Mathematisches Forschungsinstitut Oberwolfach

Report No. 22/2019

DOI: 10.4171/OWR/2019/22

# Statistical and Computational Aspects of Learning with Complex Structure

Organized by
Sara van de Geer, Zürich
Markus Reiß, Berlin
Philippe Rigollet, Boston

5 May – 11 May 2019

Abstract. The recent explosion of data that is routinely collected has led scientists to contemplate more and more sophisticated structural assumptions. Understanding how to harness and exploit such structure is key to improving the prediction accuracy of various statistical procedures. The ultimate goal of this line of research is to develop a set of tools that leverage underlying complex structures to pool information across observations and ultimately improve statistical accuracy as well as computational efficiency of the deployed methods. The workshop focused on recent developments in regression and matrix estimation under various complex constraints such as physical, computational, privacy, sparsity or robustness. Optimal-transport based techniques for geometric data analysis were also a main topic of the workshop.

*Mathematics Subject Classification (2010):* 62xx.

## Introduction by the Organizers

The workshop Statistical and Computational Aspects of Learning with Complex Structure, organized by Sara van de Geer (ETH Zürich), Markus Reiß (Humboldt Universität) and Philippe Rigollet (MIT) was held May 5th – May 11th, 2019. The aim of this workshop was to highlight recent achievements in modern statistical problems where more and more complex structure arise by bringing together statisticians, mathematicians and computer scientists that work on the cutting-edge of these questions. These goals were largely achieved.

The workshop was well attended by diverse pool of 51 participants (20% were women) with broad geographic representation from eight countries. The workshop featured twenty one hour-long presentations of noted excellent quality: the speakers made an effort to reach out to the diverse audience of the workshop which fostered sustained discussions between the participants. On Monday evening, PhD students offered short lightening talks to present their most recent achievements. The talks can be roughly clustered into the following topics, which the workshop was focused on.

*Complex signals in structured models:* Florentina Bunea discussed a regression model that accomodates a latent low-dimensional structure. John Duchi presented recent results in collaboration with Apple on the statistical limitations associated to private learning. Richard Nickl introduced a new model for the estimation of a signal known to evolve according to a non-abelian PDE. David Donoho derived the optimal spectral threshold for high-dimensional principal component analysis. Alexandre Tsybakov derived minimax rates for functional estimation in various sparse regression models. Elizaveta Levina presented a natural model for network data and an associated hierarchical clustering method with strong statistical guarantees. Alexandra Carpentier discussed support estimation in sparse linear regression.

*Computational aspects of structured learning:* Ankur Moitra discussed a computationally efficient estimation of Gaussian graphical models. Francis Bach and Andrea Montanari gave complementary presentations on the computational aspects of two-layer neural networks and the limitations of some existing simplifications of such neural networks. Soledad Villar presented a semidefinite relaxation for the classification-aware dimension reduction.

*Learning from heterogeneous datasets:* A noticeably growing theme in statistics is that of data integration whereby several datasets believed to contain similar information are combined to boost statistical efficiency. For this task, optimal transport has emerged as a powerful tool and several talks discussed mathematical and statistical aspects of optimal transport (Axel Munk, Jonathan Weed, Alexandra Suvorikova, Facundo Memoli and Jean-Michel Loubes). Other techniques for similar questions were also discussed by Genevera Allen and Peter Bühlmann.

*Inference from wild data:* Another theme featured in this workshop was the possibility of learning from data that violates commonly assumed assumptions such as independence. In this context, new techniques still allow to perform reliable inference. To achieve these goals, Emmanuel Candes and Rina Foygel-Barber presented new methods for conformal prediction under minimal assumptions and Alessandro Rinaldo discussed the inference from data collected by a bandit algorithm.

*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-1641185, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Merle Behr in the "Simons Visiting Professors" program at the MFO.

## Workshop: Statistical and Computational Aspects of Learning with Complex Structure

## Table of Contents

# Abstracts

## Learning GGMs without Condition Number Bounds

Ankur Moitra
(joint work with Jonathan Kelner, Frederic Koehler and Raghu Meka)

A Gaussian Graphical Model (GGM) in $n$ dimensions is a probability distribution with density

$$p(X = x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-(x-\mu)^T \Sigma^{-1}(x-\mu)/2\right)$$

where $\mu$ is the mean and $\Sigma$ is the covariance matrix. Let $\Theta = \Sigma^{-1}$, which is called the precision matrix. We can associate a graph to $\Theta$ which connects two nodes $i, j$ when $\Theta_{ij} \neq 0$. Now each node $i$ only interacts directly with its neighbors in the sense that $X_i$ is conditionally independent of every other node in the graphical model given its neighbors $(X_j)_{i \sim j}$. An important measure of complexity for a GGM is its sparsity $d$, which measures the largest number of non-zero off-diagonal entries in $\Theta$ in any row.

GGMs have wide-ranging applications in machine learning and the natural and social sciences where they are one of the most popular ways to model the statistical relationships between observed variables. For example, they are used to infer the structure of gene regulatory networks and to learn functional brain connectivity networks. In most of the settings in which they are applied, the number of observed samples is much smaller than the dimension. This means it is only possible to learn the GGM in a meaningful sense under some sort of sparsity assumption.

There is a vast literature on learning the sparsity pattern of $\Theta$ assuming some sort of lower bound on the strength of non-zero interactions. A popular approach is the Graphical Lasso [2] which solves the following convex program:

$$\max_{\Theta \succ 0} \log\det(\Theta) - \langle \widehat{\Sigma}, \Theta \rangle - \lambda \|\Theta\|_1$$

where $\widehat{\Sigma}$ is the empirical covariance matrix and $\|\Theta\|_1$ is the $\ell_1$ norm of the matrix as a vector. Ravikumar et al. [4] showed that under various incoherence assumptions that Graphical Lasso succeeds in recovering the sparsity pattern from $O((1/\alpha^2)d^2 \log(n))$ samples where $\alpha$ is an incoherence parameter (omitting the dependence on some additional terms, and assuming the non-zero entries are bounded away from 0 and the variances are $O(1)$). Another popular approach is the CLIME estimator [1] which solves the following linear program

$$\min_{\Theta} \|\Theta\|_1 \text{ s.t. } \|\widehat{\Sigma}\Theta - I\|_\infty \leq \lambda$$

The analysis of CLIME assumes a bound $M$ on the maximum $\ell_1$-norm of any row of the inverse covariance (given that the $X_i$'s are standardized to unit variance).

This is also a type of condition number assumption, although of a different nature than RE. It succeeds at structure recovery when given

$$m \geq CM^4 \log n$$

samples, again assuming the $\Theta_{ij}$ are either 0 or bounded away from 0.

While these works show that sparse GGMs can be estimated when the number of samples is polylogarithmic in the dimension, there is an important caveat in their guarantees. They need to assume that $\Theta$ is in some sense well-conditioned. However in the high-dimensional setting, this is a strong assumption which is violated by simple and natural models (e.g. a graphical model on a path), where these bounds turn out to be polynomial in the dimension. Furthermore, it is a fragile assumption that behaves poorly even under a seemingly benign operation like rescaling the variables.

We show that for some popular and widely-used classes of GGMs, it is possible to achieve both logarithmic sample complexity (the truly high-dimensional setting) and computational efficiency, even when $\Theta$ is ill-conditioned. First we study the class of attractive GGMs, in which the off-diagonal entries of $\Theta$ are non-positive. In terms of the correlation structure, this means that the variables are positively associated. A well-studied special case is the discrete Gaussian Free Field (GFF) where $\Theta$ is a principal submatrix of a graph Laplacian (i.e. we set some non-empty set of reference variables to zero as their boundary condition). This is a natural model because the Laplacian encourages "smoothness" with respect to the graph structure — if we think of the samples as random functions on the graph, then by integration by parts we see the log-likelihood of drawing a function is proportional to the $L^2$ norm of its discrete gradient. In the GFF setting, $\Theta$ will be ill-conditioned whenever some pair of vertices have large *effective resistance* between them (e.g., paths, rectangular grids, etc.,) as for example happens whenever there are nested sparse cuts which when collapsed lead to a long path resulting in variables having large (polynomial in $n$) variance.

We show that for attractive GGMs the conditional variance of some variable $X_i$ when we condition on a set $X_S$ is a monotonically decreasing and supermodular function of $S$. This fact was previously only observed in the GFF case. We give a new, short proof of this fact using a walk expansion, which can be derived using just basic linear algebra. Using this key result, we show the following:

**Theorem 1** (Informal)**.** [3] *Fix a $\kappa$-nondegenerate attractive GGM. There is an algorithm that runs in polynomial time and returns the true neighborhood of every node $i$ with high probability with $m \geq C(d+1/\kappa^2)d\log(1/\kappa)\log(n)$ samples, where $C$ is a universal constant.*

In fact our algorithm achieves the information-theoretically optimal sample complexity, up to constant factors as long as $d = O(1/\kappa^2)$ (a natural assumption, as the *average* degree is always $O(1/\kappa)$) and otherwise is close to optimal. In order to achieve this essentially optimal sample complexity, we need to carefully analyze the alignment between the true decrement of conditional variance in one step, $\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}})$, and the noisy empirical decrement

$\widehat{\mathrm{Var}}(X_i|X_S) - \widehat{\mathrm{Var}}(X_i|X_{S\cup\{j\}})$ without assuming too much accuracy on the estimates $\widehat{\mathrm{Var}}(X_i|X_S)$ themselves; the key insight here is to relate these decrements to the population risk of Ordinary Least Squares (OLS) and then use a suitable non-asymptotic risk bound. We also need to use an electrical argument, based on the SDD to Laplacian reduction and effective resistances, to bound the conditional variance after the first step of greedy, so that only a bounded number of iterations of greedy are required to learn a superset of the neighborhood.

While attractive GGMs are natural in some contexts, in others they are not. For example, in Genome Wide Association Schemes (GWASs) genes typically have inhibitory effects too. Walk-summable models are known to be a strict generalization, and to include other important cases like *pairwise normalizable* and *non-frustrated* models. A number of equivalent definitions are known for walk-summability — perhaps the easiest to work with is that making all off-diagonal entries negative preserves the fact that $\Theta$ is positive definite. We observe a key equivalence that, rather surprisingly, does not seem to be known in the literature: Walk-summable GGMs are exactly those that can be made SDD under an appropriate rescaling of coordinates. We prove this through elementary Perron-Frobenius theory.

Using the reduction from SDD to generalized Laplacians we are able to give algorithms for learning all, even ill-conditioned, walk-summable models (using that our greedy algorithms are naturally scale-invariant).

**Theorem 2** (Informal)**.** [3] *Fix a walk-summable, $\kappa$-nondegenerate GGM. There is an algorithm that runs in polynomial time and returns the true neighborhood of every node $i$ with high probability with $m \geq C(d^2/\kappa^4)\log(n)$ samples, where $C$ is a universal constant.*

We show examples of walk-summable GGMs where, unlike for attractive GGMs, the variance of $X_i$ conditioned on $X_S$ is not a supermodular function of $S$. Nevertheless, through some detailed calculations (and using properties of effective resistances) we are able to show that the greedy algorithm makes enough progress in each step that we quickly learn a superset of the neighborhood of each node, at which point we can do some post processing to find the true neighborhood, by iteratively trying out removing a variable and seeing if the conditional variance changes noticeably.

## References


[1] T. Cai, W. Liu and X. Luo. A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[2] J. Friedman, T. Hastie and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[3] J. Kelner, F. Koehler, R. Meka and A. Moitra. Learning Some Popular Gaussian Graphical Models without Condition Number Bounds. *ArXiv:1905.01282*, 2019.

[4] P. Ravikumar, M. Wainwright, G. Raskutti and B. Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

# Essential Regression

Florentina Bunea
(joint work with Xin Bing and Marten Wegkamp)

## 1. Introduction

We introduce the *Essential Regression* (E-Regression) model, as an alternative to the ubiquitous sparse high-dimensional linear regression on $p$ variables. It is a class of regression models tailored to applications where the relation between the dependent variable $Y$ and *representatives* of groups of components of the independent variables $X$, rather than between $Y$ and the components of $X$, is of main interest. A specific challenge addressed within the E-Regression framework is the definition of representatives in a mathematically coherent and practically interpretable way.

Formally, E-Regression is a new variant of the more classical factor regression model, introduced by [5], which postulates the existence of an unobserved, zero mean, random vector $Z \in \mathbb{R}^K$, for some *unknown* $K < p$, that is connected to the observed pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ via the model

$$Y = Z^T\beta + \varepsilon \tag{1}$$

$$X = AZ + W. \tag{2}$$

The dimension $K$, matrix $A \in \mathbb{R}^{p\times K}$ and vector $\beta \in \mathbb{R}^K$ are unknown, and $Z$, $\varepsilon$ and $W$ are independent. Furthermore, $\varepsilon$ and $W$ have zero mean, and unknown variance $\sigma^2$ and diagonal covariance matrix $\Gamma$, respectively. In contrast to sparse regression, where only few components of the observable $X$ are assumed to directly influence $Y$, our framework allows for all $p$-components of $X$ to influence $Y$, but mediated through the lower dimensional random vector $Z$. The mediator $Z$ is not observed, and made *interpretable* via a modeling assumption through which each component of $Z$ is given the physical meaning of a small group of the $X$-variables.

Factor regression models, and their many variants have been introduced to improve the prediction of $Y \in \mathbb{R}$ from $X \in \mathbb{R}^p$, when $p$ is very large and the components of $X$ are highly correlated. For this purpose, the matrix $A$ in (2) need only be unique up to generic invertible matrix transformations. This no longer suffices for the primary goal of this work, inference on the lower dimensional vector $\beta$, when two other aspects become important:

- $Z$ must be interpretable so that regression model (1) is interpretable;
- $\beta$ must be uniquely defined.

Both desiderata are met by placing the following assumptions on $A$ and the covariance matrix $\Sigma^z$ of $Z$, which differ from popular assumptions in the factor analysis literature.

**Assumption 1.**

(A0) $\|A_{j\cdot}\|_1 \le 1$ *for all* $j \in [p]$.

(A1) *For every* $k \in [K]$, *there exists at least two* $j \neq \ell \in [p]$, *such that* $|A_{jk}| = |A_{\ell k}| = 1$ *and* $A_{jk'} = A_{\ell k'} = 0$ *for any* $k' \neq k$.

(A2) *The covariance matrix* $\Sigma^z := Cov(Z)$ *is positive definite. There exists a constant* $\nu > 0$ *such that*

$$\min_{1\le a<b\le K} ( \Sigma^z_{aa} \wedge \Sigma^z_{bb} - |\Sigma^z_{ab}| ) > \nu.$$

Assumption 1, first introduced in [4], guarantees that $A$ and $\Sigma^z$ are identifiable, up to signed permutations. To the best of our knowledge, factor *regression* models under Assumption 1 have not been studied.

1.1. **Our contributions.** We begin by summarizing the model parameters, the nature of the data, as well as the relation between parameter dimensions and sample size. Throughout this work we assume that we have access to an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.

We denote by $I \subseteq \{1, \dots, p\}$ the index set of the pure $X$-variables. The following quantities are unknown and will be estimated from the data, under the Essential Regression model: $A$, $I$, $K$, $\beta$, $\sigma^2$, $\Sigma^w$, $\Sigma^z$. We allow for $p > n$, while $K < p$. In this work, we consider the case of non-sparse $\beta$, and $K < n$, but allow $K$ to grow with the sample size $n$. The complementary cases of $K > n$ and $\beta$ sparse will be studied in a follow-up work.

1.1.1. *Estimation and inference for* $\beta$. Under E-Regression, the coefficient $\beta$ satisfies

$$\beta = (\Sigma^z)^{-1}(A_I^T A_I)^{-1} A_I^T \mathrm{Cov}(X_I, Y), \tag{3}$$

where $A_I$ is the sub-matrix of $A$ with row indices corresponding to indices in the pure variable set $I$. We use the representation (3) and plug-in estimators of the unknown quantities to construct our proposed estimator $\widehat{\beta}$. We employ the LOVE algorithm developed in [4] to estimate $I$, its partition, and $K$.

To benchmark the quality of estimation of $\beta$, under the Essential Regression framework, we prove that the minimax optimal rate of estimating $\beta$ in the $\ell_2$-norm in $\mathbb{R}^K$ is $(1 \vee \|\beta\|/\sqrt{m})\sqrt{K/n}$ in our model with $K < n$. The quantity $m$ is the size of the smallest group of pure variables. We show that the proposed estimator $\widehat{\beta}$ is minimax rate optimal, up to logarithmic factors in $n$ and $p$. Moreover, $\widehat{\beta}$ is component-wise asymptotically normal. Its asymptotic variance agrees in order with the information bound in our Essential Regression model and can be consistently estimated. The analysis of $\widehat{\beta}$ relies on being able to consistently identify the pure variables. This is done by using the sample $X_1, \dots, X_n$ alone, *without* using $Y_1, \dots, Y_n$, and consequently, inference for $\beta$, at the coarser resolution level provided by the *essence* $Z$, is valid *uniformly* over $\beta$. This is in contrast with inference in direct sparse regression of $Y$ on $X$, after consistently estimating the support of $\beta$, which is valid only for regression coefficients above the minimax optimal $O(\sqrt{\log p/n})$ level, see for instance [1, 2, 3].

1.1.2. *Prediction of $Y$ from $X$ via Essential Regression.* In general factor regression (FR) models (1) – (2), at the population level, the best linear predictor of $Y$ takes the form

$$Y^*_{FR} = X^T A[\mathrm{Cov}(A^T X)]^{-1}\mathrm{Cov}(A^T X, Y). \tag{4}$$

We use the above expression, combined with a plug-in estimate of $A$, to construct in-sample predictors $\widehat{Y}$ of the observed data vector $Y \in \mathbb{R}^n$. The in-sample prediction risk bound is

$$\frac{1}{n}\mathbb{E}\left[\|\widehat{Y} - Z\beta\|^2\right] \leq C \times \frac{K}{n}\sigma^2 + \frac{\|\beta\|^2}{\Lambda_{\min}}\left\{1 + s_J\,\frac{\log(p \vee n)}{n}\right\}. \tag{5}$$

Analysis of this bound reveals improvements in the prediction risk that are possible under the E-Regression model, relative to model-free prediction via the principal components of $X$.

1.1.3. *Essential Regression as Regression with Clustered Predictors.* E-Regression can be used as a vehicle for model-based clustering and subsequent regression on cluster-related quantities. Within our E-Regression framework, we distinguish between two post-clustering problems: inference and prediction. We can interpret the matrix $A$ as a cluster allocation matrix and the inference carried out at the level of the latent factors $Z$, as inference carried out at the level of the cluster *centers*, but caution against replacing components of $Z$ by cluster averages. Indeed, we prove that replacing $Z$ by weighted averages $\bar{X}$ and subsequently regressing on $\bar{X}$, would *not* estimate $\beta$. However, this can be immediately corrected by regression on predictors $\widetilde{Z}$ of $Z$, obtained from appropriate cluster averages, exercising care when clusters overlap. With this correction, we obtain exactly the estimator of $\beta$ analyzed above, and we can interpret the newly developed inferential tools as tools for post-clustering inference in regression. Prediction of $Y$ requires less care as the cluster (weighted) averages $\bar{X}$ have the same prediction error as that relative to $\widetilde{Z}$. The resulting predictor corresponds to the one already explained above and our model formally justifies prediction from cluster averages. Moreover, prediction with clustered variables, whenever appropriate, provides an alternative to prediction via sparse regression in high dimensions, with differences particularly pronounced whenever the level of sparsity is not high and when the multi-collinearity among the $X$-variables is strong.

## References


[1] F. Bunea, *Honest variable selection in linear and logistic regression models via $\ell_1$ and $\ell_1$ + $\ell_2$ penalization*, Electronic Journal of Statistics **2** (2008), 1153–1194.

[2] P. Buhlmann and S. van de Geer, *Statistics for High-Dimensional Data*, Springer-Verlag (2011).

[3] C. Giraud, Introduction to High-Dimensional Statistics, Chapman Hall (2015).

[4] X. Bing, F. Bunea, Y. Ning and M. Wegkamp, *Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering*, The Annals of Statistics **to appear** (2019).

[5] J. Stock and M. Watson, *Forecasting Using Principal Components from a Large Number of Predictors*, Journal of the American Statistical Association, **97** (2002), 1167–1179.

## Integrated Principal Components Analysis

Genevera I. Allen
(joint work with Tiffany M. Tang)

Data integration, or the strategic analysis of multiple sources of data simultaneously, can often lead to discoveries that may be hidden in individualistic analyses of a single data source. We develop a new statistical data integration method named Integrated Principal Components Analysis (iPCA), which is a model-based generalization of PCA and serves as a practical tool to find and visualize common patterns that occur in multiple datasets. The key idea driving iPCA is the matrix-variate normal model, whose Kronecker product covariance structure captures both individual patterns within each dataset and joint patterns shared by multiple datasets:

$$\mathbf{X}_k \sim N(\boldsymbol{\mu_k}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta_k})$$

where for data source $\mathbf{X}_k$, $\boldsymbol{\Sigma}$ represents the common sample / row dependencies shared across all $k$ data sources and $\boldsymbol{\Delta}_k$ represents the separate column / feature dependencies unique to each data source. We then define the integrated principal components to be:

$$\mathbf{U} \leftarrow \text{eigenvectors}(\boldsymbol{\Sigma})$$
$$\mathbf{V}_k \leftarrow \text{eigenvectors}(\boldsymbol{\Delta}_k)$$

where $\mathbf{U}$ gives the shared iPCs and $\mathbf{V}_k$ gives the iPC loadings unique to each data set.

Building upon this model, we develop several penalized (sparse and non-sparse) covariance estimators for iPCA and study their theoretical properties. We show that our sparse iPCA estimator consistently estimates the underlying joint subspace, and using geodesic convexity, we prove that our non-sparse iPCA estimator converges to the global solution of a non-convex problem. We also demonstrate the practical advantages of iPCA through simulations and a case study application to integrative genomics for Alzheimer's Disease. In particular, we show that the joint patterns extracted via iPCA are highly predictive of a patient's cognition and Alzheimer's diagnosis.

### References

[1] Tiffany M. Tang and Genevera I. Allen. *Integrated Principal Components Analysis.* arXiv:1810.00832, 2018.

## On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport

Francis Bach
(joint work with Lénaïc Chizat)

Many tasks in machine learning and signal processing can be solved by minimizing a convex function of a measure. This includes sparse spikes deconvolution or training a neural network with a single hidden layer. For these problems, we study a simple minimization method [1]: the unknown measure is discretized into a mixture of particles and a continuous-time gradient descent is performed on their weights and positions. This is an idealization of the usual way to train neural networks with a large hidden layer. We show that, when initialized correctly and in the many-particle limit, this gradient flow, although non-convex, converges to global minimizers. The proof involves Wasserstein gradient flows, a by-product of optimal transport theory. Numerical experiments show that this asymptotic behavior is already at play for a reasonable number of particles, even in high dimension. While our results are qualitative, there are already works showing a quantitative analysis of the required number of neurons needed to reach the mean-field limit [3].

Moreover, in a series of recent theoretical works, it has been shown that strongly over-parameterized neural networks trained with gradient-based methods could converge linearly to zero training loss, with their parameters hardly varying. In this note, our goal is to exhibit the simple structure that is behind these results. In a simplified setting, we prove that "lazy training" essentially solves a kernel regression. We also show that this behavior is not so much due to over-parameterization than to a choice of scaling, often implicit, that allows to linearize the model around its initialization. These theoretical results complemented with simple numerical experiments make it seem unlikely that lazy training is behind the many successes of neural networks in high dimensional tasks [2].

### References

[1] L. Chizat, F. Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport.* Advances in Neural Information Processing Systems (NeurIPS), 2018.

[2] L. Chizat, F. Bach. *A Note on Lazy Training in Supervised Differentiable Programming.* Technical report, arXiv-1812.07956, 2018.

[3] S. Mei, A. Montanari, and P.-M. Nguyen. *A mean field view of the landscape of two-layers neural networks.* Technical Report 1804.06561, arXiv, 2018.

## Optimality in locally private estimation

John Duchi

I present new optimality results for estimation problems in local privacy models, when data is kept private even from the collector of data. I will give both fundamental limits - lower bounds - building out of communication complexity, allowing

the limits to apply to any level of desired privacy and any mode of data collection. I will also give commensurate optimality results, showing in particular applications to large scale estimation currently employed in a number of real-world scenarios.

*Note: abstract copied by the reporter from the book of abstracts at the MFO.*

## Consistent inversion of noisy non-Abelian $X$-ray transforms

Richard Nickl
(joint work with F. Monard, G.P. Paternain)

We discuss the results obtained in the recent preprint [3].

For $M$ a simple surface, the non-linear and non-convex statistical inverse problem of recovering a matrix field $\Phi : M \to \mathfrak{so}(n)$ from discrete, noisy measurements of the $SO(n)$-valued scattering data $C_\Phi$ of a solution of a matrix ODE is considered ($n \geq 2$). Injectivity of the map $\Phi \mapsto C_\Phi$ was established by [Paternain, Salo, Uhlmann; Geom. Funct. Anal. 2012, [4]].

A statistical algorithm for the solution of this inverse problem based on Gaussian process priors is proposed, and it is shown how it can be implemented by infinite-dimensional MCMC methods. It is further shown that as the number $N$ of measurements of point-evaluations of $C_\Phi$ increases, the statistical error in the recovery of $\Phi$ converges to zero in $L^2(M)$-distance at a rate that is algebraic in $1/N$, and approaches $1/\sqrt{N}$ for smooth matrix fields $\Phi$. The proof relies, among other things, on a new stability estimate for the inverse map $C_\Phi \to \Phi$.

Key applications of our results are discussed in the case $n = 3$ to *polarimetric neutron tomography*, see [Desai et al., Nature Sc. Rep. 2018, [1]] and [Hilger et al., Nature Comm. 2018, [2]].

### References

[1] N. Desai, W. Lionheart, M. Sales, S. Schmidt, S. Strobl et al., *Three Dimensional Polarimetric Neutron Tomography of Magnetic Fields*, Nature Scientific Reports (2018)

[2] A. Hilger, I. Manke, N. Kardiljov, M. Osernberg, H. Markotter, J. Banhart, *Tensorial neutron tomography of threedimensional magnetic vector fields in bulk materials.* Nature Communications (2018).

[3] F. Monard, R. Nickl, G.P. Paternain, *Consistent inversion of noisy non-Abelian X-ray transforms*, arXiv preprint, https://arxiv.org/abs/1905.00860.

[4] G. P. Paternain, M. Salo, G. Uhlmann, *The attenuated ray transform for connections and Higgs fields*, Geom. Funct. Anal. **22** (2012) 1460–1489.

## Central Limit Theorems for Wasserstein Distance between empirical distributions

Jean-Michel Loubes
(joint work with E. del Barrio)

Consider $P$, $Q$ probabilities on $\mathbb{R}^d$ and $c(x,y) = \|x-y\|^p$, $p \geq 1$.

$$\mathcal{W}_p^p(P,Q) = \min_{\pi \in \Pi(P,Q)} \int \|x-y\|^p d\pi(x,y)$$

$\Pi(P,Q)$ probabilities on $X \times Y$ with marginals $P$ and $Q$.
$\mathcal{W}_p$ is a metric on $\mathcal{F}_p$, probabilities on $\mathbb{R}^d$ with finite $p$-th moment.

Set $X_1, \ldots, X_n \in \mathbb{R}^d$, $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ Our aim is to provide a Central Limit Theorem for the Empirical transportation cost: $\mathcal{W}_p^p(P_n, Q)$ and extend it to $\mathcal{W}_p^p(P_n, Q_m)$.

For $d = 2$, (Ajtai-Komlos-Tusnady, 1984; Talagrand & Yukich, 1993)

$$c(p)\left(\frac{\log n}{n}\right)^{1/2} \leq E(\mathcal{W}_p(P_n, U([0,1]^2))) \leq C(p)\left(\frac{\log n}{n}\right)^{1/2}.$$

For $d \geq 3$, Talagrand, Yukich, 1992-1994

$$E(\mathcal{W}_p(P_n, U([0,1]^d))) \leq C(d,p)\frac{1}{n^{1/d}}.$$

Extensions to compactly supported $P$ with 'regular' density

If $d = 1$ and $P \sim f$ s.t. $\int_0^1 \left(\frac{(t(1-t))^{1/2}}{f(F^{-1}(t))}\right)^p dt < \infty$ (E. del Barrio, E. Giné and C. Matrán, 1999 and 2005)

$$\sqrt{n}\mathcal{W}_p(P_n, P) \to_w \left[\int_0^1 \left(\frac{B(t)}{f(F^{-1}(t))}\right)^p dt\right]^{1/p},$$

$B(t)$ Brownian bridge on $[0,1]$

No results for $P \neq Q$

CLT : $\left.\begin{array}{c} r_n\big(\mathcal{W}_p^p(P_n,Q) - a_n\big) \\ r_{n,m}\big(\mathcal{W}_p^p(P_n,Q_m) - a_{n,m}\big) \end{array}\right\} \Rightarrow$ Computation of approximate $p-values$

- $d = 1$, $p = 2$: A. Munk and C. Czado (1998) $\mathcal{W}_2$ for trimmed version
- $d \geq 1$, $p \geq 1$:
  - M. Sommerfeld and A. Munk (2018): $P, Q$ finitely supported
  - A. Tameling, M. Sommerfeld and A. Munk (2018): $P, Q$ countable support

In the Gaussian case results in Kroshnin, Spokoiny and Suvorikova (2019)

Here CLTs for $\mathcal{W}_2^2(P_n, Q)$ and $\mathcal{W}_2^2(P_n, Q_m)$ for general $P$, $Q$ and $d$
Assume $P$, $Q \in \mathcal{F}_2$ and

(1) $Q$ has a positive density in the interior of its convex support.

**Theorem** If $P, Q \in \mathcal{F}_{4+\delta}$ and satisfy (1), $\varphi_0$ o.t. potential from $P$ to $Q$ and $P_n$ empirical measure on $X_1, \ldots, X_n$, i.i.d. $P$ r.v.'s then

$$n\mathrm{Var}(\mathcal{W}_2^2(P_n, Q)) \to \sigma^2(P, Q)$$

with

$$\sigma^2(P,Q) = \int_{\mathbb{R}^d} (\|x\|^2 - 2\varphi_0(x))^2 dP(x) - \Big( \int_{\mathbb{R}^d} (\|x\|^2 - 2\varphi_0(x)) dP(x) \Big)^2$$

and

$$\sqrt{n}\big(\mathcal{W}_2^2(P_n, Q) - E\mathcal{W}_2^2(P_n, Q)\big) \underset{w}{\to} N(0, \sigma^2(P, Q))$$

Furthermore, if $Q_m$ empirical measure on $Y_1, \ldots, Y_m$ i.i.d. $Q$ r.v.'s, independent of the $X_i$'s, $n \to \infty$, $m \to \infty$ with $\frac{n}{n+m} \to \lambda \in (0,1)$, then

$$\tfrac{nm}{n+m}\mathrm{Var}(\mathcal{W}_2^2(P_n, Q_m)) \to (1-\lambda)\sigma^2(P,Q) + \lambda\sigma^2(Q,P)$$

$$\sqrt{\tfrac{nm}{n+m}}\big(\mathcal{W}_2^2(P_n, Q_m) - E\mathcal{W}_2^2(P_n, Q_m)\big) \underset{w}{\to} N(0, (1-\lambda)\sigma^2(P,Q) + \lambda\sigma^2(Q,P))$$

$$d = 1: \quad \mathcal{W}_p^p(F_n, G) = \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} |X_{(i)} - G^{-1}(t)|^p dt.$$

Other tools : strong approximation of quantile process

**Theorem**[Central Limit Theorem for $\mathcal{W}_p$ with $p > 1$] Assume that $F, G \in \mathcal{F}_{2p}$ and $G^{-1}$ is continuous on $(0,1)$ and $p > 1$. Then

(i) If $X_1, \ldots, X_n$ are i.i.d. $F$ and $F_n$ is the empirical d.f. based on the $X_i$'s

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - E\mathcal{W}_p^p(F_n, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) If, furthermore, $F^{-1}$ is continuous, $Y_1, \ldots, Y_m$ are i.i.d. $G$, independent of the $X_i$'s, $G_m$ is the empirical d.f. based on the $Y_j$'s and $\frac{n}{n+m} \to \lambda \in (0,1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - E\mathcal{W}_p^p(F_n, G_m)) \to_w N(0, (1-\lambda)\sigma_p^2(F,G) + \lambda\sigma_p^2(G,F)).$$

Assume $p \geq 2$. Under the assumptions of the CLT,

(i) if $F$ satisfies I) to IV) then $\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G))$.
(ii) if, furthermore, $G$ satisfies I) to IV) and $\frac{n}{n+m} \to \lambda \in (0,1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G)) \to_w N(0, (1-\lambda)\sigma_p^2(F,G) + \lambda\sigma_p^2(G,F)).$$

Results are taken from [1], [2], [3].

References

[1] E Del Barrio, P Gordaliza, H Lescornel, JM Loubes, Central limit theorem and bootstrap procedure for Wasserstein's variations with an application to structural relationships between distributions, Journal of Multivariate Analysis,169 (2019), 341-362

[2] E Del Barrio, JM Loubes, Central limit theorems for empirical transportation cost in general dimension The Annals of Probability 47 (2) (2019), 926–951.

[3] E del Barrio, P Gordaliza, JM Loubes A Central Limit Theorem for transportation cost with applications to Fairness Assessment in Machine Learning (2018) arXiv preprint arXiv:1807.06796

## Distributional Replicability*

Peter Bühlmann

(joint work with Dominik Rothenhäusler, Nicolai Meinshausen, Jonas Peters)

The common notion of replicability in statistics quantifies how well a finding from one data set generalizes to a new unseen data set having the same data-generating distribution as the original one. Typically, the quantification is in terms of statistical uncertainties. We consider here the problem when the new data set comes from a different distribution than the one generating the observed data. It is related to distributional robustness, and hence called "distributional replicability".

Distributional robustness deals with the problem of estimating and optimizing an unknown parameter (or function) $\theta$ with respect to a worst case risk over a class of probability distributions

$$\hat{\theta} = \mathrm{argmin}_\theta \sup_{P \in \mathbf{F}} \mathbb{E}_P[\ell(Z;\theta)], \tag{1}$$

where $\ell(.;.)$ is a loss function and $Z$ represents a random variable generating a data point. The choice of the class $\mathbf{F}$ is important as it encodes models for data-generating distributions of a new unseen data set. We focus here on regression where the class $\mathbf{F}$ is induced by causal-type (structural equation) models estimated from observed data. The latter is assumed to be heterogeneous with different observed environments or regimes which allows to construct $\mathbf{F}$ in a data-driven way. The optimization in (1) can be achieved by *causal regularization* in terms of the observed data distribution(s). The corresponding methodology is called *anchor regression* [1]. Anchor regression itself is motivated by *invariant causal prediction* [2], a conceptual framework and methodology which connects causality with distributional invariance (and hence) robustness properties.

*Anchor regression* provides a novel methodology for distributional (and predictive) robustness and replicability, with provable guarantees; and it leads to interesting results on a large-scale data set from proteomics.

* Dedicated to Sara van de Geer whose birthday has been celebrated during the workshop.

## References

[1] D. Rothenhäusler, N. Meinshausen, P. Bühlmann and J. Peters (2018). *Anchor regression: heterogeneous data meets causality*, Preprint arXiv:1801.06229 (2018).

[2] J. Peters, P. Bühlmann and N. Meinshausen *Causal inference using invariant prediction: identification and confidence intervals (with discussion)*, Journal of the Royal Statistical Society, Series B **78** (2016), 947–1012.


## Estimation of functionals in sparse vector model


Alexandre Tsybakov

(joint work with Laëtitia Comminges, Olivier Collier, Mohamed Ndaoud)


Assume that we have the observations $y_i = \theta_i + \varepsilon\xi_i,\ i = 1, \dots, d$, where $\theta = (\theta_1, \dots, \theta_d) \in \mathbf{R}^d$ is a vector of unknown parameters, $\varepsilon > 0$, and $\xi_i$ are independent identically distributed (i.i.d.) random variables. Assume also that $\theta$ belongs to the class $B_0(s)$ of all $s$-sparse vectors, that is, vectors in $\mathbf{R}^d$ with not more than $s$ non-zero components, $s \in \{1, \dots, d\}$. We first consider the problem of estimation of $\|\theta\|_\gamma = \left(\sum_{i=1}^d |\theta_i|^\gamma\right)^{1/\gamma}$, $\gamma > 0$, based on observations $y = (y_1, \dots, y_d)$. We prove that, if $\varepsilon > 0$ is known and $\xi_i$ are i.i.d. standard Gaussian variables, the minimax risk for estimation of $\|\theta\|_\gamma$ under the squared loss on the class $B_0(s)$ satisfies (cf. [1]):

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s)} \mathbf{E}_\theta[(\hat{T} - \|\theta\|_\gamma)^2/\varepsilon^2] \asymp \begin{cases} s^{2/\gamma}\log(1 + d/s^2), & \text{if } s \le \sqrt{d}, \\ \dfrac{s^{2/\gamma}}{\log(1 + s^2/d)}, & \text{if } s > \sqrt{d} \text{ and } \gamma \notin E, \\ d^{1/\gamma}, & \text{if } s > \sqrt{d} \text{ and } \gamma \in E, \end{cases}$$

where $E$ is the set of all even integers, and $\mathbf{E}_\theta$ denotes the expectation with respect to the distribution of $y$, and $\inf_{\hat{T}}$ is the infimum over all estimators. We also construct estimators achieving this minimax rate, see [1]. This generalizes the previous results of [2] (case $\gamma = 2$) and of [3] (case $\gamma = 1$ and $s = d^a, a > 1/2$).

Next, for the same sparse vector model, when the noise is not necessarily Gaussian and $\varepsilon$ is not necessarily known, we consider adaptive estimation of $\theta$, of the norm $\|\theta\|_2$ and of the noise variance $\varepsilon^2$. We construct adaptive estimators and establish the optimal rates when adaptation is considered with respect to the triplet "noise level - noise distribution - sparsity". We consider classes of noise distributions with polynomially and exponentially decreasing tails as well as the case of Gaussian noise. The obtained rates turn out to be different from the minimax non-adaptive rates when the triplet is known. A crucial issue is the ignorance of the noise variance. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise variance can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. Estimation of noise variance in our setting

can be viewed as an adaptive variant of robust estimation of scale in the contamination model, where instead of fixing the nominal distribution in advance, we assume that it belongs to some class of distributions.

## References

[1] L. Comminges, O. Collier, A.B. Tsybakov, *On estimation of nonsmooth functionals of sparse normal means*, arxiv1805.10791
[2] O. Collier, L. Comminges, A.B. Tsybakov, *Minimax estimation of linear and quadratic functionals on sparsity classes*, Annals of Statistics **45** (2017), 923–958.
[3] T.T. Cai, M.G. Low, *Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional*, Annals of Statistics **39** (2011), 1012–1041.
[4] L. Comminges, O. Collier, M. Ndaoud, A.B. Tsybakov, *Adaptive robust estimation in sparse vector model*, 1802.04230

## Hierarchical community detection by recursive partitioning

Elizaveta Levina
(joint work with Tianxi Li, Lihua Lei, Sharmodeep Bhattacharyya, Purnamrita Sarkar, Peter J. Bickel)

Network data have become increasingly common in many fields, with interesting scientific phenomena discovered through the analysis of biological, social, ecological, and various other networks. Among various network analysis tasks, community detection (the task of clustering network nodes into groups with similar connection patterns) has been one of the most studied, due to the ubiquity of communities in real-world networks and the appealing mathematical formulations that lend themselves to analysis. For the most part, community detection has been formulated as the problem of finding a single partition of the network into some "correct" number of communities. However, it is both well known in practice and supported by theory that nearly all the algorithms and models proposed for this type of community detection do not work well when the number of communities is large. We argue that for large networks, a hierarchy of communities is preferable to such a partition, since multiple partitions at different scales frequently make more sense in real networks, and the hierarchy can be scientifically meaningful, like an evolutionary tree. A hierarchical tree, with larger communities subdivided into smaller ones, offers a natural and very interpretable representation of community structure, and simplifies the problem of estimating the potentially large number of communities from the entire network. In addition, a hierarchy gives us much more information than any "flat" partition, by indicating how close communities are through their tree distance. Finally, recursive splitting is more computationally efficient, and, as we show, in some settings is more accurate. In particular, we show that even when the full community structure corresponding to the leaves of the tree is below the recovery threshold, we can still consistently recover the top levels of the tree as long as they are well separated, giving us partial but accurate information where a flat partition method would fail.

Many existing algorithms for hierarchical clustering can be modified to apply to networks. We adopt a simple top-down recursive partitioning algorithm, once popular in the clustering literature. It requires two tools that, in turn, can be chosen among many existing methods: an algorithm to partition a given network into two, and a stopping rule to decide whether there is more than one community in a given subnetwork. Given these two tools, the recursive (bi-)partitioning algorithm proceeds by starting with all nodes in one community, applying the stopping rule to decide whether a split is needed, applying the splitting algorithm to split into two communities if so, and continuing to apply this to every resulting subnetwork until the stopping rule indicates there are no further splits to make. This class of algorithms can be made model-free and tuning-free, and is computationally efficient, with the computational cost growing logarithmically in the number of communities rather than linearly, which is the case for most flat partition methods. We implement recursive partitioning by using regularized spectral clustering as the splitting rule, and the Bethe-Hessian estimator of the number of communities as the stopping rule, although any other consistent method can be used instead.

We analyze the algorithm's theoretical performance under a natural framework for this setting, the binary tree stochastic block model. Under this model, we prove that the algorithm correctly recovers the entire community tree under mild growth assumptions on the average degree, allowing for sparse networks. Further, the assumptions to recover each level of the tree, which we make explicit, get strictly stronger as we move down the tree, illuminating the regime where recursive partitioning can correctly recover mega-communities at the higher levels of the hierarchy even when it cannot recover every community at the bottom of the tree. We show that in practice recursive partitioning outperforms "flat" spectral clustering on multiple performance metrics when the number of communities is large, and illustrate the algorithm on a dataset of statistics papers, constructing a highly interpretable tree of statistics research communities.

### References

[1] T. Li, L. Lei, S. Bhattacharyya, P. Sarkar, P. J. Bickel, and E. Levina. *Hierarchical community detection by recursive partitioning*, arXiv:1810.01509, 2018.

## Conformalized Quantile Regression

Emmanuel Candes
(joint work with Evan Patterson, Yaniv Romano)

Conformal prediction is a technique for constructing prediction intervals that attain valid coverage in finite samples, without making distributional assumptions. Despite this appeal, existing conformal methods can be unnecessarily conservative because they form intervals of constant or weakly varying length across the input space. In this talk we propose a new method that is fully adaptive to heteroscedasticity. It combines conformal prediction with classical quantile regression,

inheriting the advantages of both. We establish a theoretical guarantee of valid coverage, supplemented by extensive experiments on popular regression datasets. We compare the efficiency of conformalized quantile regression to other conformal methods, showing that our method tends to produce shorter intervals.

Suppose we are given $n$ training samples $\{(X_i, Y_i)\}_{i=1}^n$ and we must now predict the unknown value of $Y_{n+1}$ at a test point $X_{n+1}$. We assume that all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are drawn exchangeably—for instance, they may be drawn i.i.d.—from an arbitrary joint distribution $P_{XY}$ over the feature vectors $X \in \mathbb{R}^p$ and response variables $Y \in \mathbb{R}$. We aim to construct a *marginal distribution-free prediction interval* $C(X_{n+1}) \subseteq \mathbb{R}$ that is likely to contain the unknown response $Y_{n+1}$. That is, given a desired miscoverage rate $\alpha$, we ask that

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$$

for any joint distribution $P_{XY}$ and any sample size $n$. The probability in this statement is marginal, being taken over all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$.

To accomplish this, we build on the method of conformal prediction [5, 6, 7, 1, 8, 9]. We first split at random the training data into two disjoint subsets, a proper training $\mathcal{I}_1$ set and a calibration set $\mathcal{I}_2$.

- We fit two quantile regressors $\hat{q}_{\alpha_{\text{lo}}}(\cdot)$ and $\hat{q}_{\alpha_{\text{hi}}}(\cdot)$ on the proper training set to obtain initial estimates of the lower and upper bounds of the prediction interval by applying tools from quantile regression. Quantile regression estimates a conditional quantile function $q_\alpha$ of $Y_{n+1}$ given $X_{n+1}{=}x$ by solving the optimization problem

  $$\hat{q}_\alpha(x) = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i, f(X_i)) + \mathcal{R}(f),$$

  where $f(x)$ is the quantile regression function and the loss function $\rho_\alpha$ is the "check function" or "pinball loss" [3, 4], defined by

  $$\rho_\alpha(y, \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0, \\ (1-\alpha)(\hat{y} - y) & \text{otherwise.} \end{cases}$$

  Above, $\mathcal{F}$ is a class of functions (either parametric or nonparametric) and $\mathcal{R}(\cdot)$ is a possible regularizer.
- Then, using the calibration set, we correct the prediction interval calculated above (we "conformalize" it). We introduce conformity score for each data point in the calibration set:

  $$E_i = \max\{\hat{q}_{\alpha_{\text{lo}}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i)\}, \quad i \in \mathcal{I}_2.$$

  These scores are signed distances to the boundaries: the score is negative if the calibration point lies within the empirical quantile range and positive if it lies outside.

Given a new input data $X_{n+1}$, we construct the prediction interval for $Y_{n+1}$ as

$$C(X_{n+1}) = \left[\hat{q}_{\alpha_{\text{lo}}}(X_{n+1}) - Q,\ \hat{q}_{\alpha_{\text{hi}}}(X_{n+1}) + Q\right],$$

where

$$Q := (1-\alpha)(1+1/|\mathcal{I}_2|)\text{-th empirical quantile of}\{E_i : i \in \mathcal{I}_2\}$$

conformalizes the plug-in prediction interval. We refer to this method as the split CQR algorithm.

Our main result is this:

**Theorem 1.** *If $(X_i, Y_i)$, $i = 1, \ldots, n+1$ are exchangeable, then the prediction interval $C(X_{n+1})$ constructed by the split CQR algorithm satisfies*

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha.$$

*Moreover, if the conformity scores $E_i$ are almost surely distinct, then the prediction interval is nearly perfectly calibrated:*

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

Our method differs from the standard method of conformal prediction [1, 9] in that we calibrate the prediction interval using conditional quantile regression, while the standard method uses only classical, conditional mean regression. The result is that our intervals are adaptive to heteroscedasticity whereas the standard intervals are not. In [2] we evaluate the statistical efficiency of our framework by comparing its miscoverage rate and average interval length with those of other methods. Based on extensive experiments across eleven datasets, we conclude that conformal quantile regression yields shorter intervals than the competing methods.

## References


[1] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[2] Y. Romano, E. Patterson, and E. J. Candès, *Conformalized quantile regression*, *arXiv:1905.03222*, 2019.

[3] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[4] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[5] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pages 444–453, 1999.

[6] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

[7] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. IntechOpen, 2008.

[8] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[9] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

## Predictive inference with the jackknife+

Rina Foygel Barber
(joint work with Emmanuel Candès, Aaditya Ramdas, Ryan Tibshirani)

Given a training data set $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, consider the problem of constructing a prediction interval at a new point $X_{n+1}$ that is likely to contain the new response value $Y_{n+1}$. Our aim is to construct prediction intervals that are *distribution-free*, meaning that we do not need to rely on any assumptions about the data distribution to ensure that the prediction intervals achieve the desired coverage probability. To introduce some notation, we would like to ensure that

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha$$

for some target error level $\alpha$. Implicitly, this probability is also taken with respect to the training data points, which are used in the construction of the data-dependent prediction interval $\widehat{C}_{n,\alpha}$. We will assume that the training and test data points $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are drawn i.i.d. from some distribution, but place no conditions on the distribution. (More generally, we can assume simply that these $n+1$ data points are exchangeable.)

Many popular prediction methods take the form

$$\widehat{C}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm (\text{margin of error}),$$

where $\widehat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is a regression function fitted to the training data,

$$\widehat{\mu} = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n)\big),$$

for some regression method $\mathcal{A}$ (such as linear regression or a neural net—we treat $\mathcal{A}$ as a black box and assume only that $\mathcal{A}$ is invariant to the ordering of the training data points). If the margin of error is calculated using the training data residuals, however, then the predictive intervals are likely to undercover. This is because many algorithms will overfit to training data, leading to training residuals that are quite low relative to the prediction error on a new unseen test point $(X_{n+1}, Y_{n+1})$. It is well known that data splitting can be used to address this problem—the regression function is fitted on one portion of the training data, while the margin of error is calculated using the remaining portion. Specifically, let $n = n_0 + n_1$, define $\widehat{\mu}_{\text{split}} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})\Big)$, define $R_i^{\text{split}} = |Y_i - \widehat{\mu}_{\text{split}}(X_i)|$ for the holdout set $i = n_0 + 1, \ldots, n$, and let

$$\widehat{C}^{\text{split}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}_{\text{split}}(X_{n+1}) \pm \Big(\text{the } \lceil (1-\alpha)(n_1+1) \rceil\text{-th smallest value of } R^{\text{split}}_{n_0+1}, \ldots, R^{\text{split}}_n\Big).$$

This method, known also as "split conformal prediction prediction" or "inductive conformal prediction", achieves distribution-free predictive coverage because the residuals on the holdout set, $R^{\text{split}}_{n_0+1}, \ldots, R^{\text{split}}_n$, are exchangeable with the test point residual $|Y_{n+1} - \widehat{\mu}_{\text{split}}(X_{n+1})|$ [1, 2, 3]. One potential problem with the data splitting approach, however, is that when the training size $n$ is not very large,

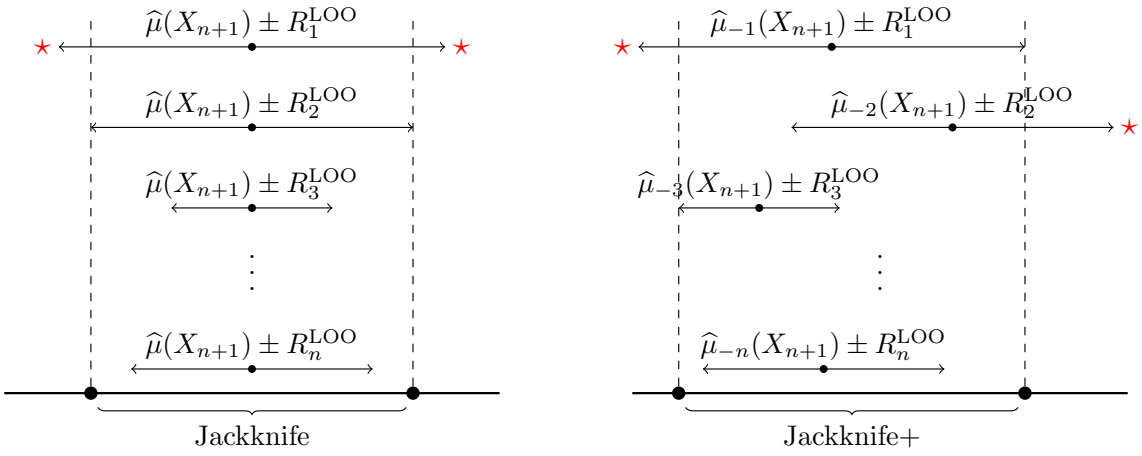


FIGURE 1. Illustration of the jackknife and jackknife+ methods.

fitting the regression function $\widehat{\mu}_{\mathrm{split}}$ on an even smaller sample size $n_0 < n$ can reduce accuracy (leading to wider prediction intervals). On the other hand, we do need to make $n_0$ substantially smaller than $n$, since an extremely small holdout set size $n_1 = n - n_0$ would lead to high variance in the margin of error.

To avoid this tradeoff we can instead consider the *jackknife* method (also known as leave-one-out cross-validation). Let $\widehat{\mu}_{-i} = \mathcal{A}\big((X_j, Y_j) : j \neq i\big)$, fitted on the training data with point $i$ removed, and let $R_i^{\mathrm{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ be the corresponding leave-one-out (LOO) residual. The jackknife interval is then given by

$$\widehat{C}_{n,\alpha}^{\mathrm{jackknife}}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \Big(\text{the } \lceil (1-\alpha)(n+1) \rceil\text{-th smallest value of } R_1^{\mathrm{LOO}}, \ldots, R_n^{\mathrm{LOO}}\Big).$$

Since point $i$ was not included when fitting the regression function $\widehat{\mu}_{-i}$, we have avoided overfitting and so in general expect to see good coverage. Unfortunately, however, coverage cannot be guaranteed without further assumptions. For example, $\mathcal{A}$ might perform better when the training data size is $n-1$ rather than $n$—while this type of behavior seems implausible, it is actually the case for "ridgeless" least squares (i.e., ridge regression with penalty parameter tending to zero), when $n < d$ [4]. On the other hand, if the algorithm is assumed to satisfy *stability* [5]—informally, this means that asymptotically we will have $\widehat{\mu}(X_{n+1}) \approx \widehat{\mu}_{-i}(X_{n+1})$, i.e., our predictions will not be sensitive to adding a single point to the training data—then this ensures asymptotically valid coverage for the jackknife [6].

In order to achieve predictive coverage without assumptions on the data distribution or the regression algorithm, we propose the jackknife+ method, which is defined as follows:

$$\widehat{C}_{n,\alpha}^{\mathrm{jackknife+}}(X_{n+1}) = \Big[\text{the } \lfloor \alpha(n+1) \rfloor\text{-th smallest value of } \widehat{\mu}_{-i}(X_{n+1}) - R^{\mathrm{LOO}_i}, \text{ the } \lceil (1-\alpha)(n+1) \rceil\text{-th smallest value of } \widehat{\mu}_{-i}(X_{n+1}) + R^{\mathrm{LOO}_i}\Big].$$

This construction differs from the usual jackknife by replacing the point estimate $\widehat{\mu}(X_{n+1})$ with the leave-one-out values $\widehat{\mu}_{-i}(X_{n+1})$. The difference between the two methods is illustrated in Figure 1. The proposed jackknife+ method is closely related to the cross-conformal method [7, 8].

Our main result proves that the jackknife+ method has, at most, a factor of 2 inflation in its error rate:

**Theorem 1.** *Let $\mathcal{A}$ be any regression method that is invariant to the ordering of the training data, and let $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$ for an arbitrary distribution $P$. Then the jackknife+ method satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha.$$

The factor of 2 cannot be removed—we verify this with an explicit matching bound:

**Theorem 2.** *For any dimension $d \geq 1$, there exists an algorithm $\mathcal{A}$ that is invariant to the ordering of the data, and a distribution $P$ on $\mathbb{R}^d \times \mathbb{R}$, such that for $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$,*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \leq 1 - 2\alpha + \mathcal{O}\left(\sqrt{\log(n)/n}\right).$$

Our paper [9] also includes generalizations of this procedure—first, to a $K$-fold cross-validation method (for jackknife+, we choose $K = n$, and use leave-one-out cross-validation), and second, to construct asymmetric intervals with signed residuals rather than using the absolute value residuals for a symmetric margin of error, with analogous theoretical results for these modifications as well.

## References


[1] Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks.* In Tools in artificial intelligence. InTech, 2008.

[2] Vladimir Vovk. *Conditional validity of inductive conformal predictors.* In Asian conference on machine learning, pages 475–490, 2012.

[3] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. *Distribution-free predictive inference for regression.* Journal of the American Statistical Association, 113(523):1094–1111, 2018.

[4] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. *Surprises in high-dimensional ridgeless least squares interpolation.* arXiv preprint arXiv:1903.08560, 2019.

[5] Olivier Bousquet and André Elisseeff. *Stability and generalization.* Journal of machine learning research, 2(Mar):499–526, 2002.

[6] Lukas Steinberger and Hannes Leeb. *Conditional predictive inference for high-dimensional stable algorithms.* arXiv preprint arXiv:1809.01412, 2018.

[7] Vladimir Vovk. *Cross-conformal predictors.* Annals of Mathematics and Artificial Intelligence, 74(1-2):9–28, 2015.

[8] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. *Cross-conformal predictive distributions.* In Conformal and Probabilistic Prediction and Applications, pages 37–51, 2018.

[9] Barber, Rina Foygel, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. *Predictive inference with the jackknife+.* arXiv preprint arXiv:1905.02928, 2019.

## Data analysis based on optimal transport: Theory, algorithms, applications

Axel Munk

(joint work with Marcel Klatt, Jörn Schrieber, Max Sommerfeld, Carla Tameling, Yoav Zemel)

The optimal transport distance (OTD) between two probability measures (see e.g., [11] or [17] for a comprehensive treatment) is a fundamental concept in mathematical sciences including probability and statistics, with respect to both theory and practice. The $p$-th OTD between two probability measures $\mu$ and $\nu$ on a Polish metric space $(\mathcal{X}, d)$ is given by

$$W_p(\mu,\nu) = \left(\inf \int_{\mathcal{X}\times\mathcal{X}} d^p(x,y) d\pi(x,y)\right)^{1/p} \tag{1}$$

for $p \in [1,\infty)$, the infimum is taken over all probability measures $\pi$ (couplings) on the product space $\mathcal{X}\times\mathcal{X}$ with marginals $\mu$ and $\nu$. Despite its long standing history in mathematics and related disciplines, such as physics and economics, statistical OTD based data analysis is a relatively new emerging field and challenged mainly by two issues:

(1) Its routine use in many real world applications as a measure to compare complex objects relies on fast computation of the empirical OTD (i.e. when the marginal measures are estimated from data by empirical counterparts $\mu_n, \nu_n$, denoted as EOTD).
(2) Methods for statistical inference are lacking to a large extent.

Addressing (2) there is a long standing interest in distributional limits for EOTD. However, most of this work is restricted to the univariate case $\mathcal{X} \subset \mathbb{R}$ (see e.g. [9, 5, 2]) A major reason of the limitation to dimension $D = 1$ is that in gerneral only for $\mathcal{X} \subset \mathbb{R}$ (or more generally a rooted tree) the coupling which solves (1) is known explicitly. For $\mathcal{X} \subset \mathbb{R}$ this can be expressed in terms of the quantile functions $F^{-1}$ and $G^{-1}$ of $\mu$ and $\nu$, respectively, as $\pi = (F^{-1} \times G^{-1})\#\mathcal{L}$, where $\mathcal{L}$ is the Lebesgue measure on $[0,1]$. For higher dimensions only in specific settings such a coupling can be computed explicitly and then can be used to derive limit laws as well (see e.g. [12]). Already for $D = 2$ it is well known that the scaling rate for the limiting distribution (if it exists) of $W_1(\hat{\mu}_n, \mu)$ when $\mu$ is the uniform measure on $\mathcal{X} = [0,1]^2$ must be of complicated nature as it is bounded from above and below by a rate of order $\sqrt{n\log(n)}$. Recently, [6] gave distributional limits for the quadratic EWD for the euclidean space in general dimension with a scaling rate $\sqrt{n}$. This yields a (non-degenerate) normal limit in the case $\mu \neq \nu$, i.e., when the data generating measure is different from the measure to be compared with (extending [9] to $D > 1$). Their result centers the EOTD with its expectation and requires $\mu$ and $\nu$ to have a positive Lebesgue density on the interior of their convex support. However, in the case $\mu = \nu$, their distributional limit degenerates to a point mass at 0, underlining the fundamental difficulty of this problem again.

An alternative approach has been advocated recently in [14, 16, 8] who restrict to finite / countable spaces $\mathcal{X} = \{x_1, \ldots, x_N\}$ ($N$ can be $\infty$). Such limit laws for the EOTD for $\mu = \nu$ then require a different scaling rate $n^{1/2p}$. The proof is based on the one hand on weak convergence of the underlying multinomial process associated with $\hat{\mu}_n$ with respect to a weighted $\ell^1$-norm (if $N = \infty$)

$$||x|| = \sum_{x \in \mathcal{X}} d^p(x, x_0)|r_x| + |r_{x_0}|, \tag{2}$$

and on the other hand to (infinite dimensional) sensitivity analysis of the corresponding linear program [3]. Here, $x_0 \in \mathcal{X}$ is fixed, but arbitrary. For the delta method to work here weak convergence in the weighted $\ell^1$-norm (2) of the underlying empirical process $\sqrt{n}(\mu_n - \mu)$ is required as the directional Hadamard differentiability is proven w.r.t. this norm. In turn the well known summability condition

$$\sum_{x \in \mathcal{X}} d^p(x, x_0)\sqrt{r_x} < \infty \tag{3}$$

is necessary and sufficient for weak convergence of the EOTD, which is known to be necessary and sufficient for the discrete empirical process $\sqrt{n}(\mu_n - \mu)$ to be Donsker according to the well known Borisov-Durst Theorem.

In summary, the discretized OT allows to obtain such limit laws in large generality and therefore offers a perspective to various statistical applications, such as OT based ANOVA or confidence statements for the OTD [14, 16].

We further argue that for many applications the transport plan itself is a quantity which can be very helpful for a meaningful data analysis. However, if it comes to asymptotic laws of the underlying transport plan, limit theorems are entirely lacking ($D = 1$ is an exception again) and only for the entropy regularized plan [4] asymptotic normality has been shown recently for finite number of support points [8]. This is based on a pertubation analysis of the regularized EOT and highlights an interesting link to a computational burden recently noticed [1, 7] when approximating the linear program underlying the original OT problem by regularized solutions. As the regularization parameter $\lambda$ vanishes, the computational effort scales quadratically. This coincides with the scaling rate we obtain for $p = 2$ [8].

Finally, we address (1) and discuss resampling schemes for the statistical computation of the EOTD. These can be combined with any state of the art back end solver [15] in a simple way. We show that subsampling allows to approximate the OTD with controllable statistical accuracy at a given computational cost which can be magnitudes smaller than for the original problem. In particular, when approximating a $d$ dimensional cube by an equidistant grid of size $N$ the expected $L^1$ loss of the EOTD is independent of the size $N$ of the underlying problem as long as the spatial dimension of the cube is $1 \leq D \leq 3$. When $D = 4$ the discretization size enters logarithmically, and for larger $d$ polynomially at a specific rate, see [15]. It is not known to us whether this is sharp, in particular for large dimensions.

## References


[1] Altschuler, J., Weed, J. and Rigollet, P., *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*, Advances in Neural Information Processing Systems (2017), 1961–1971.

[2] Bobkov, S. and Ledoux, M., *One-dimensional Empirical Measures, Order Statistics and Kantorovich Transport Distances*, Lecture Notes, (2014).

[3] Bonnans, J.F. and Shapiro, A., *Perturbation Aanalysis of Optimization Problems*, Springer (2000).

[4] Cuturi, M. and Peyré, G., *A smoothed dual approach for variational Wasserstein problems*, SIAM Journal on Imaging Sciences, **9** (2016), 320–343.

[5] del Barrio, E., Giné, E. and Matrán, C., *Central limit theorems for the Wasserstein distance between the empirical and the true distributions*, Annals of Probability, **27** (1999), 1009–1071.

[6] del Barrio, E., Loubes, J.M., *Central limit theorems for empirical transportation cost in general dimension*, The Annals of Probability **47** (2019), 926-951.

[7] Dvurechensky, P., Gasnikov, A. and Kroshnin, A., *Computational optimal transport: Complexity by accelerated gradient descent Is better than by Sinkhorn's algorithm*,Preprint arXiv:1802.04367 (2018).

[8] Klatt, M., Tameling, C., Munk, A., *Empirical regularized optimal transport: Statistical theory and applications*, arXiv:1810.09880 (2019).

[9] Munk, A. and Czado, C., *Nonparametric validation of similar distributions and assessment of goodness of fit*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **60** (1998), 223–241.

[10] Panaretos, V. M. and Zemel, Y., *Statistical aspects of Wasserstein distances*, Annual Review of Statistics and Its Application, **6** (2019), 4015-431.

[11] Rachev, S. T. and Rüschendorf, L., *Mass Transportation Problems: Volume I: Theory*, Springer Science & Business Media (1998).

[12] Rippl, T., Munk, A., Sturm, A., *Limit laws of the empirical Wasserstein distance: Gaussian distributions*, Journal of Multivariate Analysis, **151** (2016), 90-109.

[13] Schrieber, J., Schuhmacher, D. and Gottschlich, C., *DOTmark–a benchmark for discrete optimal transport*, IEEE Access, **5** (2017), 271–282.

[14] Sommerfeld, M. and Munk, A., *Inference for empirical Wasserstein distances on finite spaces*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **80** (2018), 219–238.

[15] Sommerfeld, M., Schrieber, J., Zemel, Y. and Munk, A., *Optimal transport: Fast probabilistic approximation with exact solvers*, Preprint arXiv:1802.05570 (2018).

[16] Tameling, C., Sommerfeld, M., Munk, A., *Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications*, arXiv:1707.00973 (2017), Annals of Applied Probability, to appear.

[17] Villani, C., *Optimal Transport: Old and New*, Springer Science & Business Media (2008).

## Wasserstein Projection Pursuit

Jonathan Weed
(joint work with Philippe Rigollet)

Given two probability measures $\mu$ and $\nu$ supported on $\mathbb{R}^d$, the Wasserstein distance $W_p(\mu, \nu)$ between them can be estimated by the empirical quantity $W_p(\hat{\mu}_n, \hat{\nu}_n)$. Standard results [3, 1] imply that if $d > 2p$, then this plug-in estimator satisfies

$$\begin{aligned}\mathbb{E}|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)| &\le \mathbb{E}W_p(\mu, \hat{\mu}_n) + \mathbb{E}W_p(\nu, \hat{\nu}_n) \\ &\le C_{d,p} n^{-1/d}\,,\end{aligned}$$

and this rate can be shown to be tight for certain $\mu$ and $\nu$. This estimator therefore suffers from the curse of dimensionality. However, this analysis does not preclude the possibility that given $n$ i.i.d. samples from $\mu$ and $\nu$, one can construct a *different* estimator $\hat{W}$ that achieves a significantly better rate of convergence. We therefore ask whether the $n^{-1/d}$ rate can be improved.

We define the minimax risk

$$R(n, \mathcal{P}) := \inf_{\hat{W}} \sup_{\mu,\nu \in \mathcal{P}} \mathbb{E}_{\mu,\nu} |\hat{W} - W_p(\mu, \nu)|\,,$$

where the infimum is taken over estimators $\hat{W}$ constructed from $n$ independent samples from $\mu$ and $n$ independent samples from $\nu$. We are aware of only one lower bound on $R(n, \mathcal{P})$ in the literature, due to [2], who prove that $R(n, \mathcal{P}) \gtrsim n^{-3/2d}$ when $\mathcal{P}$ is the set of measures supported on $[0, 1]^d$. Our main result is to sharpen this bound considerably.

**Theorem 1.** *Let $\mathcal{P}$ be the set of distributions supported on $[0, 1]^d$. If $d > 2p$, then*

$$R(n, \mathcal{P}) \ge C_{d,p}(n \log n)^{-1/d}\,.$$

This result immediately implies that the plug-in estimator is optimal up to logarithmic factors.

### 1. Proof technique

The core idea of our lower bound is to relate estimating the Wasserstein distance to the problem of estimating total variation distance, sharp rates for which are known [11, 4]. Lower bounds for the total variation estimation problem are generally obtained by constructing a pair of distributions $p$ and $q$ as well as a reference distribution $u$ on $[m]$ such that $\mathrm{tv}(p, u) \ge 2\varepsilon$ and $\mathrm{tv}(q, u) < \varepsilon$ but such that $n$ samples from $p$ and $q$ are indistinguishable. The existence of such a pair implies that no estimator can obtain accuracy better than $\varepsilon$.

In order to apply this technique to the Wasserstein distance, a first attempt is to embed $p$, $q$, and $u$ in $[0, 1]^d$ by letting them be supported on an equally spaced grid of size $[m]$. Given two distributions $\mu$ and $\nu$ on such a grid, it is easy to see that

$$m^{-1/d}\mathrm{tv}(\mu, \nu)^{1/p} \lesssim W_p(\mu, \nu) \lesssim \mathrm{tv}(\mu, \nu)^{1/p}\,.$$

Unfortunately, the lower and upper bounds in the above embedding differ by a factor of $m^{1/d}$, and neither inequality can be sharpened in general. As a result, such an embedding is too coarse to be useful in constructing lower bounds. However, we show that this approach can be salvaged when one of the distributions is the uniform measure on $[m]$ as long as the $\chi^2$-divergence between the two distributions is not large.

**Proposition 1.** *Assume $d > 2p \geq 2$, and let $m$ be a positive integer. Let $u$ be the uniform distribution on $[m]$. There exists a random function $F : [m] \to X$ such that for any distribution $q$ on $[m]$,*

$$cm^{-1/d}\mathrm{tv}(q,u)^{\frac{1}{p}} \leq W_p(F_\sharp q, F_\sharp u) \leq C_{d,p}m^{-1/d}(\chi^2(q,u))^{1/d}\mathrm{tv}(q,u)^{\frac{1}{p}-\frac{2}{d}}$$

*with probability at least .9.*

We then show a modified lower bound for a testing problem involving the total variation distance over the class of distributions on $[m]$ close to the uniform measure in $\chi^2$ divergence, inspired by a strategy of [10] and [12]. Combining these results yields the lower bound.

## 2. Spiked Transport Model

We propose a model analogous to the spiked covariance model, where we plant low-dimensional structure in an otherwise high-dimensional model. Let us fix a subspace $\mathcal{U} \subseteq \mathbb{R}^d$ of dimension $k \ll d$. Let $X^{(1)}$ and $X^{(2)}$ be two random variables supported on $\mathcal{U}$ with arbitrary distribution, and let $Z$ be a third random variable, independent of $X^{(1)}$ and $X^{(2)}$, such that $Z$ is supported on $\mathcal{U}^\perp$, the orthogonal complement of $V$. We let $\mu^{(1)}$ and $\mu^{(2)}$ be the law of $X^{(1)} + Z$ and $X^{(2)} + Z$, respectivey. Though $\mu^{(1)}$ and $\mu^{(2)}$ are high-dimensional distributions, they differ only on the low-dimensional subspace $\mathcal{U}$. We call this the *spiked transport model*.

This model suggests the following estimator. Given any probability distribution $\mu$ on $\mathbb{R}^d$ and a $k \times d$ matrix $U$ with orthonormal rows, if $Y \sim \mu$, we write $\mu_U$ for the distribution of $UY$. Given samples from $\mu^{(1)}$ and $\mu^{(2)}$, we define

$$\hat{W} := \sup_{U \in \mathcal{V}_k} W_p(\hat{\mu}_U^{(1)}, \hat{\mu}_U^{(2)}) \,. \tag{1}$$

where the maximization is taken over the *Stiefel manifold* $\mathcal{V}_k$ of $k \times d$ matrices with orthonormal rows. We call this procedure *Wasserstein projection pursuit*. This procedure has also been considered by [8] and [5].

We show that under the spiked transport model, our proposed estimator performs well.

**Theorem 2.** *Let $p \in [1,2]$. Under the spiked transport model, if $\mu^{(1)}$ and $\mu^{(2)}$ satisfy $T_p(\sigma^2)$, then the estimator $\hat{W}$ defined in* (1) *satisfies*

$$\mathbb{E}|\hat{W}_p - W_p(\mu^{(1)}, \mu^{(2)})| \lesssim \sigma \cdot \left(n^{-1/k} + \sqrt{\frac{d\log n}{n}}\right).$$

Strikingly the rate $n^{-1/d}$ achieved by the naïve estimator has been replaced by $n^{-1/k}$—in other words, this estimator enjoys the rate typical for $k$-dimensional rather than $d$-dimensional measures.

Our assumption is that the measures in question satisfy a *transport inequality*: $\mu$ satisfies $T_p(\sigma^2)$ if

$$W_p(\nu, \mu) \leq \sqrt{2\sigma^2 D(\nu\|\mu)} \quad \forall \nu \in \mathcal{P}(\mathbb{R}^d)\,.$$

Since the pioneering work of [6, 7] and [9], transport inequalities have played a central role in the analysis of the concentration properties of high-dimensional measures.

## References


[1] S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(4):1183–1203, 2013.

[2] K. Do Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory Comput. Syst.*, 48(2):428–442, 2011.

[3] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

[4] J. Jiao, Y. Han, and T. Weissman. Minimax estimation of the $L_1$ distance. *IEEE Trans. Inform. Theory*, 64(10):6672–6706, 2018.

[5] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde. Generalized sliced Wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019.

[6] K. Marton. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

[7] K. Marton. A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.*, 6(3):556–571, 1996.

[8] F.-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. *arXiv preprint arXiv:1901.08949*, 2019.

[9] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, 6(3):587–600, 1996.

[10] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:183, 2010.

[11] G. Valiant and P. Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 403–412. IEEE Computer Soc., Los Alamitos, CA, 2011.

[12] Y. Wu and P. Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.*, 47(2):857–883, 2019.


## The high-dimensional behavior of linearized neural networks


Andrea Montanari

(joint work with Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz)


Consider the classical statistical learning problem, whereby we are given independent and identically distributed (i.i.d.) pairs $(y_i, \boldsymbol{x}_i)$, $i \leq n$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ a feature vector and $y_i \in \mathbb{R}$ a label or response variable. We assume a simple model in which feature vectors are uniformly distributed over the sphere with radius $\sqrt{d}$ in $\mathbb{R}^d$, $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and labels aregiven by $y_i = f_*(\boldsymbol{x}_i)$, for some unknown function $f_* \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$. We would like to construct a function $f$ which allows us

to predict future responses. The quality of a predictor $f$ is measured via its square prediction error (risk): $\mathbb{E}\{(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\}$.

Current practice supports the use of neural networks, the simplest example being two-layers neural networks:

$$\text{(NN)} \qquad \mathcal{F}_{\mathsf{NN}} \equiv \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \ : \quad a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \ \forall i \le N \Big\}.$$

Here $N$ is the number of neurons, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function. While it is well understood that $\mathcal{F}_{\mathsf{NN}}$ can approximate a broad class of functions $f_*$, it is unclear what subset of these functions can be learnt using practical algorithms, and in particular using stochastic gradient descent.

Over the last several years, considerable attention has been devoted to two classes of models that can be regarded as linearizations of two-layers networks. The first class is the random features model of Rahimi and Recht [6], which only optimizes over the weights $a_i$'s, while keeping the first layer fixed:

$$\text{(RF)} \qquad \mathcal{F}_{\mathsf{RF}}(\boldsymbol{W}) \equiv \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \le N \Big\}.$$

Here $\boldsymbol{W} \in \mathbb{R}^{n\times d}$ is a matrix whose $i$-th row is the vector $\boldsymbol{w}_i$. In the RF model, this is chosen randomly, and independent of the data.

The second model is the neural tangent kernel of Jacot, Gabriel and Hongler [5], which we define as

$$\text{(NT)} \qquad \mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \equiv \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \ : \quad \boldsymbol{a}_i \in \mathbb{R}^d \ \forall i \le N \Big\}.$$

Again, $\boldsymbol{W}$ is a matrix of weights that is not optimized over, but instead drawn at random. Further $\sigma'$ is the (weak) derivative of the activation function with respect to its argument.

Both $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ and $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$ are proper subsets of $\mathcal{F}_{\mathsf{NN}}$, that are tractable: optimizing the square loss over these classes reduces to a least squares problem. We define the minimum population risk over these classes by

$$(1) \qquad R_{\mathsf{M}}(f_*, \boldsymbol{W}) = \inf_{f\in\mathcal{F}_{\mathsf{M}}(\boldsymbol{W})} \mathbb{E}\big[(f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2\big], \quad \mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}\}.$$

Notice that this is a random variable because of the random features encoded in the matrix $\boldsymbol{W} \in \mathbb{R}^{n\times d}$. We assume $(\boldsymbol{w}_i)_{i\le N} \sim_{iid} \mathsf{Unif}(S^{d-1})$. For $\ell \in \mathbb{N}$, we denote by $\mathsf{P}_{\le \ell} : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d})$ the orthogonal projector onto the subspace of polynomials of degree at most $\ell$. (We also let $\mathsf{P}_{>\ell} = \boldsymbol{I} - \mathsf{P}_{\le \ell}$.) In other words, $\mathsf{P}_{\le \ell} f$ is the function obtained by linear regression of $f$ onto monomials of degree at most $\ell$.

The tradeoff between prediction error and the number of hidden units $N$ in the model RF has been been studied in the past, see in particular [2, 1, 3, 7]. In the paper recent paper [4] we obtain a sharp characterization holding in the high-dimensional regime $N, d \to \infty$, both for the RF and the NT model.

For RF, assuming $N \leq d^{\ell+1+\delta}$ for an integer $\ell$ and any $\delta > 0$, we have

$$R_{\mathsf{RF}}(f_*, \boldsymbol{W}) = R_{\mathsf{RF}}(\mathsf{P}_{\leq \ell} f_*, \boldsymbol{W}) + \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_{\mathbb{P}}(\|f_*\|_{L^2}^2)\,.$$

This result holds under minimal conditions on the activation function $\sigma$.

For NT, assuming $N \leq d^{\ell+1+\delta}$ for an integer $\ell$ and any $\delta > 0$, we have

$$R_{\mathsf{NT}}(f_*, \boldsymbol{W}) = R_{\mathsf{NT}}(\mathsf{P}_{\leq \ell+1} f_*, \boldsymbol{W}) + \|\mathsf{P}_{>\ell+1} f_*\|_{L^2}^2 + o_{\mathbb{P}}(\|f_*\|_{L^2}^2)\,.$$

This result holds under a technical condition on the Hermite coefficient of $\sigma$, which can be checked to hold for activation functions of common use.

## References


[1] Ahmed El Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[2] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

[3] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[4] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv:1904.12191*, 2019.

[5] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[6] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[7] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.


## On Central limit theorem for Bures-Wasserstein barycenters and beyond


Alexandra Suvorikova

(joint work with A.Kroshnin, V.Spokoiny)


Space of finite-dimensional Hermitian operators $\mathbb{H}(d)$ provides a powerful toolbox for data representation. For instance, in quantum mechanics it is used for mathematical description of physical properties of a quantum system, also known as observables. A subspace $\mathbb{S}(d) \subset \mathbb{H}(d)$ of real-valued symmetric matrices is also of great interest: points in $\mathbb{S}(d)$ are widely used for description of systems in engineering applications, medical studies, neural sciences, evolutionary biology e.t.c. Usually such data sets are considered to be randomly sampled from an unknown distribution $\mathbb{P}$ [1, 2, 3], and statistical characteristics of $\mathbb{P}$ such as, in particular, mean and variance, appear to be of interest for further planning of an experiment and analysis of obtained results, used for further development of natural science models. The talk is mainly based on the paper [5]. There we investigate in more details the concept of Bures-Wasserstein barycenter $Q_*$, that is essentially a Fréchet mean of some distribution $\mathbb{P}$ supported on a subspace of positive semi-definite Hermitian operators $\mathbb{H}_+(d)$ endowed with Bures-Wasserstein distance introduced

in a seminal paper [4]. Namely, for any pair of positive semi-definite matrices $Q, S \in \mathbb{H}_+(d)$ it is written as:

$$d^2_{BW}(Q,S) = \mathrm{tr}Q + \mathrm{tr}S - 2\mathrm{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2}.$$

We allow a barycenter to be constrained to some affine subspace of $\mathbb{H}_+(d)$, $\mathbb{A}$, and provide conditions ensuring its existence and uniqueness. Given some $\mathbb{P}$, $\mathrm{supp}(\mathbb{P}) \subseteq \mathbb{H}_+(d)$, its Bures-Waserstein barycenter is written as

$$Q_* \in \underset{Q\in\mathbb{H}_+(d)\cap\mathbb{A}}{\mathrm{argmin}} \int\limits_{\mathrm{supp}(\mathbb{P})} d^2_{BW}(Q,S)d\mathbb{P}(S).$$

Moreover, given an i.i.d. set of matrices $S_1, ..., S_n$ sampled from $\mathbb{P}$, we can construct its empirical counterpart $Q_n$

$$Q_n \in \underset{Q\in\mathbb{H}_+(d)\cap\mathbb{A}}{\mathrm{argmin}} \frac{1}{n}\sum_i d^2_{BW}(Q,S_i).$$

In the first part of the talk we investigate convergence and concentration properties of an empirical counterpart of $Q_*$ in both Frobenius norm and Bures-Wasserstein distance, and explain, how obtained results are connected to optimal transportation theory and can be applied to statistical inference in quantum mechanics. The second part of the talk is based on a working paper "Geometry of multiplier bootstrap in the space of Hermitian matrices". There we explain how the framework of classical resampling techniques (see e.g. [6]) can be extended to the case of Bures-Wasserstein space, and introduce some geometrical intuition behind the idea. The work develops an idea presented in [7].

## References


[1] Goodnight, Charles J., and James M. Schwartz. *A bootstrap comparison of genetic covariance matrices*, Biometrics (1997): 1026–1039.

[2] Calsbeek, Brittny, and Charles J. Goodnight. *Empirical comparison of G matrix test statistics: finding biologically relevant change*, Evolution: International Journal of Organic Evolution **63.10** (2009): 2627–2635.

[3] Gonzalez, Oscar, et al. *Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA* Multiscale Modeling & Simulation **15.3** (2017): 1073–1107.

[4] Bhatia, Rajendra, Tanvi Jain, and Yongdo Lim. *On the Bures-Wasserstein distance between positive definite matrices*, Expositiones Mathematicae (2018).

[5] Kroshnin, Alexey, Vladimir Spokoiny, and Alexandra Suvorikova. *Statistical inference for Bures-Wasserstein barycenters*, arXiv preprint arXiv:1901.00226 (2019).

[6] Mammen, Enno. *Bootstrap and wild bootstrap for high dimensional linear models*, The annals of statistics **21.1** (1993): 255-285.

[7] Ebert, Johannes, Vladimir Spokoiny, and Alexandra Suvorikova. *Construction of non-asymptotic confidence sets in 2-Wasserstein space*, arXiv preprint arXiv:1703.03658 (2017).

## The Gromov-Wasserstein distance and distributional invariants of datasets

Facundo Mémoli

In many applications datasets can be regarded as metric measure spaces (mm-spaces for short): triples $\mathcal{X} = (X, d_X, \mu_X)$ where $(X, d_X)$ is a compact metric space and $\mu_X$ is a fully supported Borel probability measure on $X$.

Two mm-spaces $\mathcal{X}$ and $\mathcal{Y}$ are said to be isomorphic if there exists an isometry $\varphi : X \to Y$ such that $\varphi_{\#}\mu_X = \mu_Y$.

Let $\mathfrak{M}$ denote the collection of all mm-spaces. One possible metric structure on $\mathfrak{M}$ is the Gromov-Wasserstein distance [3]: for each $p \geq 1$,

$$d_{GW,p}(\mathcal{X}, \mathcal{Y}) := \frac{1}{2} \inf_{\mu} \mathrm{dis}_p(\mu)$$

where $\mu$ ranges over all couplings $\mu$ between $\mu_X$ and $\mu_Y$ and for each coupling $\mu$,

$$\mathrm{dis}_p(\mu) := \left( \iint_{X\times Y\times X\times Y} |d_X(x,x') - d_Y(y,y')|^p \, \mu(dx \times dy)\, \mu(dx' \times dy') \right)^{1/p}$$

is called the $p$-distortion of $\mu$. This definition can be extended to the case $p = +\infty$.

**Remark 1.** *Note that when $X$ and $Y$ are finite, the above minimization leads to a quadratic functional on the linearly constrained variable $\mu$.*

**Theorem 1** ([3])**.** *For each $p \in [1, +\infty]$ $d_{GW,p}$ defines a metric $\mathfrak{M}$ modulo isomorphism.*

There is a distinguished object in $\mathfrak{M}$: the one point mm-space $*$; it's underlying set is the one point set $\{*\}$, the metric is (0), and the reference probability measure is $\delta_*$. Notice that given any other $\mathcal{X} \in \mathfrak{M}$ there exists exactly one measure coupling between $\mu_X$ and $\delta_*$: the product measure $\mu_X \otimes \delta_*$. The $p$-distortion of this coupling is therefore

$$\left( \iint_{X\times X} (d_X(x,x'))^p \, \mu(dx)\, \mu_X(dx') \right)^{1/p}$$

which only depends on $\mathcal{X}$. This quantity is clearly an isomorphism invariant of $\mathcal{X}$, is called the $p$-diameter of $\mathcal{X}$, and is denoted as $\mathrm{diam}_p(\mathcal{X})$. Thus, $d_{GW,p}(\mathcal{X}, *) = \frac{1}{2}\mathrm{diam}_p(\mathcal{X})$, and, since by Theorem 1 $d_{GW,p}$ satisfies the triangle inequality, we find that as a map $\mathfrak{M} \to \mathbb{R}_+$, $\mathrm{diam}_p$ is 2-Lipschitz:[1]

$$2 \cdot d_{GW,p}(\mathcal{X}, \mathcal{Y}) \geq |\mathrm{diam}_p(\mathcal{X}) - \mathrm{diam}_p(\mathcal{Y})|.$$

There are other more interesting invariants of mm-spaces. One would expect that the collection $\{\mathrm{diam}_p(\mathcal{X})\}_{p\in[0,\infty]}$ of $p$-diameters of $\mathcal{X}$ coincides with the moments of some distributional invariant associated to $\mathcal{X}$. This motivates the definition of the so called global distribution of distances of $\mathcal{X}$:

$$dH_{\mathcal{X}} := (d_X)_{\#}\, \mu_X \otimes \mu_X,$$

[1] One also more informally says that the $p$-diameter invariant is stable.

a probability measure on the real line. The global distribution of distances is also stable:

**Proposition 2** ([3]). *For all $\mathcal{X}, \mathcal{Y} \in \mathfrak{M}$,*

$$2 \cdot d_{GW,p}(\mathcal{X}, \mathcal{Y}) \geq d_{W,p}^{\mathbb{R}_+}(dH_{\mathcal{X}}, dH_{\mathcal{Y}}).$$

*(Above, the right hand side is the p-Wasserstein distance between the respective global distance distributions.)*

Global distance distributions have been used extensively in applications (see [4] and references therein). One question that arises is whether these invariants are injective, perhaps inside some suitably restricted class of mm-spaces. It is known that without further qualification the answer is negative [3, Section 5]. There are even finite subsets $X$ and $Y$ of the real line, with the same cardinality such that when endowed with uniform measure and with the euclidean distance they have the same global distribution of distances. One example follows from a construction described in [1]: Let $X = \{0, 1, 4, 10, 12, 17\}$ and $Y = \{0, 1, 8, 11, 13, 17\}$. Then, one can check that

$$dH_{\mathcal{X}} = dH_{\mathcal{Y}} = \frac{3}{18}\delta_0 + \frac{1}{18}\sum_{a \in A} \delta_a$$

where $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17\}$.

Recent work has attempted to clarify whether one may be able to achieve injectivity by restricting $dH_\bullet$ to suitable (but interesting) subclasses $\mathfrak{C} \subset \mathfrak{M}$, or even a notion of local injectivity: given a class $\mathfrak{C}$ identify the mm-spaces $\{\mathcal{X}_\alpha \in A\} \subset \mathfrak{C}$ such that if some other $\mathcal{Y} \in \mathfrak{C}$ satisfies $dH_{\mathcal{Y}} = dH_{\mathcal{X}_\alpha}$ for some $\alpha \in A$, then $\mathcal{Y}$ must be isomorphic to $\mathcal{X}_\alpha$.

For example, in [2] is is proved that if $\mathfrak{C}$ denotes the collection of all smooth simple plane curves $C$ (endowed with euclidean distance and normalized arc length measure), then $dH_C = dH_{\mathbb{S}^1}$ if and only if $C$ is isometric to the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$.

However, the following interesting phenomenon was also observed in [2]:

**Proposition 3.** *For arbitrary $\epsilon > 0$, there exist two curves $C_\epsilon$ and $C'_\epsilon$ in $\mathfrak{C}$ such that:*

1. *both $C_\epsilon$ and $C'_\epsilon$ are $\epsilon$-close to $\mathbb{S}^1$ in say the Gromov-Hausdorff sense,*
2. *$dH_{C_\epsilon} = dH_{C'_\epsilon}$, but*
3. *$C_\epsilon$ is not isomorphic to $C'_\epsilon$.*

## References


[1] G. S. Bloom, *A counterexample to a theorem of S. Piccard*, J. Comb. Theory, Ser. A **22(3)**, 378–379 (1977).

[2] F. Mémoli and T. Needham, *Gromov-Monge quasi-metrics and distance distributions*, arXiv preprint arXiv:1810.09646, (2018).

[3] F. Mémoli, *Gromov-Wasserstein distances and the metric approach to object matching*, Foundations of computational mathematics **11.4** (2011): 417-487.

[4] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, *Shape distributions*, ACM Trans. Graph. **21(4)**, 807–832 (2002).

# Label aware dimensionality reduction with applications to genetic marker selection

Soledad Villar
(joint work with B. Dumitrascu, C. McWirther, D. Mixon, B. Engelhardt)

## 1. Label-aware dimensionality reduction by convex programming

Structure-preserving dimensionality reduction techniques are central to data science and had been long-studied. They take many forms, from Johnson-Lindenstrauss projections to manifold learning. In this work we focus on a dimensionality reduction technique that preserves the classification structure of the data.

**Problem.** Projection factor recovery
Let $\Pi$ denote the orthogonal projection onto some unknown subspace $T \subseteq \mathbb{R}^d$ of some unknown dimension. What conditions on $f\colon T \to [k] := \{1, \dots, k\}$ and $\mathcal{X} \subseteq \mathbb{R}^d$ enable exact or approximate recovery of $\Pi$ from data of the form $\{(x, f(\Pi x))\}_{x\in\mathcal{X}}$?

In words, assuming the classification function factors through some unknown orthogonal projection operator $\Pi$ (i.e. labels $y = f(\Pi x)$), the objective is to reconstruct $\Pi$. Consider a sequence of labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ in $\mathbb{R}^d \times [k]$ and denote $\mathcal{Z}(\mathcal{D}) := \{x_i - x_j : i, j \in \mathcal{I}, y_i \neq y_j\}$. The following program finds the best orthogonal projection for factor recovery purposes:

$$(1) \quad \text{minimize} \quad \operatorname{rank}\Pi \ \text{ subject to } \ \|\Pi z\| \geq \Delta \ \forall z \in \mathcal{Z}(\mathcal{D}), \ \ \Pi^\top = \Pi, \ \ \Pi^2 = \Pi$$

Here, $\Pi$ is the decision variable, whereas $\Delta > 0$ is a parameter that prescribes a desired minimum distance between projected points $\Pi x_i$ and $\Pi x_j$ with differing labels. This parameter reflects a fundamental tension: $\Delta$ should be large so as to enable classification, but also $\operatorname{rank}\Pi$ should be small so that we can reduce the dimension. Since it is not clear how to tractably implement (1), together with collaborator Dustin Mixon and Culver McWhirter we propose a convex relaxation referred to as **SqueezeFit**:

$$(2) \quad sqz(\mathcal{D}, \Delta): \ \text{minimize } \operatorname{tr} M \ \text{ subject to } \ z^\top M z \geq \Delta^2 \ \forall z \in \mathcal{Z}(\mathcal{D}), \ 0 \preceq M \preceq I$$

If $\mathcal{Z}(\mathcal{D})$ is finite, then $sqz(\mathcal{D}, \Delta)$ is a semidefinite program, otherwise $sqz(\mathcal{D}, \Delta)$ is a semi-infinite program. Figure 1 illustrates how SqueezeFit is well suited for projection factor recovery.

When formulating SqueezeFit, we took inspiration from the metric learning literature.

### 1.1. Theoretical analysis of SqueezeFit.

In [6] provide theoretical guarantees for SqueezeFit in the context of projection factor recovery. The analysis considers geometric features of SqueezeFit and derives conditions that allow for recovering the projection factor successfully. The geometric properties of SqueezeFit are used to characterize when the SqueezeFit semi-infinite program satisfies strong duality. A sensitivity analysis of its dual certificate provides theoretical guarantees for exact recovery under certain models of random data where a true projection $\Pi$

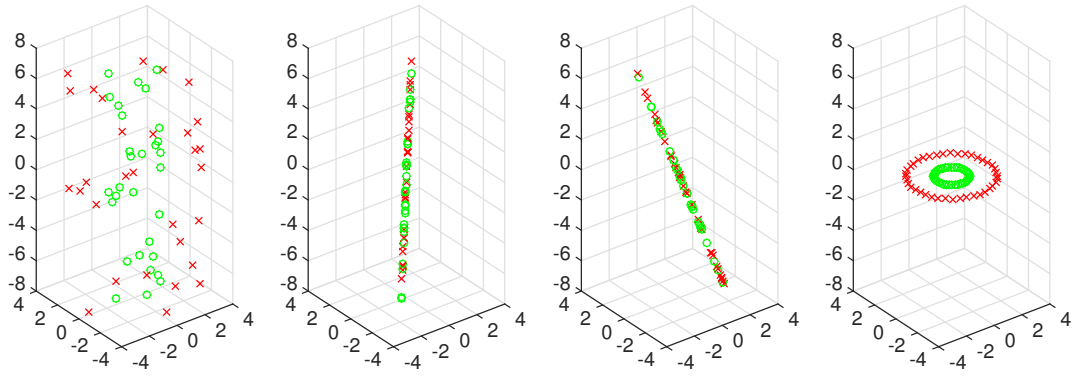

FIGURE 1. **(far left)** Plot of 60 data points in $\mathbb{R}^3$, half in one class, half in another. These points were drawn according to a random model with an unknown planted projection factor $\Pi$. **(middle left)** Principal component analysis (PCA) suggests one-dimensional structure in the data. Projecting onto this subspace (which was identified without regard for the points' classes) results in an undesirable mixture of the classes. **(middle right)** Unlike PCA, linear discriminant analysis (LDA) actually considers which class each point belongs to. Since there are two classes, the result is projection onto a 1-dimensional subspace, obtained by applying the classes' inverse covariance matrix to the difference of class centroids. Unfortunately, the result is again an unhelpful mixture of classes. **(far right)** Unlike PCA and LDA, SqueezeFit finds a low-rank projection that maintains some amount of distance between points from different classes. The resulting projection is a close approximation to the planted projection factor.

is planted and noise is added. The SNR thresholds found are optimal up to logarithmic factors.

1.2. **Scientific relevance: marker selection.** Recent technological developments in genetics and molecular biology have generated a wealth of data allowing researchers to measure and quantify RNA levels of individual cells. Compared to traditional bulk RNA sequencing (RNA-seq) in which information from thousands of cells is averaged, single-cell RNA sequencing (scRNA-seq) studies yield invaluable insights regarding cell type. Such information is critical to understanding complex human diseases and to understanding cell trajectories underlying cell development [9].

Data from single-cell RNA sequencing presents itself as matrices of size $d \times n$, where $n$, the number of cells queried, ever increasing, can now span hundreds of thousands, and where $d$, the number of genes, building blocks in the genetic alphabet, is close to $40,000$. Analyzing scRNA-seq itself is a very active field of research which has spun many dimensionality reduction methods exist for analyzing and clustering. However, the cell type information alone can not cast light into the

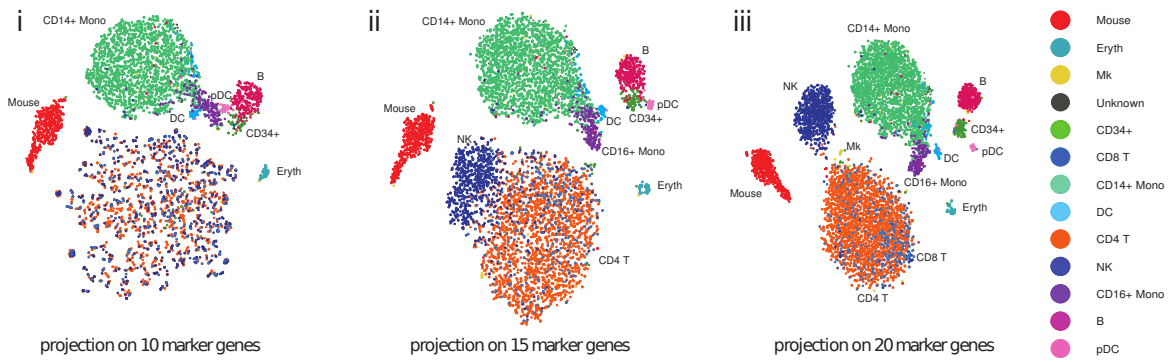


FIGURE 2. t-SNE visualization of results from single-cell expression profiles of cord blood mononuclear cells (CBMC, [8]) given a partition of labels. SqueezeFit-LP reveals that 15 marker genes are sufficient to distinguish 13 distinct cell populations.

spatial information underlying how the cell types are organized within tissues, or within tumor cells. To address this issue, imaging methods have been recently developed to visualize cells at unprecedented resolutions in a spatial setting [3, 1, 5]. These methods rely on a technique called single-molecule fluorescence in situ hybridization (smFISH), in which fluorescent probes bind near genes of interest called markers [7]. When the genes bound by probes are expressed, fluorescence can be detected using microscopy. Visualizing diverse and easily distinguishable sets of cells using this technique is often challenged by choosing the probes for the appropriate gene markers. State of the art methods can usually employ a set of at most 20 markers [2], and choosing these among the total of $40,000$ genes is a combinatorially challenging problem.

Together with Bianca Dumitrascu, Dustin Mixon and Barbara Engelhardt [4] we propose a linear programming version of SqueezeFit towards the problem of identifying genetic markers that separates labeled data. For instance, in order to identify the markers that separate a test group from a control group, one can write a projection factor recovery problem where the decision variable $M$ is diagonal. The SqueezeFit relaxation greatly simplifies to a linear program and can be implemented efficiently. See Figure 2 for a visualization of our results.

## REFERENCES

[1] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[2] S. Codeluppi, L. E. Borm, A. Zeisel, G. La Manno, J. A. van Lunteren, C. I. Svensson, and S. Linnarsson. Spatial organization of the somatosensory cortex revealed by cyclic smfish. *bioRxiv*, page 276097, 2018.

[3] N. Crosetto, M. Bienko, and A. Van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 16(1):57, 2015.

[4] B. Dumitrascu, S. Villar, D. G. Mixon, and B. E. Engelhardt. Optimal gene selection for cell type discrimination in single cell analyses. *BioRxiv*, page 599654, 2019.

[5] E. Lein, L. E. Borm, and S. Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, 2017.
[6] C. McWhirter, D. G. Mixon, and S. Villar. Squeezefit: label-aware dimensionality reduction. *in preparation*.
[7] A. Raj, P. Van Den Bogaard, S. A. Rifkin, A. Van Oudenaarden, and S. Tyagi. Imaging individual mrna molecules using multiple singly labeled probes. *Nature methods*, 5(10):877, 2008.
[8] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865, 2017.
[9] L. Zhu, J. Lei, B. Devlin, K. Roeder, et al. A unified statistical framework for single cell and bulk rna sequencing data. *The Annals of Applied Statistics*, 12(1):609–632, 2018.

## On the bias and risk of sample means in multi-armed bandits

Alessandro Rinaldo
(joint work with Jaehyeok Shin and Aaditya Ramdas)

Mean estimation is one of the most fundamental problems in statistics. In the classic non-adaptive setting, the target of estimation is the true mean $\mu$, assuming it exists, of a fixed distribution that is chosen in advance. In this case, if a fixed number of observations are sampled in an i.i.d. manner, then the sample mean is arguably the most natural choice for an estimator due to its favorable properties. In particular, it is unbiased, consistent, and converges almost surely to $\mu$. Under tail assumptions such as sub-Gaussian or sub-exponential conditions, the sample mean is tightly concentrated around $\mu$. Lastly, the sample mean has minimax optimal risk with respect to suitable loss functions such as the Kullback-Leibler (KL) loss for distributions in a natural exponential family.

However, in many cases the data are collected and analyzed in an adaptive manner, a prototypical example being the stochastic multi-armed bandits (MAB) framework. During the data collecting stage, in each round an analyst can draw a sample from one among a finite set of available distributions (arms) based on the previously observed data (*adaptive sampling*). The data collecting procedure can also be terminated based on a data-driven stopping rule rather than at a fixed time (*adaptive stopping*). Further, the analyst can choose a specific target distribution based on the collected data (*adaptive choice*), for example choosing to focus on the arm with the largest empirical mean at the stopping time. Lastly, in hindsight, the analyst may wonder what the bias of the sample mean of the chosen arm was at some past time (*adaptive rewinding*).

We decouple these different sources of selection bias: adaptive sampling of arms, adaptive stopping of the experiment, and adaptively choosingw hich arm to study. Through a new notion called "optimism" that captures certain natural monotonic behaviors of algorithms, we provide sufficient conditions for adaptive procedures to lead to biased sample means, with the bias taking either sign, depending on whether their combined effect is "monotone". To complement the qualitative characterization of the sign of the bias, we derive sharp bounds on teh risk (expected

Bregman divergence between the sample and true mean) under adaptive sampling and stopping. Next, we derive sharp risk bounds for sample means in the fully adaptive setting that includes an adaptive arm choice and adaptive rewinding. These bounds hold for a large class of underlying distributions, including all univariate exponential families, sub-Gaussian and sub-exponential distributions. Finally, we specify sufficient conditions for the consistency of a sequence of sample means in the fully adaptive setting

## Sparsity testing in the linear regression model

Alexandra Carpentier
(joint work with Nicolas Verzelen)

We consider the problem of sparsity testing in the high-dimensional linear regression model, as done in the Gaussian vector model in [2].

Let us write the random design high-dimensional linear regression model

$$Y = \mathbf{X}\theta^* + \sigma\epsilon \ , \tag{1}$$

where the unknown parameter $\theta^*$ belongs to $\mathbb{R}^p$, the noise vector $\epsilon \in \mathbb{R}^n$ follows a standard normal distribution and where the rows of $\mathbf{X}$ are i.i.d. sampled according to the normal distribution $\mathcal{N}(0, \mathbf{S}igma)$. In the sequel, $\mathbb{P}_{\theta^*,\mathbf{S}igma,\sigma}$ stands for the distribution of $(Y, \mathbf{X})$ in (1).

Given a non-negative integer $k_0 \in [0, p]$, write $\mathbb{B}_0[k_0] = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k_0\}$ for the set of $k_0$-sparse vectors $\theta$. Rephrasing our aim, we want to test whether $\theta^*$ belongs to $\mathbb{B}_0[k_0]$.

We aim at characterizing the smallest distance $\rho$, such that some tests achieve a small type I error probability and reject the null with high probability whenever $\min_{u\in\mathbb{B}_0[k_0])} \|\theta^* - u\|_2$ is larger than $\rho\sigma$, under the assumption that $\theta^* \in \mathbb{B}_0[k_0 + \Delta]$ for a fixed integer $\Delta > 0$. The following tables provide the rates in two different settings. Table 1 displays the rate in the case where $\mathbf{\Sigma} = I_p$ and $\sigma = 1$. Table 2 displays the rate in the case where $\mathbf{\Sigma}$ has eigenvalues upper and lower bounded by a constant and $\sigma$ is unknown.

Along the way, we also build a variable selection method based on iteratively projected square-root Lasso. This variable selection scheme is of independent interest. Up to our knowledge, it is the first provably polynomial time scheme that correctly selects the non-zero entries of $\theta^*$ whenever all of them are large compared to $\sigma\sqrt{\frac{\log(p)\log(n)}{n}}$ (see [4, 1] for a discussion), this uniformly over the class of covariance matrices with bounded eigenvalues upper and lower bounded by a constant. We then provide another test based on this construction, which is running in polynomial time (depending on $p, n$).

TABLE 1. Square minimax separation distances in the case where $\mathbf{\Sigma} = \mathbf{I}_p$ and $\sigma = 1$, when $p \geq n^{1+\zeta}$ with a fixed $\zeta > 0$.

| $k_0$ | $\Delta$ | $\rho^{*2}_\gamma[k_0, \Delta]$ |
|---|---|---|
| $k_0 \leq p^{1/2-\zeta}$ | $1 \leq \Delta \leq k_0 + \frac{\sqrt{n}}{\log(p)}$ | $\frac{\Delta \log(p)}{n}$ |
| | $k_0 + \frac{\sqrt{n}}{\log(p)} \leq \Delta \leq p - k_0$ | $\frac{1}{\sqrt{n}} + \frac{k_0 \log(p)}{n}$ |
| $p^{1/2+\zeta} \leq k_0 \leq c_\gamma \frac{n}{\log(p)}$ | $1 \leq \Delta \leq k_0 p^{-\zeta}$ | $\frac{\Delta \log(p)}{n}$ |
| | $k_0 \leq \Delta \leq p - k_0$ | $\frac{k_0}{n \log(p)}$ |

TABLE 2. Square minimax separation distances in the case where $\mathbf{\Sigma}$ has eigenvalues upper and lower bounded by a constant and $\sigma$ is unknown. We report in this table only the case where $n^{1+\zeta} \leq p \leq n^{2-\zeta}$, where $\zeta \in (0, 1)$ can be chosen arbitrarily small. LB stands for Lower bound and UB stands for upper bound.

| $k_0$ | $\Delta$ | $\boldsymbol{\rho}^{*2}_{g,\gamma}[k_0, \Delta]$ |
|---|---|---|
| $k_0 \leq p^{1/2-\zeta}$ | $1 \leq \Delta \leq p^{1/2-\zeta} \wedge k_0$ | $\frac{\Delta \log(p)}{n}$ |
| | $p^{1/2+\zeta} \wedge k_0 \leq \Delta \leq p - k_0$ | $\frac{\sqrt{p}}{n}$ |
| $p^{1/2+\zeta} \leq k_0 \leq c_\gamma \frac{n}{\log(p)}$ | $1 \leq \Delta \leq k_0 p^{-\zeta}$ | $\frac{\Delta \log(p)}{n}$ |
| | $k_0 \leq \Delta \leq p - k_0$ | LB : $\frac{k_0}{n \log(p)}$ |
| | | UB : $\frac{k_0 \log(p)}{n}$ |

## REFERENCES


[1] Ery Arias-Castro and Karim Lounici, *Estimation and variable selection with exponential weights*, *Electronic Journal of Statistics*, 8(1):328–354, 2014.

[2] Alexandra Carpentier and Nicolas Verzelen, *Adaptive estimation of the sparsity in the Gaussian vector model*, *Annals of Statistics, Volume 47, Number 1, 93-126*, 2019.

[3] Alexandra Carpentier and Nicolas Verzelen, *Adaptive estimation of the sparsity in the Gaussian vector model*, *arXiv:1901.08802*, 2019.

[4] Long Feng and Cun-Hui Zhang, *Sorted Concave Penalized Regression*, *arXiv:1712.09941*, 2017.


*Reporter: Francesco Ortelli*

# Participants

**Prof. Dr. Genevera Allen**
Department of Statistics
Rice University / MS-138
6100 Main Street
Houston, TX 77005
UNITED STATES

**Dr. Randolf Altmeyer**
Institut für Mathematik
Humboldt Universität zu Berlin
Unter den Linden 6
10117 Berlin
GERMANY

**Prof. Dr. Francis Bach**
INRIA - SIERRA project team
Departement d'Informatique de'l
Ecole Normale Superieure
Voie DQ12
2, rue Simone Iff
75012 Paris Cedex
FRANCE

**Prof. Dr. Merle Behr**
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

**Franz Besold**
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10117 Berlin
GERMANY

**Dr. Natalia Bochkina**
School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
King's Buildings
Peter Guthrie Tait Road
Edinburgh EH9 3FD
UNITED KINGDOM

**Prof. Dr. Peter Bühlmann**
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

**Prof. Dr. Florentina Bunea**
Department of Statistical Science
Cornell University
Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

**Prof. Dr. Cristina Butucea**
CREST - ENSAE
5, Avenue Henry Le Chatelier
91120 Palaiseau Cedex
FRANCE

**Prof. Dr. Emmanuel J. Candès**
Department of Statistics
Stanford University
Sequoia Hall
Stanford CA 94305-4065
UNITED STATES

**Prof. Dr. Alexandra Carpentier**
Fakultät für Mathematik
Otto-von-Guericke-Universität
Magdeburg
Postfach 4120
39016 Magdeburg
GERMANY

**Domagoj Ćevid**
Departement Mathematik
ETH-Zentrum
Rämistrasse 101
8092 Zürich
SWITZERLAND

**Yuansi Chen**
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

**Prof. Dr. Arnak Dalalyan**
ENSAE / CREST
École Nationale de la Statistique et de l'Administration Économique
5, Avenue Henry Le Châtelier
91120 Palaiseau Cedex
FRANCE

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

**Prof. Dr. David L. Donoho**
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

**Prof. Dr. John Duchi**
Department of Statistics and Electrical Engineering
Stanford University
Sequoia Hall
Stanford CA 94305-4065
UNITED STATES

**Prof. Dr. Rina Foygel Barber**
Department of Statistics
The University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637-1514
UNITED STATES

**Prof. Dr. Laszlo Györfi**
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Stoczek u. 2
1521 Budapest
HUNGARY

**Dr. Olga Klopp**
ESSEC Business School
CS 50105 Cergy
3, Avenue Bernard Hirsch
95021 Cergy-Pontoise / Cedex
FRANCE

**Prof. Elizaveta Levina**
Department of Statistics
University of Michigan
323 West Hall
1085 S. University Avenue
Ann Arbor MI 48109-1107
UNITED STATES

**Prof. Dr. Po-Ling Loh**
Department of Statistics
Wisconsin Institute for Discovery
University of Wisconsin, Madison
1300 University Avenue
Madison,WI 53715
UNITED STATES

**Prof. Dr. Jean-Michel Loubes**
Laboratoire de Statistique et Probabilités
Université Paul Sabatier
118, route de Narbonne
31062 Toulouse Cedex 4
FRANCE

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

**Dr. Facundo Mémoli**
Department of Mathematics
The Ohio State University
100 Mathematics Building
231 West 18th Avenue
Columbus, OH 43210-1174
UNITED STATES

**Prof. Dr. Ankur Moitra**
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
UNITED STATES

**Prof. Andrea Montanari**
Department of Electrical Engineering
and Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

**Dr. Nicole Mücke**
Institut für Stochastik und
Anwendungen
Universität Stuttgart, Raum 8.554
Pfaffenwaldring 57
70569 Stuttgart
GERMANY

**Prof. Dr. Sayan Mukherjee**
Department of Statistical Science
Duke University
112 Old Chemistry Building
P.O. Box 90251
Durham NC 27710
UNITED STATES

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

**Prof. Dr. Richard Nickl**
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

**Prof. Dr. Robert D. Nowak**
Department of Electrical and
Computer Engineering
University of Wisconsin-Madison
Engineering Hall # 3627
1550 Engineering Drive
Madison WI 53706
UNITED STATES

**Francesco Ortelli**
Departement Mathematik
ETH-Zentrum
Rämistrasse 101
8092 Zürich
SWITZERLAND

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

**Prof. Dr. Philippe Rigollet**
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge MA 02139-4307
UNITED STATES

**Prof. Dr. Alessandro Rinaldo**
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

**Prof. Dr. Angelika Rohde**
Fakultät für Mathematik
Albert-Ludwigs-Universität Freiburg
LST für Stochastik
Ernst-Zermelo-Strasse 1
79104 Freiburg i. Br.
GERMANY

**Prof. Dr. Johannes Schmidt-Hieber**
Department of Applied Mathematics
University of Twente
P.O.Box 217
7500 AE Enschede
NETHERLANDS

**Dr. Karin Schnass**
Institut für Mathematik
Universität Innsbruck
Technikerstrasse 13
6020 Innsbruck
AUSTRIA

**Prof. Dr. Vladimir G. Spokoiny**
Weierstrass-Institute for Applied Analysis
and Stochastics (WIAS)
Mohrenstrasse 39
10117 Berlin
GERMANY

**Bernhard Stankewitz**
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

**Dr. Alexandra Suvorikova**
Institut für Mathematik
Universität Potsdam
Postfach 601553
14415 Potsdam
GERMANY

**Dr. Yan Shuo Tan**
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

**Prof. Dr. Mathias Trabs**
Department Mathematik
Universität Hamburg
Bundesstrasse 55
20146 Hamburg
GERMANY

**Prof. Dr. Alexandre B. Tsybakov**
CREST, ENSAE
5, Avenue Henry Le Châtelier
91120 Palaiseau Cedex
FRANCE

**Prof. Dr. Sara van de Geer**
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

**Dr. Soledad Villar**
NYU Center for Data Science
Office 621
60 5th Avenue
New York, NY 10011
UNITED STATES

**Dr. Martin Wahl**
Institut für Mathematik
Humboldt Universität zu Berlin
Unter den Linden 6
10117 Berlin
GERMANY

**Dr. Jonathan Weed**
Department of Mathematics
Massachusetts Institute of Technology
77, Massachusetts Avenue
Cambridge, MA 02139-4307
UNITED STATES

**Sven Wang**
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

**Prof. Dr. Ji Zhu**
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
UNITED STATES