Mathematisches Forschungsinstitut Oberwolfach

Report No. 13/2011

DOI: 10.4171/OWR/2011/13

Mini-Workshop: Level Sets and Depth Contours in High Dimensional Data

Organised by
Mia Hubert, Katholieke Universiteit Leuven
Jun Li, University of California, Riverside
Wolfgang Polonik, University of California, Davis
Robert Serfling, University of Texas, Dallas

February 27th – March 5th, 2011

Abstract. Extraction of information about the distribution underlying a high-dimensional data set is a formidable, complex problem dominating modern nonparametric statistics. Two general strategies are (i) to extract merely qualitative information, such as modality or other shape information, and (ii) to consider relatively simple inference problems, such as binary classification. One approach toward (i) and (ii) is based on measuring qualitative information via mass concentration functions. Another approach is based on multivariate depth functions and inherently addresses issues of robustness. Having different orientations and aims, these approaches have evolved in parallel with little interaction. Yet they both in common implicitly involve level set estimation as a major tool. This mini-workshop was the first serious attempt to study and exploit such interconnections between these approaches. Researchers from both areas exchanged ideas toward forging a novel, synergistic approach that fruitfully strengthens the roles of mass concentration and depth methods in statistical inference for multivariate data. Foundations for level set estimation as a general statistical method were explored. Deeper understanding of the so-called generalized quantiles approach was pursued. Application to binary classification, a pervasive problem in modern statistics, received intensive special attention.

Mathematics Subject Classification (2000): Primary: Nonparametric Inference 62G99, Secondary: Multivariate Analysis 62J99.

Introduction by the Organisers

The mini-workshop *Level Sets and Depth Contours in High Dimensional Data*, organized by Mia Hubert (Katholieke Universiteit Leuven), Jun Li (UC Riverside), Wolfgang Polonik (UC Davis) and Robert Serfling (UT Dallas) on February 27–March 5th, 2011 brought together 18 participants with diverse geographic, demographic and with research expertise in level sets and/or depth methodology.

Statistical methodology for high-dimensional data problems faces many challenges, many of them relate to *geometry*. One general strategy for dealing with related dilemmas is to consider less complex goals. Here one approach is to obtain *qualitative* information about *shape*, with monotonicity, modality, or *mass concentration* of the underlying distribution being specific instances. Another approach involves the use of *multivariate quantiles* and *depth functionals*. In fact, the notion of a quantile is closely related to the notion of *outlyingness*, which in turn connects with *robustness* of multivariate statistical methodology.

Interestingly, the two above-mentioned approaches both entail *the estimation of level sets*: (a) level sets of depth functionals, or depth contours, provide a measure outlyingness of outlyingness of multivariate data; (b) several mass concentration functions of a multivariate distribution can be considered as functionals of level sets of the corresponding probability density function; (c) information about modality of a distribution is reflected in the shape of level sets of the probability density function; (d) in a classification context, the Bayes decision boundary of the optimal classifier is given by a level set of the regression function.

A second general strategy in dealing with the challenges posed by high dimensionality is to confine to relatively simple statistical inference procedures. A very important example is the *classification or discrimination problem*. Again we find level set estimation as an underlying tool: for example, the optimal (Bayes) classifier is a level set of the corresponding regression function (conditional expectation). In fact, typically (binary) classifiers are characterized by a decision boundary which may be represented as $\{x : g(x) = 0\}$ for some decision function g. In other words, if an observation falls into the level set $\{x : g(x) \geq 0\}$ then it is classified into one class, and otherwise it is classified into the other class.

Although different in their orientation and goals, and having evolved with relatively little interaction, depth-based methods and the mass concentration approach are connected via technical commonalities revolving around the general theme of *level set estimation* and through certain applications such as the classification problem. Formal investigation and systematic exploitation of such connections among these quite distinct statistical settings would be novel and fruitful. It is the chief target of the proposed workshop, and other spin-offs are anticipated as well.

This mini-workshop brought together representatives of the depth and mass concentration groups along with interested statisticians from related areas. This was a first serious attempt to forge a new synergy yielding a deeper understanding of level set and depth contour estimation and their applications and to spawn new research directions.

Mini-Workshop: Level Sets and Depth Contours in High Dimensional Data

Table of Contents

Abstracts

(Level) Sets in Statistics

Wolfgang Polonik

The simple notion of a level set of a function h, i.e. the set $\{x : h(x) \geq \lambda\}$ for some threshold value λ, turns out to be a cornerstone for many multivariate statistical methodologies. Such methodologies include notions of stochastic ordering (including majorization with close relation to excess mass estimation), clustering, classification, data depth, mode hunting, multivariate quantiles, nonparametric maximum likelihood estimation under order restriction, goodness-of-fit testing, testing for modality and partially identified models with applications in anomaly detection, astronomy, chemometrics, econometrics, intrusion detection, medical imaging, and more.

Various types of functions h have been considered, such as probability densities, depth functionals, regression functions, and weighted differences of probability densities. In the context of partially identified models, the function h is a function of the parameter of the underlying parametric model, i.e. the corresponding level set is a subset of the parameter space.

The construction of the statistical methods requires the estimation of level sets. Direct estimation methods include the excess mass and the minimum volume approach. These methods require the specification of a model for level sets in terms of a class of candidate sets (e.g. convex sets). Should the model be misspecified, then the corresponding direct estimation methods lead to certain generalizations of level sets, such as minimum volume sets or generalized λ-clusters. Indirect methods are based on plug-in estimators of the corresponding function h.

In this talk we will discuss the indicated statistical methodology and their relations under the point of view of (generalized) level sets, and we will discuss relations among these methodologies.

References

[1] Bugni, F. (2010): Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set. *Econometrica* **76**, 735-753.

[2] Duong, T., Koch, I. and Wand, M.P. (2009). Highest density difference region estimation with application to flow cytometric data. *Biometrical Journal* **51**(3), 504-521.

[3] Hardy, G.H. Littlewood, J.E., and Pólya (1929): Some simple inequalities satisfied by convex functions. *Messenger Math.* **58**, 145 - 152.

[4] Hartigan, J.A. (1987). Estimation of a convex density contour two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

[5] Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27** 783–858.

[6] Manski, C.F. (1989). Anatomy of the selection problem. *J. Hum. Resources* **24**(3), 343 - 360.

[7] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

[8] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *Ann. Statist.* **23** 855–881.
[9] Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69** 1 - 24.
[10] Polonik, W. (1999): Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribution free methods. *Ann. Statist.* **27** 1210–1229.
[11] Polonik, W. and Wang, Z. (2005). Estimation of regression contour clusters - An application of the excess mass approach to regression. *J. Multivariate Anal.* **94** 227 – 249.
[12] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.

Depth, Outlyingness, Quantile, and Rank Functions

Robert Serfling

Outlyingness, quantile, and rank functions are well understood in the univariate data setting. In the multivariate setting, these along with depth functions have been developed in recent years into a new nonparametric multivariate statistical analysis methodology that is better tuned to the geometric nature of data in higher dimensions. Contours and level sets play a focal role here, but to date the potential application of some general theory on level sets that has developed in parallel to the depth function development has not been exploited well. This Workshop is designed to establish strong connections between the level sets and depth functions communities. In particular, this is an opening talk providing an overview of the landscape of depth functions, broadly considered. It provides a general framework for the various talks at this Workshop on specific "depth function" topics, a general orientation for the "level sets" community at this Workshop, and a prelude to discussion of a "level sets–depth functions synergy" at this Workshop. Key topics in this talk are:

- Relations among depth, outlyingness, quantile, and rank functions in $\mathbb{R}^d$.
- Desirable properties, especially affine invariance and equivariance.
- Applications.
- Computation.
- Convergence theory for sample versions.
- Extensions to abstract settings, for a broader range of inference problems.

References

[1] Chaudhuri, P., *On a geometric notion of quantiles for multivariate data*, Journal of the American Statistical Association **91** (1996), 862–872.
[2] Dang, X. and Serfling, R., *Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties*, Journal of Statistical Planning and Inference **140** (2010), 198–213.
[3] Dang, X., Serfling, R. and Zhou, W., *Influence functions of some depth functions, and application to depth-weighted L-statistics*, Journal of Nonparametric Statistics **21** (2009), 49–66.
[4] Ilmonen, P., Nevalainen, J., and Oja, H., *Characteristics of multivariate distributions and the invariant coordinate system*, Preprint (2010).

[5] Einmahl, J. H. J. and Mason, D. M., *Generalized quantile processes*, Annals of Statistics, **20** (1992), 1062–1078.
[6] Liu, R. Y., *On a notion of simplicial depth*, Proceedings of the National Academy of Science USA **85** (1988), 1732–1734.
[7] Liu, R. Y., Parelius, J. M. and Singh, K., *Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion)*, Annals of Statistics **27** (1999), 783–858.
[8] Lopez-Pintado, S. and Romo, J., *On the concept of depth for functional data*, Journal of the American Statistical Association **104** (2009), 718–734.
[9] Mizera, I., *On depth and deep points: a calculus*, Annals of Statistics **30** (2002), 1681–1736.
[10] Möttönen, J. and Oja, H., *Multivariate spatial sign and rank methods*, Journal of Nonparametric Statistics **5** (1995), 201–213.
[11] Rousseeuw, P. J. and Hubert, M., *Regression depth (with discussion)*, Journal of the American Statistical Association, **94** (1999), 388–433.
[12] Serfling, R., *Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis*, Journal of Multivariate Analysis **83** (2002), 232–247.
[13] Serfling, R., *Depth functions in nonparametric multivariate analysis*, In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications* (R. Y. Liu, R. Serfling, D. L. Souvaine, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science **72** American Mathematical Society (2006), 1–16.
[14] Serfling, R., *On strong invariant coordinate system (SICS) functionals*, Working paper (2009).
[15] Serfling, R., *Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization*, Journal of Nonparametric Statistics **22** (2010), 915–936.
[16] Serfling, R. and Zuo, Y., *Some perspectives on multivariate quantile and depth functions*, Annals of Statistics **38** (2010), 676–684. [Invited and refereed discussion]
[17] Tukey, J. W., *Mathematics and the picturing of data*, In *Proceedings of the International Congress of Mathematicians, Vancouver 1974* (R. D. James, ed.) **2** (1975), 523–531.
[18] Zhang, J., *Some extensions of Tukey's depth function*, Journal of Multivariate Analysis **82** (2002), 134–165.
[19] Zhou, W. and Serfling, R., *Multivariate spatial U-quantiles: a Bahadur-Kiefer representation, a Theil-Sen estimator for multiple regression, and a robust dispersion estimator*, Journal of Statistical Planning and Inference **138** (2008), 1660–1678.
[20] Zuo, Y. and Serfling, R., *General notions of statistical depth function*, Annals of Statistics **28** (2000), 461–482.
[21] Zuo, Y. and Serfling, R., *Structural properties and convergence results for contours of sample statistical depth functions*, Annals of Statistics **28** (2000), 483–499.

Detection of Spatial Clusters with the Scan and the Average Likelihood Ratio

Guenther Walther
(joint work with Hock Peng Chan)

We are concerned with the problem of detecting a deterministic signal with unknown spatial extent against a noisy background. The standard statistical tool to address this problem is the scan statistic (maximum likelihood ratio statistic), which considers the maximum of local likelihood ratio statistics on certain subsets of the data. There is a large body of work on scan statistics, see e.g. [5]. But there is also empirical evidence that the scan statistic is suboptimal, see e.g. [7, 2].

[10, 4] propose to use the average of the likelihood ratio statistics instead of their maximum. In the first part of the talk we present the results of a theoretical comparison of these two methods in the prototypical univariate sampled data model with white Gaussian noise and we obtain the following results:

The scan statistic possesses optimal detection power only for signals with the smallest spatial extent. Otherwise the scan statistic is suboptimal, and the loss of power can be considerable for signals having a large spatial extent. In the case of the average likelihood ratio (ALR) statistic, these conclusions hold in reversed order: The ALR possesses optimal detection power for signals having large spatial extent, but is suboptimal for signals with small spatial extent. However, the loss of power in the latter case is so small that it is unlikely to be of concern, at least for most sample sizes considered today.

Next we consider simple modifications to obtain universal optimality for both the ALR and the scan. The ALR averages the likelihood ratios pertaining to $\sim n^2$ stretches of the data, where n is the sample size, resulting in an $O(n^2)$ algorithm. Thus the use of the ALR is computationally infeasible even for moderate sample sizes. We introduce a condensed ALR that averages only a certain subset of the likelihood ratios and we show that this condensed ALR possesses optimal detection power for signals having arbitrary spatial extent. Furthermore, this condensed ALR can be computed in almost linear time, viz. with an $O(n \log^2 n)$ algorithm. In light of the preceding discussion, it is arguably this improvement in computation time rather than the small gain in detection power that is the main advantage of this modification. We note that typically, an approximation introduced to make a procedure computationally less intensive will on the flip side degrade its performance somewhat. It is thus noteworthy that in the case of the ALR, our computationally efficient modification will actually lead to an improved (in fact: optimal) performance.

In the case of the scan, optimality obtains by employing critical values that depend on the size of the region under consideration. Two possible ways to implement this idea are via penalization as introduced in [6] or by grouping regions that have about the same size into blocks as in [11]. Various efficient algorithms for computing a good approximation to the scan statistic have been introduced in [8, 1, 11, 9]. Unlike the ALR, constructing a computationally efficient approximation for the scan will not lead to universally optimal power. Rather, statistical optimality for the scan derives from employing size-dependent critical values.

We then consider the detection problem in the multivariate Bernoulli model. Our examination suggests that the main conclusions from above extend to the multivariate case, and we present sharp results for the special case of the Bernoulli model in $\mathbf{R}^2$.

References

[1] Arias-Castro, E., Donoho, D.L., and Huo, X., *Near-optimal detection of geometric objects by fast multiscale methods*, IEEE Trans. Inform. Th. **51** (2005), 2402–2425.

[2] Chan, H.P., *Detection of spatial clustering with average likelihood ratio test statistics*, Ann. Statist. **37** (2009), 3985–4010.

[3] Dumbgen, L. and Spokoiny, V.G., *Multiscale testing of qualitative hypotheses*, Ann. Statist. **29** (2001), 124–152.
[4] Gangnon, R.E. and Clayton, M.K., *The weighted average likelihood ratio test for spatial disease clustering*, Statistics in Medicine **20** (2001), 2977–2987.
[5] Glaz, J., Poznykov, V. and Wallenstein, S., *Scan Statistics: Methods and Applications* (2009). Birkhauser, Boston.
[6] Dümbgen, L. Aan Walther, G., *Multiscale inference about a density.* Ann. Statist. **36** (2008), 1758-1785.
[7] Neill, D., *An empirical comparison of spatial scan statistics for outbreak detection*, Internat. Journal of Health Geographics **8** (2009), 1–16.
[8] Neill, D. and Moore, A., *A fast multi-resolution method for detection of significant spatial disease clusters*, Adv. Neur. Info. Proc. Sys. **10** (2004), 651–658.
[9] Rufibach, K. and Walther, G., *The block criterion for multiscale inference about a density, with applications to other multiscale problems*, Journal of Computational and Graphical Statistics **19** (2010), 175–190.
[10] Siegmund, D., *Is peak height sufficient?*, Genetic Epidemiology **20** (2001), 403–408.
[11] Walther, G., *Optimal and fast detection of spatial clusters with scan statistics*, Ann. Statist. **38** (2010), 1010-1033.

Data Depth and Multivariate Spacings, Ordering and Beyond

Regina Y. Liu
(joint work with Juan A. Cuesta-Albertos, Jun Li)

Spacings derived from univariate order statistics have been the foundation for much of the development in statistical inference and nonparametric statistics. The excellent treatise by R. Pyke (1965) as well as the references therein and thereafter all attest to the importance of spacings. In his paper R. Pyke lamented,

> Perhaps the most significant restrictions of this paper has been our concern with one-dimensional spacings. There are many applications in which samples are drawn from two- or even three-dimensional space and for which it is important to study the spacings of the observations.

Although research on spacings has continued, his call for multivariate spacings has remained largely unanswered. The main difficulty in generalizing the univariate spacings to multivariate settings is the lack of suitable ordering schemes for multivariate observations. Using the multivariate order statistics derived from data depth, we introduce and develop multivariate spacings. Specifically, the spacing between two consecutive order statistics is the region which bridges the two order statistics, in the sense that the region contains all the points whose depth values fall between the depth values of the two consecutive order statistics. These multivariate spacings can be viewed as a data-driven realization of the so-called "statistically equivalent blocks". These spacings assume a form of center-outward layers of "shells" ("rings" in the two-dimensional case), for which the shapes of the shells follow closely the underlying probabilistic geometry.

We discuss the properties and applications of these spacings. For example, we use the spacings to construct multivariate tolerance regions. The construction

is nonparametric and completely data driven, and the resulting tolerance region reflects the true geometry of the underlying distribution. This is different from the existing approaches which require that the shape of the tolerance region be specified in advance.

Finally, we also discuss several families of multivariate goodness-of-fit tests based on the proposed spacings.

References

[1] R. Pyke, *Spacings*, Journal of the Royal Statistical Society. Series B (Methodological) **27** (1965), 395-449.

[2] J. Li, and R. Liu, *Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions*, Annals of Statistics **36** (2008), 1299-1323.

[3] J. Cuesta-Albertos, J. Li, and R. Liu, *Multivariate spacings based on data depth: II. Goodness-of-fit tests*, preprint.

Weighted Generalisation of Halfspace Depth

Daniel Hlubinka

(joint work with Lukáš Kotík)

The data depth is capturing global properties of the underlying probability distribution unlike the probability density function which is purely local. Both the global and the local view to the distribution have its advantages and disadvantages in applications. In our lecture we propose a way how to include local behaviour of the probability density function into a (generalised) halfspace depth function. The idea is very simple but it allows surprisingly flexible shapes for the central depth regions which need not to be convex, star-shaped or even connected.

The main idea is to introduce the *weight function* to the depth calculation. Consider a function $w(x,s) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ and define the depth of point $\theta \in \mathbb{R}^d$ w.r.t. the probability distribution P as

$$(1) \quad D(\theta, P) = \inf_{s,\|s\|=1} \mathrm{E}w(X-\theta, s), \quad D_N(\theta, P_n) = \inf_{s,\|s\|=1} \frac{1}{N}\sum_{i=1}^{N} w(X_i-\theta, s)$$

where s represents a normal vector to the halfspace border hyperplane and X is a random variable distributed according to the probability distribution P, D_N being the sample version of D based on iid sample of size N from P. In some sense *generalised halfspaces* are replacing the usual halfspace in (1).

There is a natural question on the possible choices of the weight function w. It is clear that the choice $w(x,s) = 1_{\langle x,s\rangle \geq 0}$, i.e., w being an indicator of a halfspace gives the usual halfspace depth. On the other hand $w(x,s) = k(x)$, k being some kernel centred at 0 (Gaussian, Epanechnikov, Uniform, ...) gives a kernel density estimation and hence the depth central regions become the plug-in density level sets estimators. Therefore choosing the weight function w one controls the "global" and the "local" properties included in the depth values.

The central regions of this generalised halfspace depth are typically smaller in the Lebesgue measure sense than the corresponding halfspace depth central regions. Also the "shape" of the underlying distribution is better captured by the generalised depth (except the elliptically symmetric distributions where depth and density contours are equivalent). The generalised depth shows quite good performance in particular in classification problems.

References

[1] D. Hlubinka, L. Kotík and O. Vencálek, *Weighted halfspacedepth*, Kybernetika **46** (2010), 125–14*.

Data Mining Methods Based on Kernelized Spatial Depth

Xin Dang

Statistical depth functions provide center-outward ordering of points with respect to a distribution or a date set in high dimensions. Of the various depth notions, the spatial depth is appealing because of its computational efficiency. However, it tends to provide circular contours and fail to capture well the underlying probabilistic geometry outside of the family of spherically symmetrical distributions. We propose a novel statistical depth, the *kernelized spatial depth* (KSD), which generalizes the spatial depth via *positive definite kernels*. By choosing a proper kernel, the KSD can capture the local structure of a data set while the spatial depth fails. We demonstrate this by the half-moon data and the triangle-shaped data. Based on the KSD, we propose a novel *outlier detection algorithm*, by which an observation with a depth value less than a threshold is declared as an outlier. The proposed algorithm is simple in structure: the threshold is the only one parameter for a given kernel. It applies to a one-class learning setting, in which "normal" observations are given as the training data, as well as to a missing label scenario where the training set consists of a mixture of normal observations and outliers with unknown labels. We give upper bounds on the false alarm probability of a depth-based detector. These upper bounds can be used to determine the threshold. We perform extensive experiments on synthetic data and data sets from real applications. The proposed outlier detector is compared with existing methods. The KSD outlier detector demonstrates competitive performance.

KSD is extended to graph data, where pairwise relationships of objects are given and represented by edges. Several graph kernels including a new proposed one, complement Laplacian kernel, are considered for ranking the "centrality" of graph vertices. An application of graph ranking to gene data is briefly discussed. Six gene expression profiles from the Gene Expression Omnibus (GEO) include three DNA reactive agents to induce genotoxic stress, two DNA non-reactive agents to induce cytotoxic stress and one control group. The goal is to identify the most important genes differentiating genotoxic compounds from the cytotoxic compounds. We first construct a bipartite graph from the Gene Ontology (GO), which describes dependent structure of genes. For each compound, add weights to the bigraph

using the gene expression data. Run the KSD algorithm on each bi-graph to develop a gene expression profile of ranked genes for each compound. Then compare the ranked gene sets to find the genes which are differently regulated between two group treatments.

A clustering algorithm based on KSD is also proposed. Preliminary results show it promising. With successes in the application, theoretical developments of KSD are demanding. The talk will be ended with questions:

(1) What properties does the KSD possess?
If the kernel is fixed, continuity of KSD $D_\kappa(\mathbf{x}, \mathbf{F})$ as a function of $\mathbf{x}$ and continuity as functional F can be established. Also, consistency and asymptotic normality of sample KSD can be established via the practice of V-statistic theory. Influence function can be used for robustness analysis. However, usual desired properties of depth functions may fail because KSD fails to provide center-outward global ordering. Kernelized spatial median and kernelized spatial quantile are also briefly discussed.
(2) What's the role of parameter in the kernel? How to choose it optimally?
(3) What is relationship between KSD and kernel density estimation?

References

[1] Y. Chen, X. Dang, H. Peng and H. Bart, *Outlier detection with the kernelized spatial depth function*, IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2) 2009, 288-305.

[2] C. Gao, X. Dang, Y. Chen and D. Wikins, *Graph ranking for exploratory gene data analysis*, BMC Bioinformatics **10**(Suppl 11) 2009.

Robust Kernel Density Estimation

Clayton Scott
(joint work with JooSeuk Kim)

This talk describes a method of nonparametric density estimation that exhibits robustness to contamination of the training sample, meaning the training sample consists of some realizations that are not from the density being estimated. This problem is motivated, for example, by anomaly detection applications. When labeled examples of anomalies are unavailable, it is common to define an anomaly detector by thresholding a density estimate based on non-anomalous data. In applications where it is difficult or impossible to obtain a pure sample (containing no anomalies), robust density estimation can mitigate the impact of contamination.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ be a random sample from a distribution F with a density f. We imagine $f = (1-\epsilon)f_0 + \epsilon f_1$, where f_1 represents the anomalous component. No assumptions are made on f_0 or f_1. The kernel density estimate of f is $\widehat{f}_{KDE}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n k_\sigma(\mathbf{x}, \mathbf{X}_i)$, where $k_\sigma(\mathbf{x}, \mathbf{X}_i)$ is for concreteness assumed to be the Gaussian kernel $k_\sigma(\mathbf{x}, \mathbf{X}_i) = (\sqrt{2\pi}\sigma)^{-d}\exp(-\|\mathbf{x} - \mathbf{X}_i\|^2/2\sigma^2)$.

For the Gaussian kernel, there exists a mapping $\Phi : \mathbb{R}^d \to H_\sigma$, where H_σ is an infinite dimensional Hilbert space, such that $k_\sigma(\mathbf{x}, \mathbf{x}') = \langle \Phi_\sigma(\mathbf{x}), \Phi_\sigma(\mathbf{x}') \rangle$. We will

assume that $\Phi_\sigma(\mathbf{x})$ is the canonical feature map, $\Phi_\sigma(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$. We also recall the reproducing property, which states that for all $g \in H_\sigma$, $g(\mathbf{x}) = \langle \Phi_\sigma(\mathbf{x}), g \rangle$ [1].

From this point of view, the KDE can be expressed as

$$\widehat{f}_{KDE}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \big\langle \Phi_\sigma(\mathbf{x}), \Phi_\sigma(\mathbf{X}_i) \big\rangle = \left\langle \Phi_\sigma(\mathbf{x}), \frac{1}{n}\sum_{i=1}^{n} \Phi_\sigma(\mathbf{X}_i) \right\rangle .$$

By the reproducing property of $\Phi_\sigma(\mathbf{x})$, $\widehat{f}_{KDE} \in H_\sigma$ can be seen as $\frac{1}{n}\sum_{i=1}^{n} \Phi_\sigma(\mathbf{X}_i)$, the sample mean of $\Phi_\sigma(\mathbf{X}_i)$'s, or equivalently, the solution of

$$\min_{g \in H_\sigma} \sum_{i=1}^{n} \|\Phi_\sigma(\mathbf{X}_i) - g\|^2_{H_\sigma}.$$

For a robust loss function $\rho(x)$ on $x \geq 0$, the robust kernel density estimate is defined as

$$\widehat{f}_{RKDE} = \arg\min_{g \in H_\sigma} \sum_{i=1}^{n} \rho\big(\|\Phi_\sigma(\mathbf{X}_i) - g\|_{H_\sigma}\big). \tag{1}$$

Well-known examples of robust loss functions are Huber's or Hampel's ρ.

It seems clear that this new estimator is a robust version of the KDE in the Hilbert space, but in what sense is the corresponding function a robust estimate of the density? Is the RKDE even a density? We address these questions as follows.

Representer Theorem: If ρ satisfies certain common assumptions, then

$$\widehat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^{n} w_i k_\sigma(\mathbf{x}, \mathbf{X}_i)$$

for some $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$. Therefore the RKDE is a density. Furthermore

$$w_i \propto \varphi(\|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\|)$$

where $\varphi(t) = \psi(t)/t$ and $\psi = \rho'$. Notice that φ is a decreasing function for a robust loss. Combining this with

$$\begin{aligned} \|\Phi_\sigma(\mathbf{x}) - \widehat{f}\|^2_{H_\sigma} &= \langle \Phi_\sigma(\mathbf{x}) - \widehat{f}, \Phi_\sigma(\mathbf{x}) - \widehat{f} \rangle_{H_\sigma} \\ &= \|\Phi_\sigma(\mathbf{x})\|^2_{H_\sigma} - 2\langle \Phi_\sigma(\mathbf{x}), \widehat{f} \rangle_{H_\sigma} + \|\widehat{f}\|^2_{H_\sigma} = (\sqrt{2\pi}\sigma)^{-d} - 2\widehat{f}(\mathbf{x}) + \|\widehat{f}\|^2_{H_\sigma}, \end{aligned}$$

the RKDE is such that lower weights are assigned to points in regions of lower (estimated) density. Thus outlying data points contribute less to the estimate.

Computation: A kernelized form of iterative re-weighted least-squares is presented. The algorithm requires $O(n^2)$ steps per iteration, and often converges in fewer than 10 iterations. For convex losses, the algorithm provably converges to the global minimizer of (1). The algorithm can be viewed as finding a fixed point of the equations in the aforementioned representer theorem.

Influence Function: We show how the influence function for an RKDE, based on the empirical distribution, can be obtained by solving a system of linear equations. These equations also reveal the insensitivity of the RKDE to outlying data.

Asymptotics: For fixed bandwidth σ, the infinite sample limits of the RKDE is $f_\sigma = k_\sigma * p_\sigma$ where

$$p_\sigma(\mathbf{x}) = \frac{w_\sigma(\mathbf{x}) f(\mathbf{x})}{\int w_\sigma(\mathbf{y}) f(\mathbf{y}) dy}$$

and $w_\sigma(\mathbf{x}) = \varphi(\|\Phi_\sigma(\mathbf{x}) - f_\sigma\|_{H_\sigma})$. For the quadratic loss, $p_\sigma = f$, whereas for a robust loss, p_σ is a modified version of f where points with lower density level are down-weighted. Thus, the RKDE introduces a bias into the kernel density estimator that down-weights the influence of outlying data.

Experiments: We take several benchmark data sets for binary classification, and use them to create a contaminated sample. We then assess the ability of the RKDE to estimate level sets of the non-anomalous component f_0, evaluated in terms of the area under the ROC. The RKDE with Hampel's loss considerably outperforms competing methods.

References

[1] I. Steinwart and A. Christmann, *Support Vector Machines*, New York: Springer, 2008.

Adaptive Density Level Set Clustering

Ingo Steinwart

A central and widely studied task in statistical learning theory is cluster analysis, where the goal is to find clusters in unlabeled data. Unlike in supervised learning tasks such as classification or regression, a key problem in cluster analysis is already a conceptionally and mathematically convincing definition of the learning goal. A widely but by no means generally accepted definition of clusters has its roots in [1], where clusters are described to be densely populated areas in the input space that are separated by less populated areas. The *non-parametric* mathematical translation of this idea usually assumes that the data $D = (x_1, \ldots, x_n) \in X^n$ is generated by some unknown probability measure P on a topological space X that has a density h with respect to some known reference measure μ on X. Given a threshold $\rho \geq 0$, the clusters are then defined to be the connected components of the density level set $\{h \geq \rho\}$.

Historically, two distinct questions have been investigated for this cluster definition. The first one is the so-called single level approach, which tries to estimate the connected components of $\{h \geq \rho\}$ for a *single and fixed* level $\rho \geq 0$. The single level approach has been studied by several authors, see, e.g., [4, 3, 7, 6, 8] and the references therein, and hence it seems fair to say that it already enjoys a reasonably good statistical understanding. Unfortunately, however, it suffers from a conceptional problem, namely that of determining a good value of ρ.

The second approach tries to address this by considering the hierarchical structure of the connected components for different levels. To be more precise, if h is a fixed density, which, for the sake of simplicity, is assumed to have *closed* density level sets, and A is a connected component of $\{h \geq \rho\}$, then, for every $r' \in [0, \rho]$, there exists exactly one connected component B of $\{h \geq \rho'\}$ with $A \subset B$. Under

some additional assumptions on μ and the density h, this then leads to a *finite* tree, in which each node B is a connected component of some level set $\{h \geq \rho'\}$ and all children of a node B are the connected components of $\{h \geq \rho\}$ for some $\rho > \rho'$ that are contained in B. We refer to [4, 9, 2, 10] for definitions and methods for estimating the structure of this tree. In particular, [2] show that in a weak sense of [5], a modified single linkage algorithm converges to this tree under some assumptions on the density h.

The goal of this work is to address the problem of the single level approach by presenting a simple algorithm that automatically approximates the smallest possible value of ρ for which the level set $\{h \geq \rho\}$ contains more than one component. In addition, the algorithm approximates the resulting components arbitrarily well for $n \to \infty$ under minimal and somewhat natural conditions, which include discontinuous densities. To this end, we first provide a definition for density level sets that make them actually independent of the chosen density. Note that this is necessary to deal with topological concepts such as connectivity without referring to a particular, and typically rather arbitrary choice of the density. This makes it mathematically rigorous to consider the infimum ρ^* over all levels ρ for which the corresponding density level sets contain more than one connected component. For simplicity, we then assume that there exists some $\rho^{**} > \rho^*$ such that the level sets for all $\rho \in (\rho^*, \rho^{**}]$ contain exactly *two* connected components. Note that the persistence of the cluster structure over a small range of levels is assumed either explicitly or implicitly in basically all density based clustering approaches. Moreover, while the restriction to *two* components seems to be quite restrictive at first glance, it turns out that it is actually more realistic than assuming more than two components. Finally, in dimensions greater than one, one more assumption on the level sets need to be made, namely one that excludes bridges and cusps that are too thin and long. However, while this is certainly unpleasant, it seems to be rather necessary, since such an assumption occurs in one form or the other in most articles dealing with density clustering. With these assumptions our main result then shows that a simple histogram based algorithm both approximates ρ^* and the resulting clusters.

References

[1] J.W. Carmichael, G.A. George, and R.S. Julius. Finding natural clusters. *Systematic Zoology*, 17:144–150, 1968.

[2] K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 343–351. 2010.

[3] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25:2300–2312, 1997.

[4] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[5] J.A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76:388–394, 1981.

[6] M. Maier, M. Hein, and U. von Luxburg. Generalized density clustering. *Optimal construction of k-nearest neighbor graphs for identifying noisy clusters*, 410:1749–1764, 2009.

[7] P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007.
[8] A. Rinaldo and L. Wasserman. Generalized density clustering. *Ann. Statist.*, 38:2678–2722, 2010.
[9] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.
[10] W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19:397–418, 2010.

Supervised Classification of Functional Data

Amparo Baíllo

The aim of supervised (binary) classification is to decide whether a random observation X (taking values in a feature space $\mathcal{F}$ endowed with a metric D) either belongs to a population P_0 or to another population P_1. The decision is based on the information provided by a training sample of correctly classified individuals $\mathcal{X}_n = \{(X_i, Y_i), 1 \leq i \leq n\}$. Here X_i, $i = 1, \ldots, n$, are independent replications of X measured on n randomly chosen individuals, $Y = 1_{\{X \in P_1\}}$ and 1_A denotes the indicator function on A.

The mathematical problem is to find a classifier $g_n(x) = g_n(x; \mathcal{X}_n)$, with $g_n : \mathcal{F} \to \{0, 1\}$, minimizing the classification error $L_n = \mathbb{P}\{g_n(X) \neq Y | \mathcal{X}_n\}$. The optimal classifier is the Bayes rule $g^*(x) = 1_{\{\eta(x) > 1/2\}}(x)$, where $\eta(x) = \mathbb{E}(Y|X = x)$. The Bayes error is $L^* = \mathbb{P}\{g^*(X) \neq Y\}$.

We are concerned here with the problem of discrimination of functional (or infinite-dimensional) data. The space $(\mathcal{F}, D)$ is assumed to be a separable metric space of functions. In [2] we review some differences between the finite- and the infinite-dimensional settings that difficult the straightforward generalization of finite-dimensional classification techniques to the infinite-dimensional framework. In [3] there is a survey on supervised and unsupervised classification with functional data.

Some existing functional classification methods are:
• adaptations of Fisher's linear rule: [11], [12], [13], [16].
• k-nearest neighbours (k-NN) rule: [5], [6].
• kernel rule: [1], [4]. Kernel and k-NN rules are particular cases of the plug-in methodology, where the unknown regression function η in the Bayes rule is replaced by an estimator $\hat{\eta}$.
• depth-based techniques. Some depth definitions applicable to the infinite-dimensional setting are: spatial depth ([7], [15]), integrated dual depth ([9], [10]), random Tukey depth ([8]), band depth ([14]) and h-modal depth ([9]).

In [2] it is shown that the Bayes rule can be explicitly computed for a class of Gaussian processes with "triangular" covariance functions. Estimating the unknown elements in the optimal rule yields parametric and nonparametric plug-in classifiers. Under certain assumptions [2] obtain convergence rates in probability to L^* for the error L_n of the nonparametric plug-in classifier. It would be

interesting to determine conditions under which depth-based procedures are also Bayes-risk consistent.

References

[1] C. Abraham, G. Biau and B. Cadre, *On the kernel rule for function classification*, Annals of the Institute of Statistical Mathematics **58** (2006), 619–633.

[2] A. Baíllo, J.A. Cuesta-Albertos and A. Cuevas, *Supervised classification for a family of Gaussian functional models*, Scandinavian Journal of Statistics (to appear) (2011).

[3] A. Baíllo, A. Cuevas and R. Fraiman, *Classification methods for functional data*, The Oxford Handbook of Functional Data Analysis, Oxford University Press (2011), 259–297.

[4] G. Biau, F. Bunea and M.H. Wegkamp, *Functional classification in Hilbert spaces*, IEEE Transactions in Information Theory **51** (2005), 2163-2172.

[5] F. Burba, F. Ferraty and P. Vieu, *k-nearest neighbors method in functional nonparametric regression*, Journal of Nonparametric Statistics **21** (2009), 453–469.

[6] F. Cérou and A. Guyader, *Nearest neighbor classification in infinite dimension*, ESAIM: Probability and Statistics **10** (2006), 340–355.

[7] P. Chaudhuri, *On a geometric notion of quantiles for multivariate data*, Journal of the American Statistical Association **91** (1996), 862-872.

[8] J.A. Cuesta-Albertos and A. Nieto-Reyes, *Functional classification and the random Tukey depth. Practical issues*, Combining Soft Computing and Statistical Methods in Data Analysis, Springer (2010), 123–130.

[9] A. Cuevas, M. Febrero and R. Fraiman, *Robust estimation and classification for functional data via projection–based depth notions*, Computational Statistics **22** (2007), 481–496.

[10] A. Cuevas and R. Fraiman, *On depth measures and dual statistics. A methodology for dealing with general data*, Journal of Multivariate Analysis **100** (2009), 753–766.

[11] P. Hall, D.S. Poskitt and B. Presnell, *A functional data-analytic approach to signal discrimination*, Technometrics **43** (2001), 1–9.

[12] G.M. James and T.J. Hastie, *Functional linear discriminant analysis for irregularly sampled curves*, Journal of the Royal Statistical Society Ser. B. **63** (2001), 533–550.

[13] B. Li and Q. Yu, *Classification of functional data: A segmentation approach*, Computational Statistics and Data Analysis **52** (2008), 4790-4800.

[14] S. López-Pintado and J. Romo, *On the concept of depth for functional data*, Journal of the American Statistical Association **104** (2009), 718-734.

[15] R. Serfling, *A depth function and a scale curve based on spatial quantiles*, Statistical data analysis based on the L_1L1-norm and related methods (2002), 25-38, Birkhäuser.

[16] H. Shin, *An extension of Fisher's discriminant analysis for stochastic processes*, Journal of Multivariate Analysis **99** (2008), 1191-1216.

Data Depth and Quantiles in Infinite Dimensional Spaces

Probal Chaudhuri

There are several versions of quantile and depth functions available in finite dimensional spaces. However, most of those finite dimensional versions do not have any meaningful and natural extension for data or distributions in infinite dimensional spaces. For instance, procedures based on simplices are restricted only to finite dimensional spaces, and procedures based on linear functions (e.g., Tukey's half-space depth and projection depth) do not have statistically meaningful extensions in infinite dimensional spaces. On the other hand, there are natural extensions of spatial quantile and associated rank and depth functions in infinite dimensional

spaces, and it can be shown that these extensions retain many of the interesting and useful properties of their finite dimensional counterparts.

Weighted-Mean Regions: Theory and Estimation

Karl Mosler
(joint work with Rainer Dyckerhoff and Pavel Bazovkin)

Weighted-mean regions are the level sets of a new class of depth functions, the weighted mean (WM) depth functions. They describe a probability distribution in Euclidean d-space regarding location, dispersion and shape, and they order given multivariate data with respect to their centrality. Also, they have a substantial interpretation in terms of multivariate set-valued risk measures that are coherent.

The talk introduces the class of weighted-mean regions and their principal properties: affine equivariance, nestedness, continuity in the parameter as well as in the distribution, subadditivity and monotonicity. The notion is illustrated with several special cases, among them the zonoid regions ([6], [8]) and the ECH (expected convex hull) trimming ([2]).

The weighted-mean regions of an empirical distribution are convex polytopes in $\mathbb{IR}^d$. A law of large numbers applies. Thus, given a sample, the empirical regions serve as natural estimates for the regions of the underlying probability distribution. In fact, the estimates can be computed for any dimension d by exact and approximate algorithms. They build on methods from computational geometry, by which the facets are characterized and their adjacency relations are found ([7]).

Applications to multivariate risk measurement ([3]) and stochastic linear programming are discussed.

The talk is based on joint work with Rainer Dyckerhoff ([4], [5]) and Pavel Bazovkin ([1]).

References

[1] P. Bazovkin, K. Mosler. *An exact algorithm for weighted-mean trimmed regions in any dimension*, submitted.

[2] I. Cascos. *The expected convex hull trimmed regions of a sample*, Computational Statistics and Data Analysis **22** (2007), 557-569.

[3] I. Cascos, I. Molchanov. *Multivariate risks and depth-trimmed regions*, Finance and Stochastics **11** (2007), 373-397.

[4] R. Dyckerhoff, K. Mosler. *Weighted-mean trimming of multivariate data*, Journal of Multivariate Analysis **102** (2011), 405-421.

[5] R. Dyckerhoff, K. Mosler. *Weighted-mean regions of a probability distribution*, submitted.

[6] G. Koshevoy, K. Mosler. *Zonoid trimming for multivariate distributions*, Annals of Statistics **25** (1997), 1998-2017.

[7] K. Mosler, T. Lange, P. Bazovkin. *Computing zonoid trimmed regions in dimension* $d > 2$, Computational Statistics and Data Analysis **53** (2009), 2500 - 2510.

[8] K. Mosler. *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach.* New York (Springer), 2002.

Data Depth for Regression and Autoregressive Models

Christine H. Müller

Data depth was generalized to regression by Rousseeuw and Hubert (1999) via the concept of nonfit. Replacing the squared or absolute residuals in the concept of nonfit by an arbitrary quality function, Mizera (2002) introduced a very general concept of depth. Using a likelihood function as quality function leads to likelihood depth, considered by Mizera and Müller (2004), Müller (2005), Denecke and Müller (2009, 2011a,b). Although a likelihood function is given by a parametric model, the resulting likelihood depth notion is often distribution-free. This holds in particular for regression as Müller (2005) showed.

Additionally, any depth notion can be used as simplicial depth as Liu (1988, 1990) did with the half space depth of Tukey (1975). The simplicial depth is an U-statistic so that in principle its asymptotic distribution is known and tests can be derived. However, this U-statistic is often degenerated so that the spectral decomposition of a conditional expectation is needed and the asymptotic distribution depends on the eigen values of this decomposition. These spectral decompositions were derived for linear and quadratic regression in Müller (2005), for general polynomial regression in Wellmann et al. (2009), for multiple regression in Wellmann and Müller (2010a), and for orthogonal regression in Wellmann and Müller (2010b). Usually it is often not easy to derive this spectral decomposition. One exception is the case of one unknown parameter as for linear regression through the origin.

However, it is also possible that the simplicial depth is not a degenerated U-statistic. This is the case when the underlying depth notion provides biased estimators as Denecke and Müller (2009, 2011a) showed. Then the asymptotic distribution is simply the normal distribution. But a bias correction is needed not only for the estimators but also for the tests to avoid very bad power of the tests for some alternatives. These bias corrections were derived in Denecke and Müller (2009, 2011b,c) for one-parametric copulas and for the Weibull distribution. In Denecke and Müller (2011a), a general theory is provided to obtain consistent tests and estimators based on such depth notions. However, up to now the general theory concerns only the case that one parameter is unknown.

I want to discuss the following extensions of the above approaches:
a) How to generalize the bias correction to the case of two or more unknown parameters. In particular, the generalization to regression with exponential or Weibull distributed observations shall be discussed since this is often used in life time experiments.
b) How to extend the approach to dependent data, in particular to autoregressive models. It seems that at least in the AR(1) model, where only one parameter is unknown, the theory for linear regression through the origin could be used.

References

[1] Denecke L. and Müller, Ch.H. (2009) *Robust estimators and tests for copulas based on likelihood depth*, in revision for CSDA.

[2] Denecke L. and Müller, Ch.H. (2011a) *Consistency and robustness of tests and estimators based on depth*, submitted.
[3] Denecke L. and Müller, Ch.H. (2011b) *Robust estimation of the parameters of the Weibull distribution in complete and censored data*, submitted.
[4] Denecke L. and Müller, Ch.H. (2011c) *Robust testing of the parameters of the Weibull distribution in complete and censored data*, submitted.
[5] Liu, R.Y. (1988) *On a notion of simplicial depth*, Proc. Nat. Acad. Sci. USA **85**, 1732-1734.
[6] Liu, R.Y. (1990) *On a notion of data depth based on random simplices*, Ann. Statist. **18**, 405-414.
[7] Mizera, I. (2002) *On depth and deep points: a calculus*, Ann. Statist. **30**, 1681-1736.
[8] Mizera, I. and Müller, Ch.H. (2004) *Location-scale depth*, J. Am. Stat. Assoc. **99**, 949-989.
[9] Müller, Ch.H. (2005) *Depth estimators and tests based on the likelihood principle with applications to regression*, J. Multivariate Anal. **95**, 153-181.
[10] Rousseeuw, P.J. and Hubert, M. (1999) *Regression depth (with discussion)*, J. Amer. Statist. Assoc. **94**, 388-433.
[11] Tukey, J.W. (1975) *Mathematics and the picturing of data*, In Proc. International Congress of Mathematicians 2, Vancouver 1974, 523- 531.
[12] Wellmann, R., Harmand, P. and Müller, Ch.H. (2009) *Distribution-free tests for polynomial regression based on simplicial depth*, J. Multivariate Anal. **100**, 622-635.
[13] Wellmann, R. and Müller, Ch.H. (2010a) *Tests for multiple regression based on simplicial depth*, J. Multivariate Anal. **101**, 824-838.
[14] Wellmann, R. and Müller, Ch.H. (2010b) *Depth notions for orthogonal regression*, to appear in J. Multivariate Anal., DOI: 10.1016/j.jmva.2010.06.008.

Smooth Tree-Based Level Set Estimation

Gabriel Chandler
(joint work with Leif Johnson)

A level set of a regression function $\mathcal{S}$ is defined as the set on which the response surface exceeds some threshold γ. We assume we observe the model $Y = m(X)+\epsilon$ where $Y \in \Re$ and $X \in \Re^d$ and E(ϵ)=0. Thus, $\mathcal{S}_\gamma = \{x : m(x) \geq \gamma\}$. Willet and Nowak (2005) proposed estimation of the level set via a tree based method based on a dyadic partition of the unit cube, which is assumed to be the support of the X. A pruning step is implemented, in which each branch (partition) of the tree (space) is considered for removal for the tree. This consideration accounts for how well this split explains the data and how small the branch is. More formally, the method minimizes a weighted average of the empirical risk $\hat{R}(T)$ and a regularization term that measures the complexity of the tree $\Phi(T)$, where T is of the class of all possible dyadic trees with a bounded number of splits in each dimension. The penalty term is weighted by a parameter ρ, which is difficult to select. We propose a non-linear mapping of a region our data space back into the d-dimensional hypercube such that relatively simple trees in this new space correspond to good approximations of the boundary of the level set in our original space. We conjecture that the selection of ρ is not so crucial should we only need relatively simple trees (few branches) to attain a good estimate. As the simplest (non-degenerate) tree has a single split in a single dimension, we would like to map our boundary, a $d-1$ dimensional manifold in such a way that it corresponds

to a $d-1$ dimensional hyperplane with height $1/2$ in the d^{th} dimension (i.e. the boundary corresponding to the simplest tree). To do this, we first take a rough tree based estimate (only choosing ρ so the resulting estimate is connected, which we assume to be true) to get an idea of the orientation of the true boundary. We then smooth this boundary slightly, and construct an invertible map of the space near the estimated boundary such that the boundary corresponds to the hyperplane mentioned above, with points lying inside (outside) the estimated set falling above (below) $1/2$ in the d^{th} dimension. In the new space, a tree based estimate is fit, and this is mapped back into our original data space. This is then itself smoothed, and the map is reapplied. This iterative procedure is continued for a fixed number of iterations with the final estimate being the one with the smallest empirical risk. The resulting estimate is not only smooth, but the smoothness is locally adaptive (as smoothness is counteracted in the new space by more complicated trees being fit), and automatic, as the only smoothing parameter chosen was the minimal smoothing applied at each step.

References

[1] R. Willett, R. Nowak *Minimax optimal level set estimation*, Proc. SPIE, Wavelets XI **16** (2005), 2965–2979.

DD-Classifier: Nonparametric Classification Procedure Based on DD-plot

Jun Li

(joint work with Juan A. Cuesta-Albertos, Regina Y. Liu)

Classification is one of the most practical subjects in statistics. It has many important applications in different fields. Many existing classification algorithms assume either certain parametric distributions for the data or certain forms of separating curves or surfaces. These parametric classifiers are suboptimal and of limited use in practical applications where little information about the underlying distributions is available a priori. In comparison, nonparametric classifiers are usually more flexible in accommodating different data structures, and are hence more desirable. In the last two decades, data depth has emerged as a powerful nonparametric analysis tool in various areas of multivariate statistics. It has offered several promising solutions to classification problems. For instance, Christmann and Rousseeuw (2001) and Christmann, et al. (2002) applied the idea of regression depth (see Rousseeuw and Huber, 1999) to classification. Ghosh and Chaudhuri (2005a) used half-space depth and regression depth to construct linear and nonlinear separating curves or surfaces. In those depth based methods, a finite dimensional parametric form (usually linear or quadratic) for the separating surface is often assumed. Thus, these classifiers are not fully nonparametric. Ghosh and Chaudhuri (2005b) subsequently proposed the maximum depth classifier, which assigns the observation to the group for which it attains the highest depth value. This classification rule is intuitively appealing and fully nonparametric, but it performs well only when the

populations differ in location only and the prior probabilities of the populations are equal. Recently, Cui et al. (2008) considered a maximum depth classifier based on a modified projection depth. However, this classifier appears to work well only under normal settings.

Using the *DD*-plot (depth-versus-depth plot), we introduce a new nonparametric classification algorithm and call it a *DD*-classifier. The algorithm is completely nonparametric, and requires no prior knowledge of the underlying distributions or of the form of the separating curve. Thus it can be applied to a wide range of classification problems. The algorithm is completely data driven and its classification outcome can be easily visualized on a two-dimensional plot regardless of the dimension of the data. Moreover, it is easy to implement since it bypasses the task of estimating underlying parameters such as means and scales, which is often required by the existing classification procedures. We study the asymptotic properties of the proposed *DD*-classifier and its misclassification rate. Specifically, we show that it is asymptotically equivalent to the Bayes rule under suitable conditions. The performance of the classifier is also examined by using simulated and real data sets. Overall, the proposed classifier performs well across a broad range of settings, and compares favorably with existing classifiers. Finally, it can also be robust against outliers or contamination.

References

[1] Christmann, A., Fischer, P., and Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, **17**, 273–287.

[2] Christmann, A. and Rousseeuw, P. J. (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis*, **37**, 65–75.

[3] Cui, X., Lin, L., and Yang, G. R. (2008). An extended projection data depth and its applications to discrimination. *Communications in Statistics-Theory and Methods*, **37**, 2276–2290.

[4] Ghosh, A. K. and Chaudhuri, P. (2005a). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, **11**, 1–27.

[5] Ghosh, A. K. and Chaudhuri, P. (2005b). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**, 327–350.

[6] Rousseeuw, P. J. and Hubert, M. (1999). Regression depth (with discussion). *Journal of the American Statistical Association*, **94**, 388–402.

On Statistical Properties of Sets Fulfilling Rollling–Type Conditions

Ricardo Fraiman

(joint work with Antonio Cuevas and Beatriz Pateiro)

Motivated by set estimation problems, we consider three closely related shape conditions for compact sets: positive reach, r-convexity and rolling condition. The first one (introduced by Federer (1959)) is maybe the most popular one, due to its relevant role in geometric measure theory. First, the relations between these shape conditions are analyzed. Second, a result of "full consistency" (i.e., consistency with respect to the usual set distances, plus boundary Hausdorff-convergence) is obtained for the estimation of sets fulfilling a rolling condition. Third, the class of

uniformly bounded compact sets whose reach is not smaller than a given constant r is shown to be a P-uniformity class (in Billingsley and Topsøe's (1967) sense) and, in particular, a Glivenko-Cantelli class. Fourth, under broad conditions, the r-convex hull of the sample is proved to be a fully consistent estimator of an r-convex support in the two-dimensional case. Moreover, its boundary length is shown to converge (a.s.) to that of the underlying support. This provides a simple efficient estimator of the boundary length based on just one inner sample. Fifth, the above results are applied to get new consistency statements for level set estimators based on the excess mass methodology (Polonik, 1995).

References

[1] Ambrosio, L., Colesanti, A. and Villa, E. (2008). Outer Minkowski content for some classes of closed sets. *Math. Ann.* **342** 727–748.

[2] Baíllo, A. and Cuevas, A. (2001). On the estimation of a star-shaped set. *Adv. in Appl. Probab.* **33** 717–726.

[3] Billingsley, P. and Topsøe, F. (1967). Uniformity in weak convergence. *Z. Wahrs. und Verw. Gebietea* **7** 1–16.

[4] Bickel, P. J. and Millar, P. W. (1992). Uniform convergence of probability measures on classes of functions. *Statist. Sinica* **2** 1–15.

[5] Colesanti, A. and Manselli, P. (2010). Geometric and isoperimetric properties of sets of positive reach in $\mathbf{E}^d$. *Preprint*

[6] Cuevas, A. and Fraiman, R. (2009). *Set estimation.* In *New Perspectives on Stochastic Geometry.* W.S. Kendall and I. Molchanov, eds. Oxford University Press, 366–389.

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

[8] Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* **36** 340–354.

[9] Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.* **28** 384–393.

[10] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491.

[11] Jiménez, R. and Yukich, J. (2010) Nonparametric estimation of surface integrals. *To appear in Ann. Statist.*

[12] Mason, D.M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142.

[13] Mattila, P.(1995) *Geometry of sets and measures in Euclidean spaces. Fractals and rectifiability* Cambridge University Press, Cambridge.

[14] Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

[15] Pateiro-López, B. and Rodríguez-Casal, A. (2010) Generalizing the convex hull of a sample: The R package alphahull. *J. Stat. Softw.* **5** 1–28.

[16] Pateiro-López, B. and Rodríguez-Casal, A. (2008) Length and surface area estimation under smoothness restrictions. *Adv. in Appl. Probab.* **40** 348–358.

[17] Perkal, J. (1956). Sur les ensembles ϵ-convexes. *Colloq. Math.* **4** 1–10.

[18] Pollard, D. (1984). *Convergence of stochastic processes.* Springer–Verlag, New York.

[19] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **33** 855–881.

[20] Polonik, W. and Wang, Z. (2005). Estimation of regression contour clusters—an application of the excess mass approach to regression *J. Multivariate Anal.* **94** 227–249.

[21] Rataj, J. (2005). On boundaries of unions of sets with positive reach. *Beiträge Algebra Geom.* **46** 397–404.

[22] Reitzner, M. (2009). *Random polytopes.* In *New Perspectives on Stochastic Geometry.* W.S. Kendall and I. Molchanov, eds. Oxford University Press, 45–75.
[23] Rodríguez-Casal, A. (2007). Set estimation under convexity-type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.* **43** 763–774.
[24] van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.
[25] Walther, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299.
[26] Walther, G. (1999). On a generalization of Blaschke's rolling theorem and the smoothing. *Math. Methods Appl. Sci.* **22** 301–316.

Robust Classification for Skewed Data

Mia Hubert
(joint work with Stephan Van der Veeken, Irène Gijbels)

In the first part of the talk, we propose several classification rules for skewed distributions. They are based on the adjusted outlyingness (AO), as introduced in Brys et al. (2005) and applied to outlier detection in Hubert and Van der Veeken (2008). The new rules combine ideas of AO with the classification methods proposed in Ghosh and Chaudhuri (2005) and Billor et al. (2008). The first classifier is described in Hubert and Van der Veeken (2010) and assigns a new observation to the group to which it attains the minimal adjusted outlyingness. The other two rules adjust for the group sizes and perform better when the group sizes are unequal.

In the second part of the talk, we investigate how we can reduce the mean squared error of the medcouple, a robust estimator of skewness, which is needed to compute the adjusted outlyingness. As already theoretically pointed out by Fernholz (1997), smoothing the empirical distribution function with an appropriate kernel and bandwidth can reduce the variance and mean squared error of some quantile-based estimators in small data sets. We apply this idea on several robust estimators of location, scale and skewness. We also propose a robust bandwidth selection procedure and show that the use of that bandwidth indeed often leads to smaller MSEs.

References

[1] N. Billor, A. Abebe, A. Turkmen and S.V. Nudurupati, *Classification based on depth transvariations*, Journal of Classification **25** (2008), 249–260.
[2] G. Brys, M. Hubert and A. Struyf, *A robust measure of skewness,* Journal of Computational and Graphical Statistics **13** (2004), 996–1017.
[3] G. Brys, M. Hubert and P.J. Rousseeuw, *A robustification of Independent Component Analysis*, Journal of Chemometrics **19** (2005), 364–375.
[4] L.T. Fernholz, *Reducing the variance by smoothing*, Journal of Statistical Planning and Inference **57** (1997), 29–38.
[5] M. Hubert and S. Van der Veeken, *Outlier detection for skewed data*, Journal of Chemometrics **22** (2008), 235–246.
[6] M. Hubert and S. Van der Veeken, *Robust classification for skewed data*, Advances in Data Analysis and Classification **4** (2010), 239–254.
[7] A.K. Ghosh and P. Chaudhuri, *On maximum depth and related classifiers*, Scandinavian Journal of Statistics. Theory and Applications **32** (2005), 327–350.

Estimation of Extreme Risk Regions Under Multivariate Regular Variation

John H. J. Einmahl

(joint work with Juan-Juan Cai, Laurens de Haan)

When considering d possibly dependent random variables with joint probability law P one is often interested in extreme risk regions, with very small probability p. We consider risk regions Q of the form $\{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta\}$, where f is the density corresponding to P and β a small number, determined by $PQ = p$. Such a region has the property that it consists of the less likely points and hence that its complement is as small as possible.

The values of p we consider are typically of order $1/n$. This means that the number of data points that fall in Q is small and can even be zero, i.e. we are extrapolating outside the sample. This lack of relevant data points makes estimation difficult. The estimation of Q is a multivariate analogue of the estimation of extreme quantiles in the univariate setting, see, e.g. de Haan and Ferreira (2006), Chapter 4. The multivariate case is much more complicated, however, since we have to estimate a set instead of a number.

Having an estimate of Q can be important in various settings. E.g., it can be used as an alarm system in risk management: if a new observation falls in the estimated Q it is a signal of extreme risk. See Einmahl, Li, and Liu (2009) for an application to aviation safety along these lines.

Given a random sample of multivariate regularly varying random vectors with law P, we construct a "statistics of extremes" estimator $\widehat{Q}_n$ of Q. When $p \to 0$, we prove that, under certain conditions,

$$P(\widehat{Q}_n \triangle Q)/p \xrightarrow{\mathbb{P}} 0, \quad \text{as } n \to \infty,$$

and hence

$$P(\widehat{Q}_n)/p \xrightarrow{\mathbb{P}} 1.$$

Obviously $\widehat{Q}_n$ depends on p. Starting from a very small $\widehat{Q}_n$ we can enlarge it until it first hits an observation. This observation is the "largest" one and it has a "p-value" attached to it. This is helpful in deciding whether an observation is the most atypical (or: an outlier). Also, by continuing this procedure we can introduce a ranking of the larger observations.

In a detailed simulation and comparison study the good performance of the procedure is demonstrated. We also apply our estimator to financial data.

References

[1] Einmahl, J.H.J., Li, J. and Liu, R.Y., *Thresholding events of extreme in simultaneous monitoring of multiple risks*, J. Amer. Statist. Assoc. **104** (2009), 982–992.

[2] de Haan, L. and Ferreira, A., *Extreme Value Theory: An Introduction*, Springer, New York.

Reporter: Mia Hubert

Participants

Prof. Dr. Amparo Baillo
Departamento de Matematicas
Universidad Autonoma de Madrid
Ciudad Universitaria de Cantoblanco
E-28049 Madrid

Prof. Dr. Gabriel Chandler
Department of Mathematics
Pomona College
610 N. College Ave.
Claremont , CA 91711-6348
USA

Prof. Dr. Probal Chaudhuri
Theoretical Statistics and
Mathematics Unit
Indian Statistical Institute
203, B.T. Road
Calcutta 700 108
INDIA

Prof. Dr. Xin Dang
Department of Mathematics
University of Mississippi
University , MS 38677
USA

Prof. Dr. John H.J. Einmahl
Department of Econometrics
Tilburg University
P. O. Box 90153
NL-5000 LE Tilburg

Prof. Dr. Ricardo Fraiman
Departamento de Economia y Matematicas
Universidad de San Andres
Vito Dumas 284
Buenos Aires
ARGENTINA

Prof. Dr. Daniel Hlubinka
Department of Probability and
Mathematical Statistics
Charles University
Sokolovska 83
18675 Praha 8
CZECH REPUBLIC

Prof. Dr. Mia Hubert
Afdeling Statistiek
Katholieke Universiteit Leuven
B-3001 Heverlee

Prof. Dr. Jun Li
Department of Statistics
University of California
Riverside , CA 92521-0138
USA

Prof. Dr. Regina Y. Liu
Department of Statistics
Rutgers University
501 Hill Center
110 Frelinghuysen Road
Piscataway , NJ 08854-8019
USA

Prof. Dr. Karl Clemens Mosler
Seminar f. Wirtschafts-u.Sozialstatistik
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln

Prof. Dr. Christine H. Müller
Fakultät für Statistik
Technische Universität Dortmund
44221 Dortmund

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

Prof. Dr. Clayton Scott
Electrical Eng. & Comp. Science Dept.
The University of Michigan
Ann Arbor , MI 48109-2122
USA

Prof. Dr. Robert J. Serfling
Department of Mathematical Sciences
University of Texas at Dallas
Richardson , TX 75080
USA

Prof. Dr. Ingo Steinwart
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
70569 Stuttgart

Kaveh Vakili
Departement Wiskunde
Faculteit der Wetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200B
B-3001 Leuven

Prof. Dr. Günther Walther
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA