and $\alpha \in (0, \pi)$, let $C_{ij}(s)$ denote the cone with tip s and opening angle α containing m_{ij}. Then $\widehat{d}_{ij}(s)$ is the Tukey depth of m_{ij} among all the data on ℓ_{ij} obtained by projecting all the X_j lying in $C_{ij}(s)$ onto ℓ_{ij}. More precisely,

$$\widehat{d}_{ij}(s) = \frac{1}{n} \min \Big\{ \big|\{k : \langle X_k, \tfrac{m_{ij}}{\|m_{ij}\|} \rangle \leq \|m_{ij}\|\big|, \ \big|\{k : \langle X_k, \tfrac{m_{ij}}{\|m_{ij}\|} \rangle \geq \|m_{ij}\|\big| \Big\},$$

so that by our definition the maximum depth of a point is not $\frac{1}{2}$, but rather $\lfloor \frac{n_{ij}(s)}{2} \rfloor$, where $n_{ij}(s)$ is the (random) number of observations in $C_{ij}(s)$.

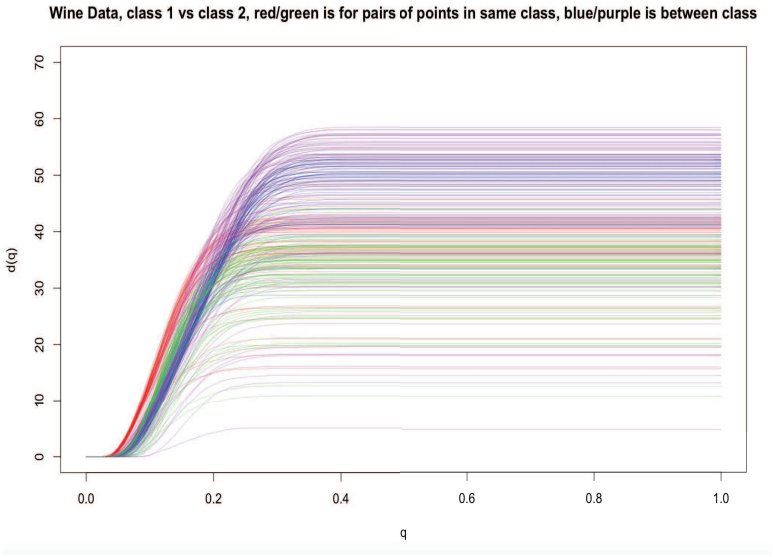


FIGURE 1. Functions $\widehat{q}_i^{(s)}(\alpha)$ for and $\widehat{q}_i^{(d)}(\alpha)$ for X_i in class 1 and class 2 of wine data, respectively.

We now pick the tip s randomly according to a distribution G, independently of the data. This leads to $\binom{n}{2}$ random variables $\widehat{d}_{ij}(S)$, and $\widehat{q}_{ij}(\alpha)$ are defined to be the quantile functions of the distributions of these random variables. As described above, for each given X_i these functions $\widehat{q}_{ij}(\alpha)$ are then averaged over all j in either the same class as X_i, or in a different class, respectively. The resulting pairs of functions $(\widehat{q}_i^{(s)}(\alpha), q_i^{(d)}(\alpha))$ are then used for classification. For that we suggest to use methods from functional data analysis, since they do not require to prespecify features in the functions. One possibility is to simply perform functional PCA on the four classes of functions $Q_k^{(s)} = \{\widehat{q}_i^{(s)}(\alpha), Y_i = k\}$, $k = 1, 2$ and $Q_k^{(d)} = \{\widehat{q}_i^{(d)}(\alpha), Y_i = k\}$, $k = 1, 2$ separately. Combining the scores for each i, results in n vectors of dimension $2k$, and these vectors are then used to train a classifier (such as a kernel SVM), which then is used to classify newly incoming unlabelled data. Other methods will be explored (see [4] and literature cited there).

3. Relations to other methods from the literature

Shorth plot. The shorth plot is proposed in [6] (see also [3]). It is a concentration measure for one-dimensional functions, geared towards mode finding. For $d = 1$, the function $\widehat{q}_{ij}(\alpha)$ can be shown to be closely related to the shorth plot, but rather than mode finding being the goal, our approach is targeting antimodes.

Local depth. Local depth has been considered in the literature, for instance, in [1, 2, 5]. The approach considered in [5] is perhaps the one closest to our approach.

FIGURE 2. First panel shows first two fPCA scores for the two groups of functions, while second panel shows standard PCA scores for original 13-dimensional wine data. It can clearly be seen that the two methods extract different types of information.

They are considering spatial depths of a data point X_j constrained to a ball of size h with midpoint X_j. In other words, rather than using cones they are using balls of varying size, and they are considering spatial depth rather than Tukey depth of the projected data. However, there are several basic methodological differences between our approach and [5]. Our approach is not based on depths of the data points themselves, but on depths of the midpoint between two pairs of points, and we consider, for each midpoint, a map to $\mathbb{R}^2$ rather than a map to $\mathbb{R}$. In fact, [5] only implicitly consider their local depths as a function in h. Out approach provides a clearer heuristic understanding for the information contained in the feature functions. Moreover, in contrast to balls, cones are not local, even though we obtain some localization effect described below. Last, but not least,

our methodology also offers an interpretation via a random set approach. On the whole, all these aspects provide a deeper understanding of the methodology.

4. Some theoretical results

Suppose that $X_1, \ldots, X_n$ form a random sample from F, and that both F and G have positive Lebesgue densities. Let $d_{ij}(s)$ and $q_{ij}(\alpha)$ be the theoretical counterparts to $\widehat{d}_{ij}(s)$ and $\widehat{q}_{ij}(\alpha)$, respectively, meaning that in finding the Tukey depth, the empirical distribution is replaced by the true distribution F. Note, however, that both $d_{ij}(s)$ and $q_{ij}(\alpha)$ are still random quantities, as they still depend on the pair (X_i, X_j).

Theorem 3. *As $n \to \infty$, we have*

$$\sqrt{\frac{n}{\log n}} \max_{1 \le i,j \le n} \sup_{s \in \ell_{ij}} \left| \widehat{d}_{ij}(s) - d_{ij}(s) \right| = O_P(1).$$

Remark. It is worth mentioning that the rate of convergence in this theorem does not depend on the dimension d. The reason for this is, that for each given pair (X_i, X_j) (or for each given line ℓ), the class of cones used in the definition of our (local) depths obtained by varying s, forms a nested class of sets as long as the orientation of the cones does not change. Since for each line we only have two different orientations, this results is a VC-class with index not depending on the dimension.

For a pair (X_i, Y_i), let F_{ij} denote the (one-dimensional) distribution on the line given by (X_i, X_j), obtained by (orthogonally) projecting all the mass of F onto this line. With this notation we have the following result:

Lemma 2. *For any pair (X_i, X_j), we have*

(i) $\lim_{\alpha \to 1} q_{ij}(\alpha) = \min\big(F_{ij}(m_{ij}), 1 - F_{ij}(m_{ij})\big)$ *is the global Tukey depth of m_{ij} for the distribution F_{ij};*

(ii) *Localization:* $\lim_{\alpha \to 0} \frac{q_{ij}(\alpha)}{\alpha^d} \to c \frac{f(m_{ij})}{g(m_{ij})}$, *where $c > 0$ is known, and g is the known pdf of G.*

As for the consistency of our depth quantile functions we have the following result:

Theorem 4. *Fix $1 \le 1 \le j \le n$. Suppose that $H_{ij}(t) = G(d_{ij}(s) \le t | X_i, X_j)$ has an inverse H_{ij}^{-1} that is continuous on a closed interval $\Delta \subset [0, 1]$. Then we have as $n \to \infty$ that*

$$\sup_{\delta \in \Delta} \left| \widehat{q}_{ij}(\delta) - q_{ij}(\delta) \right| = o_P(1).$$

References

[1] Agostinelli, C. and Romanazzi, M. (2011): Local Depth. *J. Statist. Plann. Inference* **141**, 817–830.

[2] Dutta, S., Chaudhuri, P. and Ghosh, A.K. (2012): Classification using Localized Spatial Depth with Multiple Localization, Technical report.

[3] Einmahl, J.H.J., Gantner, M. and Sawitzki, G. (2010): The Shorth Plot. *J. Comput. Graph. Statist.* **19**, 62–73.
[4] Jacques, J. and Preda, C. (2014): Model-based clustering for multivariate functional data. *Comp. Statist. Data Anal.* **71**, 92–106.
[5] Peindaveine, D. and van Bever, G. (2012): From Depth to Local Depth: A Focus on Centrality. *Technical report*
[6] Sawitzki, G (1994): Diagnostic Plots for One-Dimensional Data. In: *Computational Statistics. Papers collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg.* (Edited by P.Dirschedl & R.Ostermann) Heidelberg, Physica, pp. 237–258.

Smooth estimation of a monotone baseline hazard in the Cox model

Eni Musta
(joint work with Hendrik P. Lopuhaä)

The semi-parametric Cox regression model is a very popular method in survival analysis that allows incorporation of covariates when studying lifetimes distributions in the presence of right censored data. Within the Cox model (see [2]), the conditional hazard rate $\lambda(x|z)$ for a subject with covariate vector $z \in \mathbb{R}^p$, is related to the corresponding covariate by

$$\lambda(x|z) = \lambda_0(x)\, e^{\beta_0' z}, \quad x \in \mathbb{R}^+,$$

where λ_0 represents the baseline hazard function, corresponding to a subject with $z = 0$, and $\beta_0 \in \mathbb{R}^p$ is the vector of the regression coefficients.

Let $X_1, \ldots, X_n$ be an i.i.d. sample representing the survival times of n individuals, which can be observed only on time intervals $[0, C_i]$ for some i.i.d. censoring times $C_1, \ldots, C_n$. The observations consists of i.i.d. triplets $(T_1, \Delta_1, Z_1), \ldots, (T_n, \Delta_n, Z_n)$, where $T_i = \min(X_i, C_i)$ denotes the follow up time, $\Delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$ is the censoring indicator, and $Z_i \in \mathbb{R}^p$ is a time independent covariate vector. Given the covariate vector Z, the event time X and the censoring time C are assumed to be independent.

The regression coefficients β_0 can be estimated by the maximum partial likelihood estimator while leaving the baseline distribution unspecified (see e.g. [3], [13]). The focus of this talk is the estimation of the baseline hazard λ_0. Although the most attractive property of this approach is that it does not assume any fixed shape on the hazard curve, there are several cases when one would like to impose order restrictions to better match the practical expectations (see e.g. [6] for an example of a decreasing hazard in large clinical trial for patients with acute coronary syndrome).

Estimation of the baseline hazard function under monotonicity constraints is considered in [1] and [10]. Traditional isotonic estimators, such as the maximum likelihood estimator and Grenander-type estimator, proposed by [10], are step functions and exhibit a non normal limit distribution at rate $n^{1/3}$. On the other hand, a long stream of research (see for instance, [11], [8] or Chapter 8 in [7]) has shown that, if one is willing to assume more regularity on the function of interest,

smooth estimators are preferred to piecewise constant ones because they can be used to achieve a faster rate of convergence to a Gaussian distribution and to estimate derivatives. In particular, kernel smoothing is a rather simple and broadly used technique. However, depending on the order of smoothing isotonization, different estimators can be obtained. In [12], three smooth and monotone estimators are introduced. Then naturally, the question arises whether these estimators exhibit a Gaussian limit distribution at the usual rate $n^{2/5}$. Our main interest is to analyze the asymptotic behavior of such methods.

Asymptotic normality of the smoothed Grenander-type estimator in the ordinary right censoring model without covariates can be easily established by using a Kiefer-Wolfowitz type of result, recently derived in [5] (see [9]). Unfortunately, the lack of a Kiefer-Wolfowitz result (or of an embedding into the Brownian motion) for the Breslow estimator, provides a strong limitation towards extending the previous approach to the more general setting of the Cox model. Therefore, alternative techniques are needed. On the other hand, a different method for finding the limit distribution of smoothed isotonic estimators, which is mainly based on uniform L_2-bounds for the distance between the non-smoothed estimator and the true function, is developed in [7]. However, applying these techniques to the Cox model is much more complicated because instead of the usual exponential bounds for tail probabilities of the inverse process (which is the key step in proving the L_2-bounds) we are able to obtain only polynomial bounds as in Lemma 2 of [4]. Our main result is the following theorem.

Theorem. *Suppose λ_0 is strictly increasing and let the bandwidth be $b = c\,n^{-1/5}$, $(c > 0)$. Then, under additional smoothness assumptions, for any $x \in (0, \tau_H)$,*

$$n^{2/5}\left\{\widetilde{\lambda}_n^{SG}(x) - \lambda_0(x)\right\} \xrightarrow{d} N(\mu, \sigma^2),$$

where

$$\mu = \frac{c^2}{2}\lambda_0''(x)\int u^2 k(u)\,du \quad \text{and} \quad \sigma^2 = \frac{\lambda_0(x)}{c\Phi_0(x)}\int k^2(u)\,du.$$

and

$$\Phi_0(x) = \int e^{\beta_0' z}\,\mathbb{1}_{\{t\geq x\}}\,d\mathbb{P}(t,\delta,z).$$

Once we have the asymptotic normality for the smoothed Grenander-type estimator, following the same lines, we obtain the limit distribution of the smoothed maximum likelihood estimator and the Grenander-type smooth estimator, which turn out to be the same. Moreover, we show also that these three estimators are also proved to be asymptotically equivalent. In addition, we also introduce the maximum smoothed likelihood estimator and then rely on techniques developed in [8]. As expected, this estimator exhibits the same variance as the previous ones but with different asymptotic bias. However, in view of the theoretical results there is no reason to prefer one method with respect to the others.

Finally, a small simulation study on pointwise confidence intervals shows that the four estimators are comparable. We also noticed that confidence intervals constructed using the asymptotic distribution are very much affected by the choice

of the smoothing parameter and usually do not provide good coverage probabilities. On the other hand, bootstrap confidence intervals seem to have a better behavior.

References

[1] D. Chung and M. N. Chang, *An isotonic estimator of the baseline hazard function in Cox's regression model under order restriction*, Statistics & Probability Letters **21** (1994), 223–228.

[2] D. R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society. Series B. Methodological **34** (1972), 187–220.

[3] D. R. Cox, *Partial likelihood*, Biometrika **62** (1975), 269–276.

[4] C. Durot, *On the $\mathbb{L}_p$-error of monotonicity constrained estimators*, The Annals of Statistics **35** (2007), 1080–1104.

[5] C. Durot and H. P. Lopuhaä, *A Kiefer-Wolfowitz type of result in a general setting, with an application to smooth monotone estimation*, Electronic Journal of Statistics **8** (2014), 2479–2513.

[6] N. van Geloven, I. Martin, P. Damman, R. J. de Winter, J. G. Tijssen and H. P. Lopuhaä, *Estimation of a decreasing hazard of patients with acute coronary syndrome*, Statistics in Medicine **32** (2013), 1223–1238.

[7] P. Groeneboom and G. Jongbloed, *Nonparametric estimation under shape constraints*, Cambridge University Press **38** (2014).

[8] P. Groeneboom, G. Jongbloed and W. I. Birgit, *Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model*, The Annals of Statistics **38** (2010), 352–387.

[9] H. P. Lopuhaä and M. Eni, *Smooth estimation of a monotone hazard and a monotone density under random censoring*, ArXiv e-prints **1512.07445** (2015).

[10] H. P. Lopuhaä and G. F. Nane, *Shape constrained non-parametric estimators of the baseline distribution in Cox proportional hazards model*, Scandinavian Journal of Statistics. Theory and Applications **40** (2013), 619–646.

[11] E. Mammen, *Estimating a smooth monotone regression function*, The Annals of Statistics **19** (1991), 724–740.

[12] G. F. Nane, *Shape Constrained Nonparametric Estimation in the Cox Model*, Ph.D. thesis, Delft University of Technology (2013).

[13] A. A. Tsiatis, *A large sample study of Cox's regression model*, The Annals of Statistics **9** (1981), 93–108.

Investigating the Cosmic Web with Persistent Homology

Jessi Cisewski

Data exhibiting complicated spatial structures are common in many areas of science (e.g. cosmology, biology), but can be difficult to analyze. Persistent homology offers a new way to represent, visualize, and interpret complex data by extracting topological features, which can be used to infer properties of the underlying structures.

Persistent homology can be thought of as finding different ordered holes in data where dimension 0 holes are connected components, dimension 1 holes are loops, dimension 2 holes are voids, and so on. The summary diagram is called a "persistence diagram" – a barcode plot conveys the same information in a different way. These topological summaries can be used as inputs in inference tasks (e.g.

hypothesis tests). The randomness in the data due to measurement error or topological noise is transferred to randomness in these topological summaries, which provides an infrastructure for inference. This allows for statistical comparisons between spatially complex datasets.

During the talk, we focused on the problem of analyzing the large-scale structure (LSS) of the Universe, which is an intricate and spatially complex web of matter. In order to understand the physics of the Universe, theoretical and computational cosmologists develop large-scale simulations that allow for visualizing and analyzing the LSS under varying physical assumptions. However, as noted above, rigorous comparisons and inference on such complicated structures can be problematic. Each point in the 3D dataset represents a galaxy or a cluster of galaxies, and persistence diagrams can be obtained summarizing the different ordered holes in the data.

The topological summaries are interesting and informative descriptors of the Universe on their own, but hypothesis tests using the persistence diagrams would provide a way to make more rigorous comparisons of LSS under different theoretical models. For example, in the received cosmological model dark matter is thought to be cold; however, while the case is strong for cold dark matter (CDM) there are some observational inconsistencies with this theory. Another possibility is that dark matter is warm, or warm dark matter (WDM). It is of interest to see if a CDM Universe and WDM Universe produce LSS that is topologically distinct.

We present several possible test statistics for two-sample hypothesis tests using persistence diagrams, carryout a simulation study to investigate the suitableness of the proposed test statistics using simulated data from a variation of the Voronoi foam model, and finally we apply the proposed inference framework to WDM vs. CDM cosmological simulation data.

On Estimation in Tournaments and Graphs under Monotonicity Constraints

Sabyasachi Chatterjee

Abstract. We consider the problem of estimating the probability matrix governing a tournament or linkage in graphs. We assume that the probability matrix satisfies natural monotonicity constraints after being permuted in both rows and columns by the same latent permutation. The minimax rates of estimation for this problem under a mean squared error loss turns out to be $O(1/n)$ upto logarithmic factors. This minimax rate is achieved by the overall least squares estimate which is perhaps impractical to compute because of the need to optimize over the set of all permutations. In this talk, we investigate in detail a simple two stage estimator which is computationally tractable. We prove that the maximum squared error risk of our estimator scales like $O(1/\sqrt{n})$ up to log factors. In addition, we prove an automatic adaptation property of our estimator, meaning that the risk of our estimator scales like $O(1/n)$ upto log factors for several sub classes of our parameter space which are of natural interest. These sub classes include

probability matrices satisfying appropriate notions of smoothness, and subsume the popular Bradley Terry Model in the tournament case and the β model and Stochastic Block Models with monotonicity, in the graph case.

1. Introduction

In this talk we consider two statistical estimation problems. We begin by describing the two set ups.

- Consider the situation of n teams playing in a league tournament where each team plays every other team once. The results of the tournament can be written as a data matrix y of zeroes and ones by setting $y_{ij} = 1$ for $i < j$ if team i wins against team j, and 0 otherwise. Let θ_{ij} be the probability that team i wins against team j with $\theta_{ji} = 1 - \theta_{ij}$ whenever $i \neq j$. Set $\theta_{ii} = 0$ for all $1 \leq i \leq n$ as a matter of convention. The upper triangular part of the data matrix y is modeled as

$$y_{ij} \sim Bern(\theta_{ij}),\ \forall 1 \leq i \leq j \leq n \tag{1}$$

  where y_{ij} in the upper triangular part is jointly independent and $Bern(p)$ refers to the standard Bernoulli distibution with success probability p. The lower triangular part of the data matrix is filled in an antisymmetric manner; that is

$$y_{ij} = 1 - y_{ji},\ \forall 1 \leq j < i \leq n. \tag{2}$$

  We are interested in the problem of estimating the parameter matrix of probabilities θ_{ij} under an assumption commonly made in the ranking literature known as Strong Stochastic Transitivity(SST) (see [4] and reference therein). This assumption posits the existence of an ordering among the teams which is unknown to the statistician. This ordering is then reflected on the probabilities θ_{ij} as follows. Let team j have a higher rank (better) than team k. Then for any team i, the probability of team i defeating team k would be no less than the probability of team i defeating team j, which gives $\theta_{ij} \leq \theta_{ik}$.

  The estimation problem described above was formally introduced in [2] and the model described above was termed as the *Nonparametric Bradley Terry Model*. The terminology is apt because it generalizes the very commonly used Bradley Terry model ubiquitous in the ranking literature (see [1]). In this talk, we also refer to this model as the anti-symmetric model, following the terminology set in [2].
- Consider now the situation of observing a random graph on n nodes with no self loops. Let θ_{ij} now be the probability of node i and node j being linked. Again we set $\theta_{ii} = 0$ for all $1 \leq i \leq n$ as a matter of convention. The random graph can be now encoded as an adjacency matrix y of zeroes and ones. Again, the upper triangular part of the adjacency matrix is modelled as

$$y_{ij} \sim Bern(\theta_{ij})\ \forall\ 1 \leq i \leq j \leq n \tag{3}$$

where y_{ij} in the upper triangular part is jointly independent. The lower triangular part of the data matrix is now filled in a symmetric manner; that is

$$y_{ij} = y_{ji} \ \forall \ 1 \leq j < i \leq n. \tag{4}$$

Inspired by the SST assumption in the ranking literature, here we assume that that the vertices can be arranged in an order (unknown to the statistician) of increasing tendency of getting linked to other vertices. This assumption will again impose monotonicity constraints on the edge probabilities θ_{ij}. For example if node j is more "active" or "popular" than node k then for any node i we must have $\theta_{ik} \leq \theta_{ij}$. For an example where such an assumption seems natural, consider a social network with n people labeled $\{1, 2, \cdots, n\}$ where the i^{th} person has a popularity parameter $p_i \in [0, 1]$. The chance that person i and person j are friends is $f(p_i, p_j)$, where f is increasing in both co-ordinates to signify that increasing popularity leads to more friendship ties. The function f also needs to be symmetric, as the chance that i and j are friends is symmetric in (i, j). Indeed, in this case there is (at least) one ordering which sorts the nodes of the network in increasing order of popularity.

We pose and study the problem of estimating the edge probability matrix θ in the above set up, and we refer to this model of random graphs as the symmetric model, differentiating it from the antisymmetric (tournament) case. Under our model assumptions, the problem of estimating the edge probabilities is very closely related to the problem of estimating graphons in the spirit of [3] where we assume monotonicity (without smoothness) of the graphon in both variables, instead of smoothness assumptions made in [3].

In this talk we look at the two estimation problems in the antisymmetric and the symmetric model in a unified way. In particular, the purpose of this talk is to introduce and study the risk properties of a very natural two step estimator which is described in subsection 1.2. The two step estimator has the same form in both the models and the technique of analyzing the risk properties of the estimator in both the models is the same.

1.1. **Formal Setup of our problem.** In this subsection we define two parameter spaces; one for the antisymmetric model and one for the symmetric model. Let us introduce some notation first. Denote S_n to be the set of all permutations on n symbols. For any $n \times n$ matrix θ and any permutation $\pi \in S_n$ we define $\theta \circ \pi$ to be the $n \times n$ matrix such that $(\theta \circ \pi)_{ij} = \theta_{\pi(i),\pi(j)}$. Let Π be the $n \times n$ permutation matrix corresponding to the permutation $\pi \in S_n$. Then we can also write $\theta \circ \pi = \Pi^T \theta \, \Pi$.

Define the space of matrices

$$\mathcal{T} = \{\theta \in [0,1]^{n \times n} : \theta_{ij} \leq \theta_{ik} \ \forall i < k < j; \quad \theta_{ji} = 1 - \theta_{ij} \ \forall \ i \neq j; \quad \theta_{ii} = 0 \ \forall i\}. \tag{5}$$

Any matrix in $\mathcal{T}$ when only looked at the upper triangular part above the diagonal is non increasing in any row (as j grows) and non decreasing in any column (as i grows). The lower triangular part is just 1 minus the upper triangular part and the diagonals are zero. Basically $\mathcal{T}$ is the space of matrices which satisfy the SST assumption with known ranking where the ranking is such that player n is best, followed by player $n-1$ and so on. Then our parameter space for the antisymmetric model can be written as

$$\Theta_{\mathcal{T}} = \{\theta \circ \pi : \theta \in \mathcal{T}, \pi \in S_n\}. \tag{6}$$

Similarly, define the space of matrices

$$\mathcal{G} = \{\theta \in \mathbb{R}^{n\times n} : \theta_{ij} \le \theta_{ik} \ \forall i < j < k; \quad \theta_{ji} = \theta_{ij} \ \forall i \ne j; \quad \theta_{ii} = 0 \ \forall i\}. \tag{7}$$

Any matrix in $\mathcal{G}$ when only looked at the upper triangular part above the diagonal is non decreasing in both rows and columns. The lower triangular part is symmetrically filled, and the diagonals are zero. Again, $\mathcal{G}$ is the space of adjacency matrices which are consistent with the monotonicity restrictions imposed by the ordering where node n is most popular followed by node $n-1$ and so on. Then our parameter space for the symmetric model can be written as

$$\Theta_{\mathcal{G}} = \{\theta \circ \pi : \theta \in \mathcal{G}, \pi \in S_n\}. \tag{8}$$

For $\Theta = \Theta_{\mathcal{T}}$ or $\Theta_{\mathcal{G}}$ we study the problem of estimating the underlying matrix of probabilities θ. The loss function we consider is the mean Frobenius squared metric defined for any two matrices θ and $\tilde{\theta}$ as follows:

$$\frac{1}{n^2}\|\tilde{\theta} - \theta\|^2 := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\theta_{ij} - \tilde{\theta}_{ij})^2. \tag{9}$$

We will also use the notation $\|A\|$ to denote the Frobenius norm of the matrix A.

1.2. **The estimator.** Our estimator consists of the following steps. The idea is to first get an estimated ranking and then use this ranking to estimate the parameter matrix.

(a) **Sorting**

Let $r_i = \sum_{j=1}^{n} y_{ij}$ be the i^{th} row sum of the data matrix y. In the first step we sort the vertices according to the row sums $(r_1, \ldots, r_n)$ of the data matrix y and obtain a permutation $\hat{\sigma}$ such that $r(\hat{\sigma}(1)) \le \cdots \le r(\hat{\sigma}(n))$, which could be thought of as an estimator of π^{-1}. We call this step the sorting step, which we use to construct a sorted data matrix $y \circ \hat{\sigma}$ defined as $(y \circ \hat{\sigma})_{ij} = (y_{\hat{\sigma}(i)}, y_{\hat{\sigma}(j)})$. In case there are ties in the vector $r = (r_1, \ldots, r_n)$, break the ties uniformly at random while obtaining the permutation $\hat{\sigma}$.

(b) **Projection**

In the second step we project the sorted data matrix $y \circ \hat{\sigma}$ onto the relevant parameter space. In the antisymmetric model, we project onto the set $\mathcal{T} \cap [0,p]^{n\times n}$ and in the symmetric model, we project onto the set $\mathcal{G} \cap [0,p]^{n\times n}$. Both $\mathcal{T}$ and $\mathcal{G}$ are closed convex sets of matrices and hence

there exists a unique projection onto them. Let the projection operator be denoted by P in both cases.

We now unsort the projected and sorted data matrix $P(y \circ \hat{\sigma})$ by applying the permutation $\hat{\sigma}^{-1}$. Finally we divide the resulting matrix by p to obtain our estimator. Thus our final estimator $\hat{\theta}$ for the parameter matrix θ can be written as

$$\hat{\theta} = \frac{1}{p}\big(P(y \circ \hat{\sigma}) \circ \hat{\sigma}^{-1}\big).$$

Note that in the antisymmetric (tournament) model, the row sums of the data matrix y correspond to the number of wins or victories for each player, and the column sums of the data matrix correspond to the number of defeats for each player. Hence our sorting step just sorts the teams according to the number of victories (or equivalently the number of defeats, as sum of victory and defeat of each player is $n-1$). Similarly, in the symmetric (graph) model, the row sums of the adjacency matrix y correspond to the empirical degrees of the nodes in the graph. Therefore, our sorting step sorts the vertices according to the empirical degrees. Our sorting step is thus perhaps the simplest way to get an idea about the underlying latent ranking.

2. Main results

Having defined our estimator, we are now ready to state our main results.

Definition 1. Henceforth we will use the notation Θ in place of $\Theta_{\mathcal{T}}$ or $\Theta_{\mathcal{G}}$. The implication is that all results hold with Θ replaced by either of the two parameter sets.

For any $\theta \in \Theta$ define the row sums $R_i(\theta) := \sum_{j\in[n]} \theta_{ij}$ for $i \in [n]$. Also define the smoothness coefficient of θ by

$$Q(\theta) := \sum_{i\in[n]} \max_{j:|R_j(\theta)-R_i(\theta)|\le 2\sqrt{2n\log n}} \|\theta_{i\cdot} - \theta_{j\cdot}\|^2.$$

Theorem 5. *There is a universal constant $C < \infty$ such that for all $\theta \in \Theta$ we have*

$$\frac{1}{n^2}\mathbb{E}\|\hat{\theta} - \theta\|^2 \le C\Big[\frac{(\log n)^2}{n} + \frac{Q(\theta)}{n^2}\Big]. \tag{10}$$

where C is some universal constant.

The strength of Theorem 5 is that it is adaptive in the parameter θ, and gives tight asymptotic bounds for several sub-parameter spaces of interest. All our results will be derived as corollaries of this theorem.

2.0.1. *Worst case risk.* As a first application of Theorem 5, we deduce the worst case risk of our estimator.

Corollary 1. *There is a universal constant $C < \infty$ such that for all $\theta \in \Theta$ we have*

$$\frac{1}{n^2}\mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\sqrt{\frac{\log n}{n}}, \tag{11}$$

where C is some universal constant.

Proof. To begin, fix $i, j \in [n]$ such that $|R_i(\theta) - R_j(\theta)| \leq 4\sqrt{n \log n}$. Then we have

$$\sum_{k\in[n]}^{n} [\theta_{ik} - \theta_{jk}]^2 \leq \sum_{k\in[n]}^{n} |\theta_{ik} - \theta_{jk}| = |R_i(\theta) - R_j(\theta)| \leq 2\sqrt{2n \log n}.$$

The above bound along with Theorem 5 completes the proof of the lemma. □

2.1. **Adaptation.** Our next theorem reveals automatic adaptation properties of our estimator. Even though our estimator provably achieves $O(1/\sqrt{n})$ rate of estimation globally, it does achieve improved rates of estimation for $\theta \in \Theta$ for subclasses of parameter space of interest.

2.1.1. *Block matrices.*

Definition 2. For $k \in [n]$, let $\Theta^{(k)} \subseteq \Theta$ denote the subset of all $k \times k$ block matrices with equal sized blocks.

Corollary 2. *There exists a universal constant $C < \infty$ such that*

$$\sup_{\theta\in\Theta^{(k)}} \frac{1}{n^2}\mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\Big[\frac{\min(k, \sqrt{n}) \log n}{n} + \frac{(\log n)^2}{n}\Big].$$

for some universal constant $C < \infty$.

2.1.2. *Holder-continuous matrices.*

Definition 3. For $\alpha, L > 0$ let $\Theta(\alpha, L) \subset \Theta$ denote the subset of all Holder continuous matrices in Θ with order α and Holder constant L, i.e. θ satisfies

$$|\theta_{i,j} - \theta_{k,l}| \leq L\Big(\Big|\frac{i-k}{n}\Big|^\alpha + \Big|\frac{j-l}{n}\Big|^\alpha\Big)$$

for all $i, j, k, l \in [n]$.

Corollary 3. *There exists a universal constant $C < \infty$ such that*

$$\sup_{\theta\in\Theta(\alpha,L)} \frac{1}{n^2}\mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\Big[\max(1, L)\frac{\log n}{n^{\frac{2\alpha+1}{2\alpha+2}}} + \frac{(\log n)^2}{n}\Big]$$

Even though the worst case risk over the class of Lipschitz matrices is $\tilde{O}(n^{-5/4})$, if we further assume that θ is lower Lipschitz as well, we get an improved MSE of $\tilde{O}(n^1)$. More generally, the same holds for lower Holder continuous as well. To make this precise we propose the following definition:

Definition 4. For $\alpha, L, L' > 0$ let $\Theta(\alpha, L, L') \subset \Theta(\alpha, L)$ denote the subset of all lower Holder continuous matrices with lower Holder constant L', i.e.

$$|\theta_{i,j} - \theta_{k,l}| \geq L'\Big(\Big|\frac{i-k}{n}\Big|^\alpha + \Big|\frac{j-l}{n}\Big|^\alpha\Big).$$

Corollary 4. *There exists a universal constant $C < \infty$ such that*

$$\sup_{\theta \in \Theta(\alpha, L, L')} \mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\Big[\Big(\frac{L}{L'}\Big)^2 \frac{\log n}{n} + \frac{(\log n)^2}{n}\Big].$$

2.1.3. *Generalized Bradley Terry+Generalized β model.*

Definition 5. Let $\Theta_{\mathcal{H},F,M} \subset \Theta_{\mathcal{H}}$ denote the subset of all tournament matrices such that $\theta_{i,j} = F(w_i - w_j)$, where F is a symmetric distribution function on $\mathbb{R}$ (i.e. $F(x)+F(-x) = 1, \forall x$) with a continuous strictly positive density function function, and $\{w_i\}_{i\in[n]}$ is a sequence of real numbers in $[-M, M]$. This is the generalized Bradley Terry model. In particular, setting $F(x) = \frac{e^x}{1+e^x}$ and $F(x) = \Phi(x)$ ($\Phi(.)$ is the normal distribution function) we get the usual Bradley Terry model and the Thurstone model respectively.

Definition 6. Let $\Theta_{\mathcal{G},F,M} \subset \Theta_{\mathcal{G}}$ denote the subset of all tournament matrices such that $\theta_{i,j} = F(w_i + w_j)$, where F is a distribution function on $\mathbb{R}$ with a continuous strictly positive density function function, and $\{w_i\}_{i\in[n]}$ is a sequence of real numbers in $[-M, M]$. In particular for the choice $F(x) = \frac{e^x}{1+e^x}$ gives the β model on networks, which has originated from Social Sciences and has been studied in Statistics. The class $\Theta_{\mathcal{G},F,M}$ generalizes the usual β model to allow for a more general class of distribution functions.

Corollary 5. *(a) There exists a universal constant $C < \infty$ such that*

$$\sup_{\theta \in \Theta_{\mathcal{H},F,M}} \mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\Big[\Big(\frac{L}{L'}\Big)^2 \frac{\log n}{n} + \frac{(\log n)^2}{n}\Big],$$

where $L := \sup_{|x|\leq 2M} f(x)$ and $L' := \inf_{|x|\leq 2M} f(x)$.

(b) There exists a universal constant $C < \infty$ such that

$$\sup_{\theta \in \Theta_{\mathcal{G},F,M}} \mathbb{E}\|\hat{\theta} - \theta\|^2 \leq C\Big[\Big(\frac{L}{L'}\Big)^2 \frac{\log n}{n} + \frac{(\log n)^2}{n}\Big],$$

where $L := \sup_{|x|\leq 2M} f(x)$ and $L' := \inf_{|x|\leq 2M} f(x)$.

References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[2] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.

[3] Chao Gao, Yu Lu, Harrison H Zhou, et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[4] Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

Dimension Reduction

Stephan Huckemann
(joint work with Benjamin Eltzner)

For reduction of non-Euclidean data to a given dimension m one would ideally use a large (parametric) family of subspaces of dimension m. For a multiscale approach one might even want to have (backward) nested families of such subspaces,

$$\{\mu\} = p^0 \subset p^1 \subset \ldots \subset p^m = Q\,. \tag{1}$$

For the sample space Q being a sphere, [9] proposed *backward nested spheres analysis* giving a sequence of (small) subspheres each of codimension one in one another. Often, the geometry of the sample space does not easily admit such a *high-dimensional* procedure, for instance, if the sample space is a *polysphere*

$$Q = \mathbb{S}^{r_1} \times \ldots \times \mathbb{S}^{r_k}, \quad r_j \in \mathbb{N}, \quad j = 1, \ldots, k, \quad k \in \mathbb{N}\,,$$

as in [10]. Already for the classical torus $\mathbb{T} = \mathbb{S} \times \mathbb{S}$ intrinsic PCA usually leads to dense PCs which cannot be used for statistical purposes.

We propose to change the geometry in a data driven way into that of a single stratified sphere which is identified with itself along some lower dimensional subspheres. For $\mathbb{T}$ this would mean, cutting open along a circle, ideally far away from data, collapsing each circular shore to a point while keeping these two points identified according to the topology of $\mathbb{T}$. This gives the geometry of a two-sphere with north and south pole identified, say.

We argue – endorsed by improved data analysis, cf. [7] – that the geometry of a sphere – possibly with stratifications – is even more suitable than the original one of a polysphere, using *composite principal nested sphere* from [10], or using *tangent space PCA*, i.e. a Euclidean geometry in tangent space. A special case is given in case of *tori* ($r_1 = 1 = \ldots = r_k$ above), which leads to an improved data analysis for RNA secondary structure (i.e. the geometry of RNA backbones, cf. [8]).

In order to investigate asymptotic properties of sequences of subspaces as in (1), where each sequence is viewed as a single descriptor of data or a population, we introduce the more general setup of backward nested families of descriptors.

Definition 7. A separable topological space Q, called the *data space*, admits *backward nested families of descriptors* (BNFDs) if

(i) there is a collection P_j $(j = 0, \ldots, m)$ of topological separable spaces with continuous maps $d_j : P_j \times P_j \to [0, \infty)$ vanishing on the diagonal;
(ii) $P_m = \{Q\}$;
(iii) every $p \in P_j$ $(j = 1, \ldots, m)$ is itself a topological space and gives rise to a topological space $\emptyset \neq S_p \subset P_{j-1}$ which comes with a continuous map

$$\rho_p : p \times S_p \to [0, \infty)\,;$$

(iv) for every pair $p \in P_j$ $(j = 1, \dots, m)$ and $s \in S_p$ there is a measurable map called *projection*

$$\pi_{p,s} : p \to s\,.$$

For $j \in \{1, \dots, m\}$ and $k \in \{1, \dots, j\}$ call a family

$$f = \{p^j, \dots, p^{j-k}\},\ \text{with } p^{l-1} \in S_{p^l}, l = j-k+1, \dots, j$$

a *backwards nested family of descriptors (BNFD) from P_j to P_{j-k}*. The space of all BNFDs from P_j to P_{j-k} is given by

$$T_{j,k} = \left\{ f = \{p^{j-l}\}_{l=0}^{k} : p^{l-1} \in S_{p^l}, l = j-k+1, \dots, j \right\} \subseteq \prod_{l=0}^{k} P_{j-l}\,.$$

For $k \in \{1, \dots, m\}$, given a BNFD $f = \{p^{m-l}\}_{l=0}^{k}$ set

$$\pi_f = \pi_{p^{m-k+1}, p^{m-k}} \circ \dots \circ \pi_{p^m, p^{m-1}} : p^m \to p^{m-k}$$

which projects along each descriptor. For another BNFD $f' = \{p'^{j-l}\}_{l=0}^{k} \in T_{j,k}$ set

$$d^j(f, f') = \sqrt{\sum_{l=0}^{k} d_j(p^{j-l}, p'^{j-l})^2}\,.$$

Definition 8 (Factoring Charts)**.** If $T_{j,k}$ and P^{j-k} carry smooth manifold structures near $f' = (p'^j, \dots, p'^{j-k}) \in T_{j,k}$ and $p'^{j-k} \in P^{j-k}$, respectively, with open $W \subset T_{j,k}$, $U \subset P^{j-k}$ such that $f' \in W$, $p'^{j-k} \in U$ and local charts

$$\psi : W \to \mathbb{R}^{\dim(W)},\ f = (p^j, \dots, p^{j-k}) \mapsto \eta = (\theta, \xi), \quad \phi : U \to \mathbb{R}^{\dim(U)},\ p^{j-k} \mapsto \theta$$

we say that the *chart ψ factors* if with the projections

$$\pi^{P^{j-k}} : T_{j,k} \to P^{j-k},\ f \mapsto p^{j-k}, \quad \pi^{\mathbb{R}^{\dim(U)}} : \mathbb{R}^{\dim(W)} \to \mathbb{R}^{\dim(U)},\ (\theta, \xi) \mapsto \theta$$

we have

$$\phi \circ \pi^{P^{j-k}}|_W = \pi^{\mathbb{R}^{\dim(U)}}|_{\psi(W)} \circ \psi\,.$$

Within this setup we can define analogs of Fréchet means.

Definition 9. Random elements $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} X$ on a data space Q admitting BNFDs give rise to *backward nested population* and *sample means* (abbreviated as BN means)

$$\{E^{f^j} : j = m, \dots, 0\}, \quad \{E_n^{f_n^j} : j = m, \dots, 0\}$$

recursively defined via $f^m = \{Q\} = f_n^m$, i.e. $p^m = Q = p_n^m$ and

$$E^{f^{j-1}} = \underset{s \in S_{p^j}}{argmin}\, \mathbb{E}[\rho_{p^j}(\pi_{f^j} \circ X, s)^2], \qquad f^j = \{p^k\}_{k=j}^{m}$$

$$E_n^{f_n^{j-1}} = \underset{s \in S_{p_n^j}}{argmin} \sum_{i=1}^{n} \rho_{p_n^j}(\pi_{f_n^j} \circ X_i, s)^2, \qquad f_n^j = \{p_n^k\}_{k=j}^{m}\,.$$

where $p^j \in E^{f^j}$ and $p_n^j \in E^{f_n^j}$ is a measurable choice for $j = 1, \ldots, m$.

We say that a BNFD $f = \{p^k\}_{k=0}^m$ gives *unique* BN population means if $E^{f^j} = \{p^j\}$ with $f^j = \{p^k\}_{k=j}^m$ for all $j = 0, \ldots, m$.

Each of the E^{f^j} and $E_n^{f_n^{j-1}}$ is also called a *generalized Fréchet mean*.

Under reasonable assumptions we show strong asymptotic consistency in the sense of [2], which in our scenario translates to the following.

Theorem 6. *Under assumptions specified in* [6], $\{E_n^{f_n^m}, \ldots, E_n^{f_n^k}\}$ *converges a.s. to* $\{p^m, \ldots, p^k\}$ *in the sense that* $\exists \Omega' \subset \Omega$ *measurable with* $\mathbb{P}(\Omega') = 1$ *such that for all* $j = k, \ldots, m$, $\epsilon > 0$ *and* $\omega \in \Omega'$, $\exists N = N(\epsilon, \omega)$ *with*

$$\bigcup_{r=n}^{\infty} E_r^{f_r^j} \subset \{p \in P_j : d_j(p^j, p) \leq \epsilon\} \quad \forall n \geq N,\ \omega \in \Omega' .$$

Also under reasonable assumptions we show asymptotic normality for entire sequences, and under factoring charts also for each single final element.

Theorem 7. *In case of unique BN population means* f'^{k-1} *with* $f'^j = \{p'^{m-1}, \ldots, p'^j\}$ *and BN sample means* $\{E_n^{f_n^{m-1}}, \ldots, E_n^{f_n^{k-1}}\}$ *due to a measurable selection* $p_n^j \in E_n^{f_n^j}$, $f_n^j = \{p_n^{m-1}, \ldots, p_n^j\}$, $j = k-1, \ldots, m-1$, *under assumptions specified in* [6],

(i) $\sqrt{n} H_\psi\big(\psi^{-1}(f_n^{k-1}) - \psi^{-1}(f'^{k-1})\big) \to \mathcal{N}(0, B_\psi)$ *with a suitable chart* ψ *and matrices* H_ψ *and* B_ψ

(ii) *If additionally* $H_\psi > 0$ *then* f_n^{k-1} *satisfies a Gaussian* $\sqrt{n}$*-CLT with*

$$\sqrt{n}\big(\psi^{-1}(f_n^{k-1}) - \psi^{-1}(f'^{k-1})\big) \to \mathcal{N}(0, \Sigma_\psi), \quad \Sigma_\psi = H_\psi^{-1} B_\psi H_\psi^{-1} .$$

(iii) *If additionally the chart* ψ *factors then also* p_n^{k-1} *satisfies a Gaussian* $\sqrt{n}$*-CLT with*

$$\sqrt{n}\big(\phi^{-1}(p_n^{k-1}) - \phi^{-1}(p'^{k-1})\big) \to \mathcal{N}(0, \Sigma_\phi), \quad \Sigma_\phi = \big(\Sigma_{\psi\, ik}\big)_{i,k=1}^{\dim(P_{k-1})}$$

In particular, we show that the "reasonable" assumptions are fulfilled under standard assumptions in the scenario of backward nested great and small spheres as well in case of intrinsic means on first principal geodesic components on manifolds and non-manifold Kendall shape spaces. We point to open problems, as these standard assumptions for manifolds and stratified spaces are still under general investigation, e.g. [4, 3, 5, 1]. Specifically, non-Gaussianity and other asymptotic rates may arise from phenomena of stickiness, possibly due to the stratification (negative curvature), and of smeariness, possibly due to cut loci (positive curvature).

References

[1] Bhattacharya, R. and L. Lin, *Omnibus CLT for Fréchet means and nonparametric inference on non-euclidean spaces*, (2016), to appear.

[2] Bhattacharya, R. N. and V. Patrangenaru, *Large sample theory of intrinsic and extrinsic sample means on manifolds I.*, The Annals of Statistics **31**(1) (2003), 1–29.

[3] Hotz, T. and S. Huckemann, *Intrinsic means on the circle: Uniqueness, locus and asymptotics*, Annals of the Institute of Statistical Mathematics, **67**(1) (2015), 177–193.
[4] Hotz, T., S. Huckemann, H. Le, J. S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer, *Sticky central limit theorems on open books*, Annals of Applied Probability, **23**(6) (2013), 2238–2258.
[5] Huckemann, S., J. C. Mattingly, E. Miller, and J. Nolen, *Sticky central limit theorems at isolated hyperbolic planar singularities*, Electronic Journal of Probability, **20** (2015), paper no. 78
[6] S. F. Huckemann, and Eltzner, B, *Backward nested descriptors asymptotics with inference on stem cell differentiation*, manuscript (2016).
[7] Eltzner, B., S. F. Huckemann, and S. Jung, *Dimension reduction on polyspheres with application to skeletal representations*, Geometric Science of Information 2015 Proceedings (2015), 22–29.
[8] Eltzner, B., S. F. Huckemann, and K. V. Mardia, *Deformed torus PCA with applications to RNA structure*, arXiv:1511.04993 (2015)
[9] Jung, S., I. L. Dryden, and J. S. Marron . *Analysis of principal nested spheres*, Biometrika **99**(3) (2012), 551–568.
[10] Pizer, S. M., S. Jung, D. Goswami, J. Vicory, X. Zhao, R. Chaudhuri, J. N. Damon, S. Huckemann, and J. Marron, *Nested sphere statistics of skeletal models*, In Innovations for Shape Analysis (2013), 93–115. Springer.

Constrained or unconstrained inference in long-range dependent locally stationary processes?

Holger Dette
(joint work with Philip Preuß , Kemal Sen)

Many time series [like asset volatility or regional temperatures] exhibit a slow decay in the sample autocorrelation function and simple stationary short-memory models can not be used to analyze this type of data. However, it was pointed out by several authors that the observation of "long memory" features in the sample autocovariance function can be as well explained by non stationarity and that it is of importance to distinguish between long-memory and non-stationarity [see [7], [9] or [4] among many others].
[6] is the first reference investigating the existence of "long memory" if non-stationarities appear in the time series. In this article a procedure to discriminate between a long-range dependent model and a process with a monotone mean functional and weakly dependent innovations is derived. Later on, [5] developed a method for distinguishing between long-memory and small trends. [8] tested the null hypothesis of a constant long-memory parameter against a break in the long-memory parameter. Furthermore, [3], [2] and [10] investigated CUSUM and likelihood ratio tests to discriminate between the null hypothesis of no long-range and weak dependence with one change point in the mean.

These procedures are only designed to discriminate between long-range dependence and a very specific change in the first-order structure, like one structural break and two stationary segments of the series. This is rather restrictive, and in this talk we fill this gap and to develop present a test for the null hypothesis of no

long-range dependence in a framework which is flexible enough to deal with different types of non-stationarity in both the first and second-order structure. Our approach uses an estimate of a (possibly time varying) long-range dependence parameter, which is derived by a sequence of approximating tvFARIMA models with a slightly enlarged parameter space. This statistic estimates a functional which vanishes if and only if the null hypothesis of a short-memory locally stationary process is satisfied. We prove consistency and asymptotic normality of a corresponding test statistic under the null hypothesis of no long-range dependence. As a consequence we obtain a nonparametric test, which is based on the quantiles of the standard normal distribution and therefore very easy to implement.
The talk is based on [1] and the method utilizes some non-intuitive features of averages of unconstrained estimators in models with a constrained parameter space. In order to make these phenomena visible we provide in the second part of the talk a heuristic motivation of our approach in the context of the classical nonparametric regression model with repeated observations.

References

[1] Dette, H. , Preuß, P. , Sen, K. (2013/2016). "Detecting long-range dependence in non-stationary time series", http://arxiv.org/abs/1312.7452v1

[2] Baek, C., Pipiras, V. (2012). *Statistical tests for a single change in mean against long-range dependence.* Journal of Time Series Analysis 33, 131-151.

[3] Berkes, I., Horvarth, L., Kokoszka, P., Shao, Q. M. (2006). *On discriminating between long-range dependence and changes in mean.* Annals of Statistics 34, 1140-1165.

[4] Chen, Y., Härdle, W., Pigorsch, U. (2010). *Localized Realized Volatility Modeling.* Journal of the American Statistical Association 105(492), 1376–1393.

[5] Heyde, C. C., Dai, W. (1996). *On the robustness to small trends of estimation based on the smoothed periodogram.* Journal of Time Series Analysis 17, 141–150.

[6] Künsch, H. (2004). *Discrimination between monotonic trends and long-range dependence.* Journal of Applied Probality 23, 1025-1030.

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

[8] Sibbertsen, P., Kruse, R. (2009). *Testing for a change in persistence under long-range dependencies.* Journal of Time Series Analysis 30, 263–285.

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

[10] Yau, C. Y., Davis, R. A. (2012). *Likelihood inference for discriminating between long-memory and change-point models.* Journal of Time Series Analysis 33, 649–664.

Volume and Perimeter Estimation Using the Sample α-Shape or the Sample α-Convex Hull

Ery Arias-Castro

(joint work with Beatriz Pateiro-López, Alberto Rodríguez-Casal)

The area of geometric and topological statistics aims at learning geometrical or topological features of a distribution based on a sample from that distribution. Examples include estimating the number of connected components of the support [6], the intrinsic dimensionality [27, 28] and the homology [29, 8, 38, 11, 34, 10, 20] of

the support, the Cheeger constant [3, 35] of the support, the gradient lines [1, 14] and ridges [12] and the decomposition into basin of attraction (Morse theory) [9, 13], the Minkowski content [16], etc.

The presentation was based on two recent papers [4, 2] that address the closely related problems of volume and perimeter of the boundary of the support. In more detail, we consider the model where we observe an IID sample $X_1, \dots, X_n \in \mathbb{R}^d$ from the density f. Based on the data, our aim is to estimate the support $S = \text{supp}(f)$.

Related work. Under this model, [23, 33, 19, 7] consider the case where convex S is *convex* and study the plug-in estimator based on the convex hull of the sample, both for the volume and the perimeter. Also under convexity, [5] derive the UMVU (uniformly of minimum variance among unbiased estimators) under Poissonization. [26, 22] consider the case where S is a *boundary fragment*

$$S = \{(z_1, \dots, z_d) \in [0,1]^d : z_d \leq h(z_1, \dots, z_{d-1})\}$$

and estimate functionals of the form

$$\int_S \varphi(z) \mathrm{d}z$$

$\varphi \equiv 1$ gives the volume of S. Minimax rates are obtained (upper and lower bounds).

Another model considered in the literature is the following. Assume that f is supported on $[0,1]^d$. One observes $(Z_1, Y_1), \dots, (Z_n, Y_n)$, where $Z_1, \dots, Z_n$ are IID uniform in $[0,1]^d$ (or a regular grid) and $Y_i | Z_i \sim \text{Bern}(p(Z_i))$ where $p(z) := \frac{1}{2} + a_n \mathbb{I}\{z \in S\}$. ($a_n > 0$ control the signal strength.) Under this other model, [17, 31, 30] estimate the Minkowski content of the boundary ∂S. [24] estimate the surface area of ∂S based on a Delaunay triangulation. [25] consider the case where S is a *boundary fragment* (in dimension $d = 2$) and estimate functionals of the form

$$\int_0^1 \psi(t, h(t), h'(t)) \mathrm{d}t$$

$\psi(a, b, c) = \sqrt{1 + b^2}$ gives the perimeter, i.e., the length of the curve $\{(t, h(t)) : t \in [0,1]\}$. Minimax rates are obtained (upper and lower bounds).

Contribution. In [4, 2] we make some smoothness assumptions on the support. Specifically, we assume that both S and S^c satisfy the *r-rolling condition*. This is equivalent to assuming that ∂S has *reach* at least r, or that S and S^c are *r-convex* [32, 21, 15, 36, 37].

For volume estimation, we use the plug-in estimate based on the *α-convex hull* of the sample [32], followed by a bias correction. We analyze the resulting estimator and find that it achieves the minimax rate for this problem (under the assumed smoothness). The bias correction step is essential for this, confirming previous findings in similar settings [26, 22].

For perimeter estimation (in $d = 2$), we use the *α-shape* the sample [18]. (We could also use the boundary of the α-convex hull of the sample. In fact, the sets of

points defining the α-convex hull and the α-shape coincide under mild assumptions on the point cloud.) We analyze the resulting estimator. We speculate, based on closely related work [25], that this estimator does not achieve the minimax rate due to its bias. However, unlike [25] (who work in the context of a boundary fragment model), we do not know how to correct the bias. This remains an open problem for future research.

References

[1] E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 2015.

[2] E. Arias-Castro, B. Pateiro-López, and A. Rodríguez-Casal. Minimax estimation of the volume of a set with smooth boundary. *arXiv preprint arXiv:1605.01333*, 2016.

[3] E. Arias-Castro, B. Pelletier, and P. Pudlo. The normalized graph cut and cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44(04):907–937, 2012.

[4] E. Arias-Castro and A. Rodríguez-Casal. On estimating the perimeter using the alpha-shape. *arXiv preprint arXiv:1507.00065*, 2015. To appear in the Annales de l'Institut Henri Poincaré.

[5] N. Baldin and M. Reiß. Unbiased estimation of the volume of a convex body. *arXiv preprint arXiv:1502.05510*, 2015. To appear in Stochastic Processes and their Applications.

[6] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280, 2007.

[7] H. Bräker and T. Hsing. On the area and perimeter of a random convex hull in a bounded convex set. *Probab. Theory Related Fields*, 111(4):517–550, 1998.

[8] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[9] J. E. Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

[10] F. Chazal, M. Glisse, C. Labruere, and B. Michel. Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv preprint arXiv:1305.6239*, 2013.

[11] F. Chazal and A. Lieutier. Weak feature size and persistant homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *Computational geometry (SCG'05)*, pages 255–262. ACM, New York, 2005.

[12] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015.

[13] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Statistical inference using the morse-smale complex. *arXiv preprint arXiv:1506.08826*, 2015.

[14] M.-Y. Cheng, P. Hall, and J. Hartigan. Estimating gradient trees. In *A festschrift for Herman Rubin*, volume 45 of *IMS Lecture Notes Monogr. Ser.*, pages 237–249. Inst. Math. Statist., Beachwood, OH, 2004.

[15] A. Cuevas, R. Fraiman, and B. Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Adv. Appl. Probab.*, 44(2):311–329, 2012.

[16] A. Cuevas, R. Fraiman, and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, 35(3):1031–1051, 2007.

[17] A. Cuevas, R. Fraiman, and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, 35(3):1031–1051, 2007.

[18] H. Edelsbrunner. Alpha shapes-a survey. *Tessellations in the Sciences*, 2010.

[19] B. Efron. The convex hull of a random set of points. *Biometrika*, 52(3-4):331–343, 1965.

[20] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[21] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

[22] G. Gayraud. Estimation of functionals of density support. *Mathematical Methods of Statistics*, 6(1):26–46, 1997.
[23] J. Geffroy. Sur un probleme d'estimation géométrique. *Publ. Inst. Statist. Univ. Paris*, 13:191–210, 1964.
[24] R. Jiménez and J. E. Yukich. Nonparametric estimation of surface integrals. *Ann. Statist.*, 39(1):232–260, 2011.
[25] J.-C. Kim and A. Korostelev. Estimation of smooth functionals in image models. *Mathematical Methods of Statistics*, 9(2):140–159, 2000.
[26] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.
[27] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, pages 777–784. MIT Press, Cambridge, Massachusetts, 2005.
[28] A. V. Little, Y.-M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. 2009.
[29] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.
[30] B. Pateiro-López and A. Rodríguez-Casal. Length and surface area estimation under smoothness restrictions. *Adv. in Appl. Probab.*, 40(2):348–358, 2008.
[31] B. Pateiro-López and A. Rodríguez-Casal. Surface area estimation under convexity type assumptions. *J. Nonparametr. Stat.*, 21(6):729–741, 2009.
[32] J. Perkal. Sur les ensembles ε-convexes. *Colloq. Math.*, 4:1–10, 1956.
[33] A. Rényi and R. Sulanke. Über die konvexe Hülle von n zufällig gewählten Punkten. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 3:138–147 (1964), 1964.
[34] V. Robins. Towards computing homology from finite approximations. In *Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999)*, volume 24, pages 503–532, 1999.
[35] N. G. Trillos, D. Slepcev, J. von Brecht, T. Laurent, and X. Bresson. Consistency of cheeger and ratio graph cuts. *arXiv preprint arXiv:1411.6590*, 2014.
[36] G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6):pp. 2273–2299, 1997.
[37] G. Walther. On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. Sci.*, 22(4):301–316, 1999.
[38] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.

Structured Nonparametric Curve Estimation

Enno Mammen

In structured nonparametrics one considers nonparametric or semiparametric models with several nonparametric components: f_1,...,f_q, where one only is interested in one component, f_1 say. Thus one thinks of applications where one wants to construct confidence bands or pointwise confidence sets for f_1 or one wants to test hypothesis on the component f_1 but where the other components f_2,...,f_q are nuissance nonparametric components that are of no specific interest.

A first class of examples is structured nonparametric regression where one observes i.i.d. $\mathbb{R}^q \times \mathbb{R}$-valued random variables (X^i, Y^i) $(i = 1, .., n)$ with

$$Y^i = G[\theta, f_1, ..., f_q](X^i) + \varepsilon_i$$

for some f_1,...,f_q belonging to some specified function classes and for a finite-dimensional parameter θ and with error variables ε_i that fulfill:

$$\mathbb{E}[\varepsilon_i | X^i] = 0.$$

Here we suppose that statistical inference focusses on one of the components f_1, ..., or f_q, but not primarily on the regression function $G[\theta, f_1, ..., f_q]$. Perhaps, the simplest example of structured nonparametric regression is the additive model where

$$G[c, f_1, ..., f_q](x) = c + f_1(x_1) + ... + f_q(x_q).$$

Other examples in structured nonparametric regression include: additive models with monotone component functions, additive models with increasing number of additive components and sparsity, generalized additive models with unknown link function $G[f_0, ..., f_q](x) = f_0(f_1(x_1) + ... + f_q(x_q))$, varying coefficient models $G[f_{j,k} : j \in \{1, ..., d\}, k \in I_j](x) = \sum_{j=1}^{d} x_j \sum_{k \in I_j} f_{j,k}(x_k)$, age-cohort-period models: $G[f_1, f_2, f_3](x) = f_1(x_1) + f_2(x_2) + f_3(x_1 + x_2)$, age-cohort-period models with operational time: $G[f_1, f_2, f_3, f_4](x) = f_1(x_1) + f_2(f_4(x_1)x_2) + f_3(x_1 + x_2)$.

In structured nonparametric density estimation one observes i.i.d. $\mathbb{R}^q$-valued random variables X^i $(i = 1, .., n)$ with density

$$G[\theta, f_1, ..., f_q](x)$$

for some f_1,...,f_q belonging to some specified function classes. A specification that we will discuss below is the nonparametric chain ladder model where $G[f_1, f_2](x) = f_1(x_1) f_2(x_2) \mathtt{I}_{x_1 + x_2 \le 1; x_1, x_2 \ge 0}$.

In our talk we report on some ongoing research projects with Young Kyung Lee (Seoul), María Dolores Martínez Miranda (Granada, London), Jens Perch Nielsen (London), Byeong Park (Seoul) where we discussed the chain-ladder model and some of its modifications. Furthermore, we we will state some asymptotic oracle results for high-dimensional additive models that were obtained in a project with Karl Gregory (Mannheim) and Martin Wahl (Berlin).

Estimates in the chain-ladder model can be based on kernel smoothing estimates $\hat{g}_1$ and $\hat{g}_2$ with kernel K and bandwidths h_1 and h_2 of g_1 and g_2 where

$$g_1(x) = \int_0^{1-x} f(x, y) w(x, y)\, dy, \qquad g_2(y) = \int_0^{1-y} f(x, y) w(x, y)\, dx$$

with some weight function $w(x, y) > 0$. Our estimators $\hat{c}_f$, $\hat{f}_1$ and $\hat{f}_2$ of c_f, f_1 and f_2 are given as solutions of the equation $\hat{\mathcal{F}}(\hat{c}_f, \hat{f}_1, \hat{f}_2) \equiv 0$ under the constraint $\int_0^1 \hat{f}_1(u) du = 1$ and $\int_0^1 \hat{f}_2(v) dv = 1$, with

$$\hat{\mathcal{F}}(c, r_1, r_2)(x, y) = \begin{pmatrix} c\, r_1(x) \frac{1}{\hat{g}_1(x)} \int_0^{1-x} r_2(v) w(x, v) dv - 1 \\ c\, r_2(y) \frac{1}{\hat{g}_2(y)} \int_0^{1-y} r_1(u) w(u, y) du - 1 \end{pmatrix}.$$

Under regularity assumptions one gets the following result.

Theorem 1 *The following expansion holds for* $\delta > 0$

$$\hat{f}_1(x) = f_1(x) - \frac{\frac{1}{nh_1}\sum_{i=1}^n K\left(\frac{X_i-x}{h_1}\right) w(X_i,Y_i) - \mathbb{E}\left[K\left(\frac{X_i-x}{h_1}\right) w(X_i,Y_i)\right]}{g_1(x)} + h_1^2 b_1(x) + o_P(\frac{1}{\sqrt{nh_1}} + h_1^2),$$

uniformly for $\delta \le x, y \le 1-\delta$ *with* b_1, b_2 *given as solutions of a deterministic linear integral equations.*

The theorem can be used to develop asymptotic distribution theory for $\hat{f}_1$. Extensions of the chain-ladder model in our project include varying time scales (operational time) and seasonal effects.

In the last part of this note we report on our asymptotic oracle results for high-dimensional additive models. We compare estimation of f_1 in the additive model

$$Y^i = c + f_1(X_1^i) + ... + f_q(X_q^i) + \varepsilon_i$$

with estimation of f_1 in the oracle model $Z^i = c + f_1(X_1^i) + \varepsilon_i$. For identification we assume that $\mathbf{E}[f_j(X_j^i)] = 0$. We will argue that f_1 can be estimated in an additive model asymptotically as well as f_1 in the oracle model. This means that not knowing f_2,...,f_q does not lead to a loss of statistical information on f_1, at least asymptotically up to first order. We now give a more precise formulation of this asymptotic equivalence result. Suppose that a smoothing estimator

$$\hat{f}_1^{oracle}(x_1) = SMOOTH_{X_1^i \to Z^i}(x_1)$$

of f_1 in the oracle model $Z^i = c + f_1(X_1^i) + \varepsilon_i$ is given. We now ask: can we construct an estimator $\hat{f}_1(x_1)$ of f_1 in the additive model $Y^i = c + f_1(X_1^i) + ... + f_q(X_q^i) + \varepsilon_i$ with

$$\|\hat{f}_1 - \hat{f}_1^{oracle}\|_\infty = o_P(\delta_n),$$

where δ_n is the rate of convergence of $\hat{f}_1^{oracle}$ to f_1. The answer is: Yes!

For the case that the number q of functions is fixed, this has been shown in Horowitz, Klemelä and Mammen (2006). For a sparse high-dimensional additive model such a result has been achieved in Gregory, Mammen and Wahl (2016). There it has been allowed that the number q of functions may grow with n, even with $q > n$, but that the number s_0 of nonzero functions may grow, but only with $s_0 < n$.

The basic idea of the construction in these papers is the observation that for a nonparametric estimator $\tilde{f}_1^{oracle}$ with low bias and high variance (undersmoothing) it holds that

$$(1) \qquad \hat{f}_1^{oracle}(x_1)(= SMOOTH_{X_1^i \to Z^i}(x_1)) = SMOOTH_{X_1^i \to \tilde{Z}^i}(x_1) + o_P(\delta_n),$$

where $\tilde{Z}^i = \tilde{f}_1^{oracle}(X_1^i)$. This means that "smoothing $\approx$ smoothing $\circ$ undersmoothing". This property holds for many smoothing estimators as kernel estimators, smoothing splines, orthogonal series estimators, Pinsker estimator, ...

Horowitz, Klemelä and Mammen (2006) and Gregory, Mammen and Wahl (2016) state conditions under which it is possible to construct undersmoothing estimators $\tilde{f}_1$, ..., $\tilde{f}_q$ in the additive model such that:

$$\tilde{Z}^i = \tilde{f}_1^{oracle}(X_1^i) = \tilde{f}_1(X_1^i) + o_P(\delta_n). \tag{2}$$

This implies for $\hat{f}_1(x_1) = SMOOTH_{X_1^i \to \tilde{Y}^i}(x_1)$ with $\tilde{Y}^i = \tilde{f}_1(X_1^i)$ that:

$$\begin{aligned} \hat{f}_1(x_1) &= SMOOTH_{X_1^i \to \tilde{Z}^i}(x_1) + o_P(\delta_n) = SMOOTH_{X_1^i \to Z^i}(x_1) + o_P(\delta_n) \\ &= \hat{f}_1^{oracle}(x_1) + o_P(\delta_n). \end{aligned} \tag{3}$$

Here is the argument again: For *one* simple undersmoothing estimator $\tilde{f}_1^{oracle}$ in the oracle model one shows the existence of an estimator $\tilde{f}_1$ with (2). Then one gets for *all* estimators $\hat{f}_1^{oracle}$ in the oracle model with (1) that (3) holds for the choice for $\hat{f}_1(x_1) = SMOOTH_{X_1^i \to \tilde{Y}^i}(x_1)$ with $\tilde{Y}^i = \tilde{f}_1(X_1^i)$. This is the oracle result for additive models with a fixed number of additive components and for sparse high-dimensional additive model. It has to be studied to which extent the oracle results of additive models carry over to other nonparametric models? The practical implementation of the two-step procedure needs some further work where also smoothing parameter selection has to be discussed.

References

[1] K. Gregory, E. Mammen and M. Wahl, *Optimal estimation of sparse high-dimensional additive models*, Preprint (2016).

[2] J. Horowitz, J. Klemelä and E. Mammen, *Optimal estimation in additive regression models*, Bernoulli **12** (2006), 271–298.

[3] Y. K. Lee, E. Mammen, J. P. Nielsen and B. U. Park, *Operational time and in-sample density forecasting*, Ann. Statist. (2016), to appear.

[4] Y. K. Lee, E. Mammen, J. P. Nielsen and B. U. Park, *Asymptotics for In-Sample Density Forecasting*, Ann. Statist. **43** (2015), 620–645.

[5] E. Mammen, M. D. Martínez Miranda, and J. P. Nielsen, *Structured density forecasting with applications to non-life insurance and mesothelioma mortality*, Insurance: Mathematics and Economics **61** (2015), 76–86.

Reporter: Max Sommerfeld

Participants

Prof. Dr. Dragi Anevski
Centre for Mathematical Sciences
University of Lund
Box 118
221 00 Lund
SWEDEN

Prof. Dr. Ery Arias-Castro
Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
UNITED STATES

Dr. Fadoua Balabdaoui
CEREMADE
Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

Prof. Dr. José Enrique Chacón Durán
Departamento de Matemáticas
Universidad de Extremadura
Centro Universitario de Mérida
Avenida Santa Teresa de Jornet, 38
06800 Mérida
SPAIN

Dr. Sabyasachi Chatterjee
Department of Statistics
The University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637
UNITED STATES

Prof. Dr. Jessi Cisewski
Department of Statistics
Yale University
P.O. Box 208290
New Haven, CT 06520-8290
UNITED STATES

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

Prof. Dr. Lutz Dümbgen
Institut für Mathematische Statistik
und Versicherungslehre
Universität Bern
Alpeneggstrasse 22
3012 Bern
SWITZERLAND

Prof. Dr. Christopher Genovese
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

Prof. Dr. Piet Groeneboom
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
NETHERLANDS

Kim Hendrickx
Center for Statistics
Hasselt University
Building D
Agoralaan
3590 Diepenbeek
BELGIUM

Prof. Dr. Stephan Huckemann
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Geurt Jongbloed
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
NETHERLANDS

Dr. Rik Lopuhaä
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
NETHERLANDS

Prof. Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. James Stephen Marron
Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
UNITED STATES

Prof. Dr. Bertrand Michel
Laboratoire de Statistique Théorique et
Appliquée
Université Pierre et Marie Curie, Paris
VI
4, place Jussieu
75252 Paris Cedex 05
FRANCE

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

Eni Musta
Department of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
NETHERLANDS

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

Prof. Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Armin Schwartzman
Division of Biostatistics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0631
UNITED STATES

Prof. Dr. Bodhisattva Sen
Department of Statistics
Columbia University
1255 Amsterdam Avenue
New York, NY 10027
UNITED STATES

Max Sommerfeld
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Anuj Srivastava
Department of Statistics
Florida State University
Tallahassee FL 32306-4330
UNITED STATES

Prof. Dr. Jon A. Wellner
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
UNITED STATES