# Some topics

Vladimir Spokoiny*

Weierstrass-Institute,
Humboldt University Berlin,
Moscow Institute of Physics and Technology
Mohrenstr. 39, 10117 Berlin, Germany,

spokoiny@wias-berlin.de

September 30, 2017

**Abstract**

The paper contains a list of open problems to study and projects to develop. The list includes the properties of the parametric bootstrap in the problem of testing a composite parametric hypothesis

*AMS 2000 Subject Classification:* Primary 62F10. Secondary
*Keywords*: test, size, critical value, composite hypothesis, Talagrand concentration bound

## Community detection with adaptive weights

Consider a random graph with the vertex set $V$ and an edge matrix $\boldsymbol{Y} = (Y_{ij})$: $Y_{ij} = 1$ means an edge between the vertices $i, j$, otherwise $Y_{ij} = 0$. The Erdos-Renyi model assumes that all $Y_{ij}$ are independent Bernoulli with the success probability $p_{ij} \in [0, 1]$. The stochastic block model means that the set of vertices $V$ can be split into blocks $V_1, \ldots, V_M$ and

$$p_{ij} = q_{st} \qquad i \in V_s,\ j \in V_t\,,$$

for some values $q_{st}$, $s, t = 1, \ldots, M$. The goal is to apply the methods of adaptive weights AWCD (community detection) to recover the community set from the observed

*

edge matrix $\boldsymbol{Y}$. Apriori the number of blocks $M$ as well as the block probability values $q_{st}$ are unknown.

Workpackages:

1. Efficient scalable implementation of the AWCD that would work for large datasets
2. Exploring the theoretical properties of the AWCD including propagation, separation, and consistency
3. Extending to the case of overlapping communities
4. Application to social, medicine, bio, and financial data

Literature:

- Efimov, Adamyan, Spokoiny (2017) Adaptive nonparametric clustering. arxiv 1709.09102.
- Cristopher Moore (2017) The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. arXiv:1702.00467
- arXiv:1507.04118 Oracle inequalities for network models and sparse graphon estimation Olga Klopp (MODAL'X, CREST), Alexandre B. Tsybakov (CREST), Nicolas Verzelen (MISTEA)

Contact: Maxim Panov, Igor Silin, Kirill Efimov, Alexey Naumov, VS

## Semisupervised learning

Suppose to be given a labeled set of objects (samples) $(X_i, Y_i)$ for $i = 1, \ldots, n$ and an unlabelled set $X_j$, $j = n{+}1, \ldots, n{+}m$. The aim is to develop a method which utilizes the unlabeled set $(X_j)$ to improve the quality of classification. The idea of label propagation suggests to perform a clustering procedure at the first step of the algorithm using the whole set $X_1, \ldots, X_{n+m}$. Then each cluster is labeled due to majority of labeled data within this cluster. This procedure AWSL (semisupervised learning) is implemented in terms of the weight matrix $W_{ij}^{(k)}$ obtained from the clustering procedure and the vector $\widetilde{\boldsymbol{\theta}}^{(k)} = (\widetilde{\theta}_i^{(k)})$ which estimates the success probabilities $\theta_i = I\!P(Y_i = 1)$ for all $i$.

Workpackages:

1. Efficient scalable implementation of the AWSL for training set of large size
2. Retraining with the testing dataset

3. Exploring the theoretical properties of the AWSL including propagation, separation, and consistency

4. Application to tracking problem: given a collection of screenshots $\boldsymbol{X}(t), \boldsymbol{Y}(t)$, track a possibly moving object descibed by the labels $\boldsymbol{Y}(t)$

5. Application to social, media, bio, medicine, financial data

Literature: Efimov, Adamyan, Spokoiny (2017) Adaptive nonparametric clustering. arxiv 1709.09102.

Contact: Maxim Panov, Igor Silin, Kirill Efimov, VS

## Comparison of random graphs

Suppose to be given two (or more) random graphs $\boldsymbol{Y} = (Y_{ij})$ and $\boldsymbol{Z} = (Z_{ij})$ over the same set of vertices $V = \{1, \ldots, n\}$. The null hypothesis of stationarity means that the underlying edge probability matrices $\boldsymbol{A} = (a_{ij}) = \big(I\!P(Y_{ij} = 1)\big)$ and $\boldsymbol{B} = (b_{ij}) = \big(I\!P(Z_{ij} = 1)\big)$ coincide. The test procedure AWGC (graph comparison) can compare the stochastic block structure for both matrices: run AWCD for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ independently and then compare the obtained results.

Workpackages:

1. Efficient scalable implementation of the AWGS based on AWCD

2. Calibration for the given first kind error and for different distances between the estimated vectors $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{B}}$

3. Exploring the theoretical properties of the AWGS including power and separation rate,

4. Application to financial, economic, and energy market

Literature: Efimov, Adamyan, Spokoiny (2017) Adaptive nonparametric clustering. arxiv 1709.09102.

Contact: Maxim Panov, Igor Silin, Kirill Efimov, VS

## Two sample test for high dimensional data using Monge-Kantorovich transform

Consider two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ in $I\!R^d$ of the same size $n$. We suppose that all observations are independent, and each sample is i.i.d. with the underlying

distribution $P$ resp. $Q$ The null hypothesis means $P = Q$. The test procedure uses the the recent idea of mapping the pooled dataset $\boldsymbol{X}, \boldsymbol{Y}$ into the uniform discrete set on the unit ball by the Monge-Kantorovich transform.

Workpackages:

1. Efficient scalable implementation of the test working for large dimension $d$
2. Application to change point detection problem
3. Application to tracking problem
4. Exploring the theoretical properties of the AWSL including propagation, separation, and consistency

Literature: Efimov, Adamyan, Spokoiny (2017) Adaptive nonparametric clustering. arxiv 1709.09102.

Contact: Alexanda Suvorikova, Alexey Kroshnin, Andrey Sobolevskii, VS

## Efficient dimension reduction

Given a collection of high dimensional vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, identify the efficient dimension $m^*$ and the projector $\Pi_{m^*}$ such that $\boldsymbol{Y}_i = \Pi_{m^*} \boldsymbol{Y}_i$ up to a measurement error for all $i$. The basic assumption is that the spectral gap $\lambda_{m^*} - \lambda_{m^*+1}$ is sufficiently large (larger that $\mathtt{C} n^{-1/2}$).

The procedure uses the idea of building a confidence set for the projector $\Pi_m$ for each candidate efficient dimension $m$ using the bootstrap procedure from Naumov, Spokoiny, Ulyanov (2017). The size of the confidence set is the crucial characteristics:

- for $m = m^*$, the eigen-subspace can be recovered with accuracy $1/\sqrt{n}$;
- for $m < m^*$, if $\lambda_m - \lambda_{m+1}$ is sufficiently big, again the confidence width is of order $n^{-1/2}$;
- for $m < m^*$ but $\lambda_m - \lambda_{m+1}$ is small or zero, then $\Pi_m$ can be recovered up to an error within the e.d.r. subspace $\mathcal{I}$ of dimension $m^*$. This yields the confidence width of order $m^*$;
- if $m > m^*$, then the confidence width is of order $(m - m^*) * (d - m^*)$

Workpackages:

1. Efficient implementation working for large dimension $d$ and big $n$

2. Exploring the theoretical properties: consistency and accuracy of estimation for the efficient dimension $m^*$ and the corresponding eigen-subspace;

3. Application financial data

4. Application bio and medical data

5. Extension to a non-Gaussian case

Literature: Naumov, Spokoiny, Ulyanov (2017) Bootstrap confidence sets for spectral projectors of sample covariance. arXiv:1703.00871.

Contact: Alexey Naumov, Igor Silin, Alexander Podkopaev, VS

## Gaussian approximation and bootstrap confidence set

Consider a parametric problem for an independent sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$: $\boldsymbol{Y} \sim I\!P \in (I\!P_{\boldsymbol{\theta}})$. The MLE is computed by maximizing $L(\boldsymbol{\theta}) = \log(dI\!P_{\boldsymbol{\theta}}/d\mu)$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell_i(Y_i, \boldsymbol{\theta})$$

where $\ell_i(Y_i, \boldsymbol{\theta})$ is the log-likelihood for one observation $Y_i$. The bootstrap log-likelihood $L^\flat(\boldsymbol{\theta})$ and the bootstrap estomate $\widetilde{\boldsymbol{\theta}}^\flat$ are defined by reweighting the log-likelihood

$$\widetilde{\boldsymbol{\theta}}^\flat = \operatorname*{argmax}_{\boldsymbol{\theta}} L^\flat(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell_i(Y_i, \boldsymbol{\theta}) w_i^\flat$$

with independent random weights $w_i^\flat$. The bootstrap confidence width $z^\flat$ is defined by the condition

$$I\!P^\flat\Big(L^\flat\big(\widetilde{\boldsymbol{\theta}}^\flat\big) - L^\flat\big(\widetilde{\boldsymbol{\theta}}\big) > z^\flat\Big) = \alpha$$

The bootstrap confidence set is

$$\mathcal{E}(z^\flat) = \big\{\boldsymbol{\theta}\colon L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le z^\flat\big\}$$

Bootstrap validity requires to evaluate the resulting coverage probability for the true parameter $\boldsymbol{\theta}^*$:

$$\big|I\!P\big(\boldsymbol{\theta}^* \in \mathcal{E}(z^\flat)\big) - \alpha\big| \le \diamondsuit_n$$

Existing results ensure quite slow approximation rate (AR) $(n/p^3)^{-1/8}$, where $p$ is the parameter dimension.

Workpackages:

1. Improve the AR to $(n/p^a)^{-1/2}$ or even $(n/p^a)^{-1}$. In particular, study the impact of the dimension $p$. The current guess is $a = 3/2$.

2. Explore some particular case including Generalized Linear Models and Instrumental Variables

3. Efficient implementation for special problems and applications to benchmark data.

Contact: Alexey Naumov, Arshak Minasyan, VS

## Adaptive topological data analysis

With the emergence of new geometric inference and algebraic topology tools, computational topology has recently seen an important development toward data analysis, giving birth to the field of Topological Data Analysis (TDA) whose aim is to infer relevant, multiscale, qualitative and quantitative topological structures directly from the data. Topological persistence, more precisely persistent homology appears as a fundamental tool for TDA. In TDA, persistent homology has found applications in many fields, including neuroscience [SIN08], bioinformatics [KAS07], shape classification [CHA09], sensor networks [DE07] or signal processing [BAU14]. It is usually computed for a filtered simplicial complex built on top of the available data, i.e. a nested family of simplicial complexes whose vertex set is the data set. The obtained persistence diagrams are then used as "topological signatures" to exhibit and compare the topological structure underlying the data. The relevance of this approach relies on stability results ensuring that close data sets have close persistence diagrams [CHA14a]. However these results are not statistical and thus only provide heuristic or exploratory uses in data analysis. This work package will pursue an approach, which extends persistent diagrams to make the method adaptive to the unknown topological feature of the data including low dimensional manifold and clustering structure.

Several recent attempts have been made to study persistence diagrams from a statistical point of view, such as [MIL11] who study probability measures on the space of persistence diagrams or [BUB12] who introduces a functional representation of persistence diagrams, the so-called persistence landscapes, allowing means and variance of persistence diagrams to be defined. [FAS14] observed that persistence diagram inference is strongly connected to the better known problem of support estimation. As far as we know, only few results about support estimation in general metric spaces have been given in the past, see e.g. [CHA14b] that allow to infer persistent homology information from data corrupted by different kind of noise. Although it is attracting more and more interest, the use of persistent homology in data analysis remains widely heuristic.

The goal is to develop a new approach to TDA, which would enable to consistently separate between topological features and topological noise and which be adaptive to the unknown topological structures like manifolds or clusters.

Workpackages:

1. Extend the structure adaptive Clustering procedure from [SPO17] to persistent diagrams (AWPD).
2. Test the method with artificial and real data, focusing on sensitivity to the structures in data and robustness to the noise.
3. Develop a scalable und numerically effizient algorithm
4. apply to real datasets like texts, images, or videos.
5. Establish some theoretical results on the properties of AWPD, in particular efficiency of structural recovery.

Literature:

BAU14 Bauer, Ulrich, et al. "Persistent homology meets statistical inference-a case study: Detecting modes of one-dimensional signals."arXiv preprint ArXiv:1404.1214(2014).

BUB12 Bubenik, Peter. "Statistical topology using persistence landscapes. arXiv preprint."arXiv preprint arXiv:1207.64373 (2012).

CHA09 Chazal, Frédéric, et al. "Gromov-Hausdorff Stable Signatures for Shapes using Persistence."Computer Graphics Forum. Vol. 28. No. 5. Blackwell Publishing Ltd, 2009.

CHA14a Chazal, Frédéric, Vin De Silva, and Steve Oudot. "Persistence stability for geometric complexes."Geometriae Dedicata173.1 (2014): 193-214.

CHA14b Chazal, Frédéric, et al. "Robust topological inference: Distance to a measure and kernel distance."arXiv preprint arXiv:1412.7197(2014).

DE07 De Silva, Vin, and Robert Ghrist. "Homological sensor networks."Notices of the American mathematical society54.1 (2007).

FAS14 Fasy, Brittany Terese, et al. "Confidence sets for persistence diagrams."The Annals of Statistics42.6 (2014): 2301-2339.

KAS07 Kasson, Peter M., et al. "Persistent voids: a new structural metric for membrane fusion."Bioinformatics23.14 (2007): 1753-1759.

MIL11 Mileyko, Yuriy, Sayan Mukherjee, and John Harer. "Probability measures on the space of persistence diagrams."Inverse Problems27.12 (2011): 124007.

POL06 Polzehl, Joerg, and Vladimir Spokoiny. "Propagation-separation approach for local likelihood estimation."Probability Theory and Related Fields135.3 (2006): 335-362.

SIN08 Singh, Gurjeet, et al. "Topological analysis of population activity in visual cortex."Journal of vision8.8 (2008): 11-11.

SPO17 Spokoiny, Vladimir, Efimov, Kirill and Adamyan Larisa "Adaptive weights clustering."Manuscript (2017).

Contact: Alexey Naumov, Kirill Efimov, VS

## Inference for HMM

- Alexandrovich, G., Holzmann, H. and Leister, A. (2016), Nonparametric identification and maximum likelihood estimation for hidden Markov models, Biometrika, in press, DOI:10.1093/biomet/asw001.
- Dannemann, J. (2012), Semiparametric hidden Markov models, Journal of Computational and Graphical Statistics, 21, 677?692.
- de Souza, C.P.E. and Heckman, N.E. (2014), Switching nonparametric regression models, Journal of Nonparametric Statistics, 26, 617?637.
- Gassiat, E., Cleynen, A. and Robin, S. (2016), Inference in finite state space non parametric Hidden Markov Models and applications, Statistics and Computing, 26, 61?71.
- Hamilton, J.D. (2008), Regime-switching models
- Yau, C., Papaspiliopoulos, O., Roberts, G.O. and Holmes, C. (2011), Bayesian non-parametric hidden Markov models with applications in genomics, Journal of the Royal Statistical Society: Series B, 73, 37?57.

## Subset selection

Butucea, C., Stepanova, N.A., and Tsybakov, A.B. (2017) Variable selection with Hamming loss

## Semisupervised learning

Seeger (2001), Zhu (2005), Rigollet (2007)

## Low rank matrix estimation

- arXiv:1509.00319 Estimation of matrices with row sparsity O. Klopp (CREST, MODAL'X), A.B. Tsybakov (CREST)
- arXiv:1412.8132 Robust Matrix Completion Olga Klopp (MODAL'X, CREST), Karim Lounici, Alexandre B. Tsybakov (CREST)
- arXiv:1011.6256 Nuclear norm penalization and optimal rates for noisy low rank matrix completion Vladimir Koltchinskii, Alexandre B. Tsybakov, Karim Lounici

## Spectral clustering

Ng et al. (2002); von Luxburg (2007)

## Affinity propagation

Frey and Dueck (2007)

## Minimal spanning tree

- Werner Stuetzle (2003) Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample
- arXiv:1610.06599 Euclidean distance matrix completion and point configurations from the minimal spanning tree Adam Rahman, Wayne Oldford
- Li, Zhang Expected Lengths of Minimum Spanning Trees for Non-identical Edge Distributions, E. J. of probability Vol. 15 (2010), Paper no. 5, pages 110141. http://www.math.washington.edu/ ejpecp/

## References

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392.

Seeger, M. (2001). Learning with labeled and unlabeled data. Technical report.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.