Universidad de los Andes Colombia | Johns Hopkins University | Google | Pensadores del Mundo, Fondo Internacional Uniandes | Columbia University in the City of New York | Wisconsin, University of Wisconsin-Madison | NSF | NSB

WORKSHOP

# STATISTICS & DATA SCIENCE

DECEMBER 9 - 12. 2025

Universidad de los Andes in Bogotá, Colombia.

# Sponsors








We would like to thank Universidad de los Andes, Columbia University, Google, Johns Hopkins University, NSF, Pensadores del Mundo (Fondo Internacional Uniandes) and University of Wisconsin-Madison for sponsoring this event.

# Schedule

## Tuesday, Dec 9

### Tuesday, December 9th

| Time | Event | | Location |
|---|---|---|---|
| 8:30 - 8:50 AM | Check-in | | **B Building** - Room 202 |
| 8:50 - 9:00 AM | Welcome Remarks | | |
| 9:00 - 10:00 AM | **Gábor Lugosi:** Recent advances in network archaeology | MINICOURSE | |
| 10:00 - 10:30 AM | Coffee Break | | |
| 10:30 - 11:30 AM | **Marina Meila:** TBD | KEYNOTE | |
| 11:30 AM - 12:00 PM | No-regret generative modeling via parabolic Monge-Ampere | Nabarun Deb | |
| 12:00 AM - 1:30 PM | Lunch Break | | |
| 1:30 - 2:00 PM | Any-dimensional machine learning | Soledad Villar | |
| 2:00 - 2:30 PM | Moving from global to local dimension estimation for manifold data | Adolfo Quiroz | |
| 2:30 - 3:00 PM | Coffee Break | | |
| 3:00 - 3:30 PM | TBD | Samory Kpotufe | |
| 3:30 - 4:00 PM | Uncertainty Quantification in Bayesian Clustering | Andres Felipe Barrientos | |
| 4:00 - 4:30 PM | Adjourn | | |
| 4:30 - 6:30 PM | **Posters / Cocktail** | | **SD Building** 7th Floor |

# Wednesday, Dec 10

## Wednesday, December 10th

| Time | Event | | Location |
|---|---|---|---|
| 8:30 - 9:00 AM | Check-in | | **Building B** - Room 402 |
| 9:00 - 10:00 AM | **Gábor Lugosi:** Recent advances in network archaeology | MINICOURSE | |
| 10:00 - 10:30 AM | Coffee Break | | |
| 10:30 - 11:30 AM | **Mauricio Velasco:** Algebraic tools for recovering measures from moments | KEYNOTE | |
| 11:30 AM - 12:00 PM | Long-Time Accuracy of Ensemble Kalman Filters for Chaotic and Machine-Learned Dynamical Systems | Daniel Sanz-Alonso | |
| 12:00 AM - 1:30 PM | Lunch Break | | |
| 1:30 - 2:00 PM | Sample Splitting and Assessing Goodness-of-fit of Time Series | Richard Davis | |
| 2:00 - 2:30 PM | Data Science under Resource Constraints: Online Fault Detection for High-dimensional Data Streams | Ana María Estrada | |
| 2:30 - 3:00 PM | Coffee Break | | |
| 3:00 - 3:30 PM | Causal inference using longitudinal modified treatment policies | Ivan Díaz | |
| 3:30 - 4:00 PM | Beyond Scores: Proximal Diffusion Models | Jeremias Sulam | |
| 4:00 - 6:00 PM | Adjourn | | |
| 6:00 - 8:00 PM | **Dinner** | | **Origen Bistro** |

# Thursday, Dec 11

## Thursday, December 11th

| Time | Event | | Location |
|---|---|---|---|
| 8:30 - 8:50 AM | Check-in | | **Building B** - Room 202 |
| 9:00 - 10:00 AM | **Bodhisattva Sen:** The James–Stein Estimator and Empirical Bayes | MINICOURSE | |
| 10:00 - 10:30 AM | Coffee Break | | |
| 10:30 - 11:30 AM | **Rebecca Willett:** TBD | KEYNOTE | |
| 11:30 AM - 12:00 PM | Quadratically Regularized Optimal Transport | Alberto Gonzalez Saenz | |
| 12:00 AM - 1:30 PM | Lunch Break | | |
| 1:30 - 2:00 PM | Statistical Properties of Rectified Flow | Gonzalo Mena | |
| 2:00 - 2:30 PM | Coherence-free Entrywise Estimation of Eigenvectors in Low-rank Signal-plus-noise Matrix Models | Keith Levin | |
| 2:30 - 3:00 PM | Coffee Break | | |
| 3:00 - 3:30 PM | Bayesian Nash Equilibrium via Machine Learning | Alvaro Riascos | |
| 3:30 - 4:00 PM | Dynamic Schrödinger Bridges beyond entropy | Camilo Hernandez | |

# Friday, Dec 12

## Friday, December 12th

| Time | Event | | Location |
|---|---|---|---|
| 8:30 - 8:50 AM | Check-in | | **Building B** - Room 402 |
| 9:00 - 10:00 AM | **Bodhisattva Sen:** The James–Stein Estimator and Empirical Bayes | MINICOURSE | |
| 10:00 - 10:30 AM | Coffee Break | | |
| 10:30 - 11:30 AM | **Mahdi Soltanolkotabi:** One Small Step, One Giant Leap: From Test-Time Tweaks to Global Guarantees | KEYNOTE | |
| 11:30 AM - 12:00 PM | Scientifically guided robust inference | Debdeep Pati | |
| 12:00 AM - 1:30 PM | Lunch Break | | |
| 1:30 - 2:00 PM | A Frequentist Theory for M-posteriors: Asymptotic and Robustness Properties | Cynthia Rush | |
| 2:00 - 2:30 PM | Robust Regression under Adversarial Contamination: Theory and Algorithms for the Welsch Estimator | Mohamed Ndaoud | |
| 2:30 - 3:00 PM | Coffee Break | | |
| 3:00 - 3:30 PM | Regularized f-Divergence Kernel Tests | Monica Ribero | |
| 3:30 - 4:00 PM | TBD | TBD | |

# Keynote Speaker Abstracts

**TBD**
Marina Mila (*University of Waterloo, Canada*)
*Time*: Tuesday, December 9, 10:30-11:30 am.
*Abstract*: TBD.

**Algebraic tools for recovering measures from moments**
Mauricio Velasco (*Universidad de la República, Uruguay*)
*Time*: Wednesday, December 11, 10:30-11:30 am
*Abstract*: The moments of a measure $\mu$ in a compact set $X \subseteq \mathbb{R}^n$ are the averages of monomials according to $\mu$. A problem with applications ranging from statistical inference to stochastic control is the (possibly approximate) recovery of a measure's density from a collection of moments. In this talk I will discuss several generalizations of the classical Christoffel-Darboux kernel method for carrying out such recovery procedures and prove several quantitative guarantees and showcase some recent applications. The talk will discuss joint work (some ongoing) with several coauthors (L. Bentancur, C. Meroni, J. Miller and D. Henrion).

**TBD**
Rebecca Willett (*University of Chicago, USA*)
*Time*: Thursday, December 12, 10:30-11:30 am
*Abstract*: TBD.

**One Small Step, One Giant Leap: From Test-Time Tweaks to Global Guarantees**
Mahdi Soltanolkotabi (*University of Southern California*)
*Time:* Friday, December 13, December 10:30 - 11:30 am
*Abstract*: Simple first-order methods like Gradient Descent (GD) remain foundational to modern machine learning. Yet, despite their widespread use, our theoretical understanding of the GD trajectory—how and why it works—remains incomplete in both classical and contemporary settings. This talk explores new horizons in understanding the behavior and power of GD across two distinct but connected fronts.

In the first part, we examine the surprising power of a single gradient step in enhancing model reasoning. We focus on test-time training (TTT)—a gradient-based approach that adapts model parameters using individual test instances. We introduce a theoretical framework that reveals how TTT can effectively handle distribution shifts and significantly reduce the data required for in-context learning, shedding light on why such simple methods often outperform expectations.

The second part turns to a more classical optimization setting: learning shallow neural networks with GD. Despite extensive study, even fitting a one-hidden-layer model to basic target functions lacks

rigorous performance guarantees. We present a comprehensive analysis of the GD trajectory in this regime, showing how it avoids suboptimal stationary points and converges efficiently to global optima. Our results offer new theoretical foundations for understanding how GD succeeds in the presence of sub-optimal stationary points.

# Minicourse Abstracts

**Recent advances in network archaeology**

Gábor Lugosi (*ICREA & Pompeu Fabra University*)

*Time*: Tuesday, December 10, 9:00-10:00 am and Wednesday, December 11, 9:00-10:00 am

*Abstract*: Large networks that change dynamically over time are ubiquitous in various areas such as social networks and epidemiology. These networks are often modeled by random dynamics, which, despite being relatively simple, give a quite accurate macroscopic description of real networks. "Network archaeology" is an area of combinatorial statistics in which one studies statistical problems of inferring the past properties of such growing networks based on present-day observations. In this minicourse, we review some simple models and recent results.

**The James–Stein Estimator and Empirical Bayes**

Bodhisattva Sen (*Columbia University*)

*Time*: Thursday, December 12, 9:00-10:00 am and Friday, December 13, 9:00-10:00 am

*Abstract*: In the normal means model $Z_i \sim N(\theta_i, 1)$, for $i = 1, \ldots, n$, the coordinatewise MLE $Z = (Z_1, \ldots, Z_n)$ is the natural estimator of $\theta = (\theta_1, \ldots, \theta_n)$. Stein's startling discovery shows that for $n \geq 3$ this estimator is inadmissible under squared error loss: the James–Stein (JS) shrinkage rule uniformly lowers risk by pulling $Z$ toward a common center (Stein, 1956; James & Stein, 1961). We will review Stein's paradox, outline the proof of inadmissibility, and introduce SURE (Stein's unbiased risk estimate) to illustrate how data-dependent shrinkage can be calibrated in practice.

We then connect JS to the Empirical Bayes (EB) perspective, treating the $\theta_i$ as i.i.d. draws from an unknown prior $G$. This viewpoint clarifies why global shrinkage arises and how estimating prior hyperparameters leads naturally to JS-type rules. We also introduce Tweedie's formula, which links posterior means to the derivative of the log marginal density, and motivates nonparametric EB. We conclude by contrasting two EB strategies: f-modeling, which estimates the marginal density of the observations, and g-modeling, which directly estimates the distribution $G$.

# Invited Speaker Abstracts

## Tuesday, Dec 10

### No-regret generative modeling via parabolic Monge-Ampere

Nabarun Deb (University of Chicago Booth)

*Abstract: We introduce a novel generative modeling framework based on a discretized parabolic Monge-Ampère PDE, which emerges as a continuous limit of the Sinkhorn algorithm commonly used in optimal transport. Our method performs iterative refinement in the space of Brenier maps using a mirror gradient descent step. We establish theoretical guarantees for generative modeling through the lens of no-regret analysis, demonstrating that the iterates converge to the optimal Brenier map under a variety of step-size schedules. As a technical contribution, we derive a new Evolution Variational Inequality tailored to the parabolic Monge-Ampère PDE, connecting geometry, transportation cost, and regret. Our framework accommodates non-log-concave target distributions, constructs an optimal sampling process via the Brenier map, and integrates favorable learning techniques from generative adversarial networks and score-based diffusion models. As direct applications, we illustrate how our theory paves new pathways for generative modeling and variational inference. This is joint work with Tengyuan Liang.*

### TBD

Samory Kpotufe (TBD)

*Abstract: TBD*

### Moving from global to local dimension estimation for manifold data

Adolfo Quiroz (Universidad de los Andes)

*Abstract: In the last two decades, the emphasis in dimension estimation for multivariate data living in manifolds, has shifted from global to local estimaion. This makes perfect sense since for certain types of data, it does happen that the dimension of the manifold changes when we move around the data set. Detecting these changes allows for a better understanding of the data and it has proved to be important in certain applications, including clustering. On the other hand, the change in goals can be theoretically supported by the fact that, for many dimension estimators, local CLTs can be proved, adapting the proofs available in the global setting. The goal of this talk is to discuss changes that must be made and evaluations required to use a global dimension estimator in a local setting. The size of the local neighborhood needs to be calibrated for each particular estimator, depending on the speed of convergence in the local CLT. In general, significantly larger neighborhoods will be required for the local estimation in comparison to the same estimator when used in the global case. We also discuss some criteria for combining local dimension estimators into a more precise one. This is joint work with Bianca Michelle Vargas.*

## Any-dimensional machine learning

Soledad Villar (Johns Hopkins University)

*Abstract: Many machine learning models have the property of being defined on a fixed set of parameters while being evaluated on objects of different sizes and dimensions without losing expressivity. Graph neural networks are a prototypical example. Recent work by Mateo Diaz and Eitan Levin introduced a formalization for any-dimensional machine learning based on representation stability. In this talk, I will discuss its applications to several models, including PDE learning.*

## Uncertainty Quantification in Bayesian Clustering

Andres Felipe Barrientos (Florida State University)

*Abstract: Bayesian clustering methods have the widely touted advantage of providing a probabilistic characterization of uncertainty in clustering through the posterior distribution. An amazing variety of priors and likelihoods have been proposed for clustering in a broad array of settings. There is also a rich literature on Markov chain Monte Carlo (MCMC) algorithms for sampling from posterior clustering distributions. However, there is relatively little work on summarizing the posterior uncertainty. The complexity of the partition space corresponding to different clusterings makes this problem challenging. We propose a post-processing procedure for any Bayesian clustering model with posterior samples that generates a credible set that is easy to use, fast to compute, and intuitive to interpret. We also provide new measures of clustering uncertainty and show how to compute cluster-specific parameter estimates and credible regions that accumulate a desired posterior probability without having to condition on a partition estimate or employ label-switching techniques. We illustrate our procedure through several applications.*

# Wednesday, Dec 11

## Long-Time Accuracy of Ensemble Kalman Filters for Chaotic and Machine-Learned Dynamical Systems

Daniel Sanz-Alonso (University of Chicago)

*Abstract: Filtering is concerned with online estimation of the state of a dynamical system from partial and noisy observations. In applications where the state is high dimensional, such as numerical weather prediction, ensemble Kalman filters are often the method of choice. This work studies the long-time accuracy of ensemble Kalman filters. We introduce conditions on the dynamics and the observations under which the estimation error remains small in the long-time horizon. Our theory covers a wide class of partially-observed chaotic dynamical systems, which includes the Navier-Stokes equations and Lorenz models. In addition, we prove long-time accuracy of ensemble Kalman filters with surrogate dynamics, thus validating the use of machine-learned forecast models in ensemble data assimilation.*

## Sample Splitting and Assessing Goodness-of-fit of Time Series

Richard Davis (Columbia University)

*Abstract: A fundamental and often final step in time series modeling is to assess the quality of fit of a proposed model to the data. Since the underlying distribution of the innovations that generate a model is often not prescribed, goodness-of-fit tests typically take the form of testing the fitted residuals for serial independence. However, these fitted residuals are inherently dependent since they are based on the same parameter estimates and thus standard tests of serial independence, such as those based on the autocorrelation function (ACF) or distance correlation function (ADCF) of the fitted residuals need to be adjusted. The sample splitting procedure in Pfister et al. (2018) is one such fix for the case of models for independent data, but fails to work in the dependent setting. In this paper sample splitting is leveraged in the time series setting to perform tests of serial dependence of fitted residuals using the ACF and ADCF. Here the first $f_n$ of the data points are used to estimate the parameters of the model and then using these parameter estimates, all of the data points are used to compute the estimated residuals. Tests for serial independence are then based on these $n$ residuals. As long as $f_n$ is asymptotically one-half the sample size, the ACF and ADCF tests of serial independence tests often have the same limit distributions as though the underlying residuals are indeed iid. In particular if the first half of the data is used to estimate the parameters and the estimated residuals are computed for the entire data set based on these parameter estimates, then the ACF and ADCF can have the same limit distributions as though the residuals were iid. This procedure ameliorates the need for adjustment in the construction of confidence bounds for both the ACF and ADCF in goodness-of-fit testing. (This is joint work with Leon Fernandes.)*

## Data Science under Resource Constraints: Online Fault Detection for High-dimensional Data Streams

Ana María Estrada Gómez (Purdue University)

*Abstract: Modern data-intensive systems—ranging from sensor networks to industrial processes—generate vast, continuous streams of high-dimensional data. In such settings, resource constraints on sensing, computation, and communication make traditional monitoring and anomaly detection methods impractical. This talk introduces a family of adaptive and active learning strategies for online fault detection that explicitly account for these limitations. By dynamically selecting which data to sample, process, or transmit, these methods achieve efficient monitoring without sacrificing statistical reliability. I will discuss applications in networked and partially observed systems, real-time anomaly detection with mobile sensors, and online batch fault diagnosis. The unifying theme is the integration of statistical learning, sequential decision-making, and uncertainty quantification to enable scalable, data-efficient online monitoring under resource constraints.*

## Causal inference using longitudinal modified treatment policies

Iván Díaz (New York University)

*Abstract: Longitudinal modified treatment policies (LMTP) have been recently developed as a novel method to define and estimate causal parameters in longitudinal settings. LMTPs allow the nonparametric definition and estimation of the joint effect of multiple categorical, ordinal, or continuous treatments measured at several time points. We present the LMTP methodology for problems in which the outcome is a time-to-event variable subject to a competing event that precludes observation of the event of interest. We discuss identification results and non-parametric locally efficient estimators that use flexible data-adaptive regression techniques to alleviate model misspecification bias, while retaining important asymptotic properties such as root-n-consistency. We present an application to the estimation of the effect of the time-to-intubation on acute kidney injury amongst COVID-19 hospitalized patients, where death by other causes is taken to be the competing event.*

## Beyond Scores: Proximal Diffusion Models

Jeremias Sulam (Johns Hopkins University)

*Abstract: Diffusion models have quickly become some of the most popular and powerful generative models for high-dimensional data. The key insight that enabled their development was the realization that access to the score—the gradient of the log-density at different noise levels—allows for sampling from data distributions by solving a reverse-time stochastic differential equation (SDE) via forward discretization, and that popular denoisers allow for unbiased estimators of this score. In this talk we will present an alternative, backward discretization of these SDEs, leading to proximal maps in place of the score. We will leverage recent results in proximal matching to learn proximal operators of the log-density and, with them, develop Proximal Diffusion Models. Theoretically, we will see that proximal-based samplers produce approximate distribution faster than score-based alternatives. Empirically, we show that two variants of proximal diffusion models achieve significantly faster convergence within just a few sampling steps compared to conventional score-matching methods.*

# Thursday, Dec 12

## Quadratically Regularized Optimal Transport

Alberto Gonzalez Sanz (Columbia University)

*Abstract: Optimal transport (OT) has become a central tool in data science and statistics. A common approach, either to design more efficient numerical methods or to mitigate the curse of dimensionality, is to regularize OT with an entropy. The most widely used choice is the logarithmic entropy, which leads to the celebrated entropic optimal transport (EOT). From a practical standpoint, EOT is successful for two main reasons: (i) the dual solutions are very smooth, and (ii) the dual problem is strongly concave. Property (i) helps overcome the curse of dimensionality in empirical estimation, while property (ii) yields algorithms with linear convergence, such as the Sinkhorn algorithm. In this talk, we will discuss alternative entropy-based regularizations of optimal transport. In particular, we will focus on the quadratically regularized optimal transport problem (QOT), which uses the $L^2$ entropy and has recently emerged as a sparse alternative to EOT. Unlike EOT, whose solutions always have full support—even for small regularization parameters—QOT solutions (or QOT plans) tend to concentrate on the support of the unregularized transport problem. However, the dual of QOT is not strongly concave, and its dual solutions are not necessarily smoother than those of classical OT. This raises two natural questions: Do the supports decrease monotonically? At what rate does sparsification occur? How quickly does the QOT cost converge to the classical OT cost? Are there algorithms with linear convergence? And does QOT still suffer from the curse of dimensionality? We will review recent theoretical results that provide answers to these questions.*

## Statistical Properties of Rectified Flow

Gonzalo Mena (Carnegie Mellon University)

*Abstract: The problem of finding a transformation mapping one distribution into another is a relevant mathematical problem with several applications in physics, genomics, etc. When this transformation is assumed to be monotonic, the above problem corresponds to finding the so-called optimal transport map, for which a rich mathematical regularity theory is available and for which a recent non-parametric estimation theory has been established. These statistical results indicate that plug-in estimators of such maps converge faster than expected for Kernel density estimators, a consequence of the extra degree of smoothness of the optimal map compared to the original densities. Moreover, a central limit theorem has been established for such estimators under suitable bandwidth selection, enabling uncertainty quantification. The main drawback is that their computation is typically intractable as it relies on solving an optimal transport problem in the continuum, for which we can only obtain approximated solutions. To deal with these issues, we propose rectified transport as an alternative to optimal transport. The rectified map (Liu et al., 2022) is a relaxation of optimal transport that recovers optimal transport if sufficient constraints are added. Unlike optimal transport, the rectified map is computed pointwise by solving an ordinary differential equation with a velocity field given by a conditional expectation. Moreover, the computation of a plug-in estimator for the rectified flow amounts to solving a sequence of non–parametric regression problems. Rectified maps are typically used in diffusion models, but little is known about their regularity and sample properties. This talk will establish an elementary regularity theory, showing that the population rectified map also has more regularity than the underlying densities. Based on this theory, we derive a bias and variance analysis for this estimator and a central limit theorem. Our results indicate that this estimator benefits from the enhanced regularity of the transport map. However, the benefits are more modest compared to optimal transport, presumably because the rectified map is typically less structured than the optimal transport. This is joint work with Arun Kumar Kuchibhotla and Larry Wasserman.*

## Coherence-free Entrywise Estimation of Eigenvectors in Low-rank Signal-plus-noise Matrix Models

Keith Levin (University of Wisconsin–Madison)

*Abstract: Spectral methods are widely used to estimate eigenvectors of a low-rank signal matrix subject to noise. These methods use the leading eigenspace of an observed matrix to estimate this low-rank signal. Typically, the entrywise estimation error of these methods depends on the coherence of the low-rank signal matrix with respect to the standard basis. In this work, we present a novel method for eigenvector estimation that avoids this dependence on coherence. Assuming a rank-one signal matrix, under mild technical conditions, the entrywise estimation error of our method provably has no dependence on the coherence under Gaussian noise (i.e., in the spiked Wigner model), and achieves the optimal estimation rate up to logarithmic factors. Simulations demonstrate that our method performs well under non-Gaussian noise and that an extension of our method to the case of a rank-r signal matrix has little to no dependence on the coherence. In addition, we derive new metric entropy bounds for rank-r singular subspace recoverys under the two-to-infinity distance, which may be of independent interest. We use these new bounds to improve the best known lower bound for rank-r eigenspace estimation under two-to-infinity distance.*

## Bayesian Nash Equilibrium via Machine Learning

Alvaro Riascos (Universidad de los Andes)

*Abstract: Spectral methods are widely used to estimate eigenvectors of a low-rank signal matrix subject to noise. These methods use the leading eigenspace of an observed matrix to estimate this low-rank signal. Typically, the entrywise estimation error of these methods depends on the coherence of the low-rank signal matrix with respect to the standard basis. In this work, we present a novel method for eigenvector estimation that avoids this dependence on coherence. Assuming a rank-one signal matrix, under mild technical conditions, the entrywise estimation error of our method provably has no dependence on the coherence under Gaussian noise (i.e., in the spiked Wigner model), and achieves the optimal estimation rate up to logarithmic factors. Simulations demonstrate that our method performs well under non-Gaussian noise and that an extension of our method to the case of a rank-r signal matrix has little to no dependence on the coherence. In addition, we derive new metric entropy bounds for rank-r singular subspace recoverys under the two-to-infinity distance, which may be of independent interest. We use these new bounds to improve the best known lower bound for rank-r eigenspace estimation under two-to-infinity distance.*

## Dynamic Schrödinger Bridges beyond entropy

Camilo Hernandez (University of Southern California)

*Abstract: Over the past decade, the Schrödinger bridge problem has emerged as a central tool for modeling the evolution of uncertainty in dynamical systems. It describes the most likely stochastic evolution connecting two observed distributions and forms the dynamic, entropy-regularized analogue of optimal transport. Unlike its static counterpart, the dynamic formulation explicitly models the temporal evolution of probability flows, a feature that has proven essential in modern generative modeling via diffusion processes, where data distributions are learned through time-dependent stochastic transformations. While the classical Schrödinger problem relies on relative entropy, which ensures analytical convenience, this choice can be restrictive: it produces diffuse stochastic trajectories and imposes a specific log-penalty structure that may not capture robustness or structural features of the system. General divergence-based penalties offer greater flexibility, allowing the model to capture alternative notions of discrepancy and sensitivity. Indeed, insights from quadratically regularized optimal transport, a static analogue of the problem considered in this project, suggest that $L^2$-type penalizations can induce sparsity in the transport plan, emphasizing the most relevant trajectories while downweighting minor fluctuations. In this work, we leverage stochastic control tools and convex duality to extend these ideas to the dynamic, path-space setting.*

# Friday, Dec 13

## Scientifically guided robust inference

Debdeep Pati (University of Wisconsin - Madison)

*Abstract: Motivated by the need for interpretable yet flexible models for modern scientific applications, we shall propose a likelihood-based inference constrained to lie close to a parametric family. We shall draw connections with nonparametric Bayesian inference constrained to lie around a base distribution and show favorable asymptotic properties of the proposed estimator. Finally, we shall explore a few applications from neuroscience, trustworthy machine learning and inference with survey data.*

## A Frequentist Theory for M-posteriors: Asymptotic and Robustness Properties

Cynthia Rush (Columbia University)

*Abstract: An M-estimator is a general class of estimators in statistics that are defined as the minimizer of an objective function, typically derived from a loss or score function. In this talk, I will introduce a theoretical framework for a wide class of generalized posteriors that can be viewed as the natural Bayesian posterior counterpart of the class of M-estimators in the frequentist world and we refer to the members of this class as M-posteriors. I will discuss asymptotic normality of the M-posteriors under mild conditions on the M-estimation loss and the prior, showing that M-posteriors contract in probability around a normal distribution centered at the M-estimators, which provides frequentist consistency and suggests some degree of robustness depending on the reference M-estimator. Moreover, I will formalize the robustness properties of the M-posteriors by providing a new characterization of the posterior influence function and a novel definition of breakdown point adapted for posterior distributions. This is joint work with Juraj Marusic and Marco Avella Medina.*

## Robust Regression under Adversarial Contamination: Theory and Algorithms for the Welsch Estimator

Mohamed Ndaoud (ESSEC Business School)

*Abstract: Convex and penalized robust regression methods often suffer from a persistent bias induced by large outliers, limiting their effectiveness in adversarial or heavy-tailed settings. In this talk, we study a smooth redescending non-convex M-estimator, specifically the Welsch estimator, and show that it can eliminate this bias whenever it is statistically identifiable. We focus on high-dimensional linear regression under adversarial contamination, where a fraction of samples may be corrupted by an adversary with full knowledge of the data and underlying model. A central technical contribution of this paper is a practical algorithm that provably finds a statistically valid solution to this non-convex problem. We establish three main guarantees: (a) non-asymptotic minimax-optimal deviation bounds under contamination, (b) improved unbiasedness in the presence of large outliers, and (c) asymptotic normality, yielding statistical efficiency as the sample size grows.*

## Regularized f-Divergence Kernel Tests

Monica Ribero (Google)

*Abstract: We propose a framework to construct practical kernel-based two-sample tests from the family of f-divergences. The test statistic is computed from the witness function of a regularized variational representation of the divergence, which we estimate using kernel methods. The proposed test is adaptive over hyperparameters such as the kernel bandwidth and the regularization parameter. We provide theoretical guarantees for statistical test power across our family of f-divergence estimates. While our test covers a variety of f-divergences, we bring particular focus to the Hockey-Stick divergence, motivated by its applications to differential privacy auditing and machine unlearning evaluation. For two-sample testing, experiments demonstrate that different f-divergences are sensitive to different localized differences, illustrating the importance of leveraging diverse statistics. For machine unlearning, we propose a relative test that distinguishes true unlearning failures from safe distributional variations.*

# Accepted Posters

## An uncertainty-aware framework for data-efficient multi-view animal pose estimation

Lenny Aharon (Columbia University)

*Multi-view pose estimation is essential for quantifying animal behavior in scientific research, yet current methods struggle to achieve accurate tracking with limited labeled data and suffer from poor uncertainty estimates. We address these challenges with a comprehensive framework combining novel training and post-processing techniques, and a model distillation procedure that leverages the strengths of these techniques to produce a more efficient and effective pose estimator. Our multi-view transformer (MVT) utilizes pretrained backbones and enables simultaneous processing of information across all views, while a novel patch masking scheme learns robust cross-view correspondences without camera calibration. For calibrated setups, we incorporate geometric consistency through 3D augmentation and a triangulation loss. We extend the existing Ensemble Kalman Smoother (EKS) post-processor to the nonlinear case and enhance uncertainty quantification via a variance inflation technique. Finally, to leverage the scaling properties of the MVT, we design a distillation procedure that exploits improved EKS predictions and uncertainty estimates to generate high-quality pseudo-labels, thereby reducing dependence on manual labels. Our framework components consistently outperform existing methods across three diverse animal species (flies, mice, chickadees), with each component contributing complementary benefits. The result is a practical, uncertainty-aware system for reliable pose estimation that enables downstream behavioral analyses under real-world data constraints.*

## Assessing The Relation between Cardiovascular Risk and Mortality Through Multivariate Logistic Regression Models

Nicolás Arango (Universidad del Quindio)

*Cardiovascular diseases (CVDs), encompassing conditions like coronary artery disease (CAD), stroke, and peripheral vascular disease, represent a major global health challenge, being the leading cause of morbidity and mortality worldwide. Cardiovascular risk quantifies the probability or likelihood of an individual experiencing a significant Cardiovascular (CV) event, either fatal or non-fatal, over a specified period (typically 5 to 10 years) [Kannel et al. (2004)]. This risk is influenced by a combination of modifiable factors (e.g., dyslipidemia, hypertension, smoking, diabetes, obesity, lifestyle) and non-modifiable factors (e.g., age, gender, family history, genetics, etc) [Moham et al. (2016)]. Assessing CV risk involves simple risk factor counts and sophisticated quantitative algorithms derived from extensive cohort studies. These algorithms often use logistic regression or Cox proportional hazards to generate risk scores based on multiple factors. Prominent examples include the Framingham Risk Score (FRS) [Kannel et al. (1976)], the PROCAM score [Assman et al. (2007)], the Systematic Coronary Risk Evaluation (SCORE) project [Conroy et al. (2003)], and the REGICOR score (a Spanish Framingham adaptation). These scores typically use variables such as age, genre, smoking habits, blood pressure, and cholesterol levels. The FRS is widely used, with adaptations like the Framingham-Colombia score (applying a 0.75 calibration factor) developed for this population [Alvarez et al. (2017)]. Developing and validating accurate risk scores for fatal or non-fatal CVD within specific populations presents considerable complexity and challenges, including the cost of large cohort studies and applicability across different genetic and environmental backgrounds [Alvarez et al. (2017)]. Consequently, data-driven approaches using machine learning (ML) on large datasets like electronic health records (EHR) have emerged, offering a potential for identifying complex patterns and improving prediction. In Colombia, CVDs are the primary cause of death, underscoring the need for local risk assessment. This study addresses this challenge by analyzing the intricate relationship between various cardiovascular risk assessment methods and actual mortality outcomes within a specific cohort of patients from Armenia-Quindío, Colombia. The data originates from individuals enrolled in a cardiovascular risk program managed by the municipal first-level healthcare network (REDSALUD Armenia) between 2013 and 2018.*

## Autoregressive Time Series Alignment

Ernesto Araya Valdivia (LMU)

*We study the problem of aligning time series databases, where a multivariate time series is observed along with a perturbed and permuted version, and the goal is to recover the unknown matching between them. To model this, we introduce a probabilistic framework in which both series follow a correlated vector autoregressive (VAR) process jointly. This generalizes the classical problem of matching independent point clouds to the time series setting, with envisaged applications in privacy and sensor fusion. We derive the maximum likelihood estimator (MLE), leading to a quadratic optimization over permutations, and theoretically analyze a simpler estimator based on linear assignment. For the linear assignment approach, we establish recovery guarantees, identifying correlation thresholds that allow for perfect or partial recovery. We also explore convex relaxations of the MLE, including relaxations over the Birkhoff polytope, which allow the joint estimation of the hidden permutation and the autoregressive process parameters. To solve it, we propose an algorithm based on alternating optimization. Empirically, we find that the linear assignment method often matches or outperforms MLE relaxations, even when the latter have oracle access to the underlying VAR parameters, for recovering the matching. These findings highlight the theoretical and practical effectiveness of efficient algorithms for structured time series alignment.*

## Random point configurations on the sphere and logarithmic energy

Federico Carrasco (UdelaR, Montevideo)

*The logarithmic energy of a configuration of points on the sphere measures its degree of repulsion. In other words, the more separated the points are from each other, the lower the energy and the better the distribution of the configuration. Such repulsive properties are highly desirable in problems of polynomial interpolation on the sphere and, more recently, in machine learning on non-Euclidean geometries. In this work, we study the average logarithmic energy of the solutions to the polynomial eigenvalue problem for Gaussian random matrices, recovering known results for the Shub-Smale polynomials and the Spherical Ensemble. Our results suggest that the solutions to this problem behave favorably for spherical interpolation, and therefore could be useful for function learning. This opens up novel approaches to function approximation and could thus play a significant role in geometric machine learning models. We will present the main result, discuss its implications, and outline possible directions for future research.*

## Effect of the metabolism of tacrolimus on renal function mediated through BK Virus in patients with kidney transplant, using causal mediation analysis

Santiago Castro (Universidad de Los Andes)

*Managing the dose of an immunosuppressant such as tacrolimus, after kidney transplant, is important to reduce the possibility of damage of renal function. Metabolism of tacrolimus varies across patients and could have an impact on renal function. We carried out an observational retrospective study, that evaluates the association of metabolism of tacrolimus with the estimated glomerular filtration rate (eGFR) measured 12 months after transplantation. Multiple regression models were used to evaluate the association between metabolism and eGFR. The evaluation of such association included a causal mediation analysis for the estimation of the direct effect of metabolism rate on renal function after renal transplantation and the indirect effect of metabolism rate on renal function through the presence of virus BK. To estimate such effects, we made use of the causal framework of potential outcomes. We made the analysis with data of a university hospital in Colombia.*

## Computational and statistical lower bounds for low-rank estimation under general inhomogeneous noise

Debsurya De (Johns Hopkins University)

*Recent work has generalized several results concerning the well-understood spiked Wigner matrix model of a low-rank signal matrix corrupted by additive i.i.d. Gaussian noise to the inhomogeneous case, where the noise has a variance profile. In particular, for the special case where the variance profile has a block structure, a series of results identified an effective spectral algorithm for detecting and estimating the signal, identified the threshold signal strength required for that algorithm to succeed, and proved information-theoretic lower bounds that, for some special signal distributions, match the above threshold. We complement these results by studying the computational optimality of this spectral algorithm. Namely, we show that, for a much broader range of signal distributions, whenever the spectral algorithm cannot detect a low-rank signal, then neither can any low-degree polynomial algorithm. This gives the first evidence for a computational hardness conjecture of Guionnet, Ko, Krzakala, and Zdeborová (2023). With similar techniques, we also prove sharp information-theoretic lower bounds for a class of signal distributions not treated by prior work. Unlike all of the above results on inhomogeneous models, our results do not assume that the variance profile has a block structure, and suggest that the same spectral algorithm might remain optimal for quite general profiles. We include a numerical study of this claim for an example of a smoothly-varying rather than piecewise-constant profile. Our proofs involve analyzing the graph sums of a matrix, which also appear in free and traffic probability, but we require new bounds on these quantities that are tighter than existing ones for non-negative matrices, which may be of independent interest.*

## Conformal Mixed-Integer Constraint Learning with Feasibility Guarantees

Mateo Dulce Rubio (New York University)

*We propose Conformal Mixed-Integer Constraint Learning (C-MICL), a novel framework that provides probabilistic feasibility guarantees for data-driven constraints in optimization problems. While standard Mixed-Integer Constraint Learning methods often violate the true constraints due to model error or data limitations, our C-MICL approach leverages conformal prediction to ensure feasible solutions are ground-truth feasible. This guarantee holds with probability at least , under a conditional independence assumption. The proposed framework supports both regression and classification tasks without requiring access to the true constraint function, while avoiding the scalability issues associated with ensemble-based heuristics. Experiments on real-world applications demonstrate that C-MICL consistently achieves target feasibility rates, maintains competitive objective performance, and significantly reduces computational cost compared to existing methods. Our work bridges mathematical optimization and machine learning, offering a principled approach to incorporate uncertainty-aware constraints into decision-making with rigorous statistical guarantees.*

## Uniqueness and Regularity of Solutions to Adversarial Binary Classification with Quadratic Cost

Yaling Hong (University of Wisconsin-Madison)

*We study the uniqueness and regularity of solutions to adversarial binary classification with quadratic cost in an agnostic learner setting. The optimal adversarial attacks and optimal robust classifiers can be recovered by solving a primal-dual optimal transport problem as shown in previous works, where the adversarial strategy can be viewed as a generalized barycenter of data distributions. We focus specifically on quadratic cost with two balanced marginals and show that under the assumptions that they are absolutely continuous with full support: 1) The optimal adversarial attacks are absolutely continuous with full support and unique. 2) The optimal robust classifiers have bounded Hessians and are unique up to a constant.*

## Classification of Separable Nonlinear Signals: Performance Limits

Pedro Izquierdo Lehmann (Johns Hopkins University)

*This paper considers the statistical problem of classifying a signal corrupted by noise. We focus on separable nonlinear signals: linear combinations of functions that depend nonlinearly on the problem's parameters. We assess how the Bayes error of the classification task worsens due to noise, considering the regularity of the data's statistical distribution and the structure of the separable nonlinear signal. Typically, the noise level and the signal's structure depend on the measuring instrument used. Therefore, our analysis provides guidance on how to prioritize error mitigation when developing new measurement devices. We demonstrate our findings by analyzing two cases: signals as sums of complex exponentials, relevant to nuclear magnetic resonance data, and signals as sums of Gaussians, relevant to microscopy.*

## Zeroth-Order Langevin Monte Carlo via SPSA under Noisy Function Measurements

Hongbo Li (Michigan State University)

*In sampling problems, gradient-based schemes such as Langevin Monte Carlo (LMC) mix faster than non-gradient-based methods, but their applicability is limited by access to gradient. In practice, gradients are often unavailable and function evaluations are noisy—e.g., stochastic simulators or black-box simulators, so we propose LMC with Simultaneous Perturbation Stochastic Approximation (LMC-SPSA) under noise, which approximates the gradient of the target log-density using two noisy function evaluations per iteration. We prove, under noisy gradient estimates, that LMC-SPSA converges in distribution by proving convergence in Wasserstein distance. Furthermore, we construct a diminishing step size schedule $h_k \to 0$ that still drives the Wasserstein error bound to convergence, extending convergence guarantees beyond the constant-step setting. Analytically, we sharpen the dominant dimension dependence of the Wasserstein error from $O(p^4)$ to $O(p^2)$ (with $p$ denoting the dimension), and support this analysis with numerical results. Numerical experiments are conducted to verify the performance of LMC-SPSA with noise.*

## Estimating Morse Information from Samples

Daniel López (Johns Hopkins University)

*Understanding the topology of manifolds from finite samples is a fundamental challenge in modern data analysis. We present a novel statistical framework for estimating Morse information—critical points and their indices—directly from point cloud data sampled on Riemannian manifolds. By exploiting the geometric structure of height functions and their critical points, our approach evades the need for advanced algebraic machinery, offering instead an elegant and computationally tractable alternative. Our method proceeds in two stages: first, we identify "witnesses"—sample points that serve as reliable proxies for true critical points—using carefully designed local geometric tests. Second, we estimate the Morse index at each witness through local Hessian approximation via $k$-nearest neighbors. We establish strong consistency guarantees, proving that both the locations of critical points and their "topological" (Morse) indices are recovered with high probability as the sample size grows. This work bridges classical Morse theory with modern statistical learning, opening new avenues for topological data analysis on manifolds without requiring persistence homology or simplicial complex constructions.*

## Weighted Random Dot Product Graphs

Bernardo Marenco (UdelaR, Montevideo)

*Modeling of intricate relational patterns has become a cornerstone of contemporary statistical research and related data science fields. Networks, represented as graphs, offer a natural framework for this analysis. This paper extends the Random Dot Product Graph (RDPG) model to accommodate weighted graphs, markedly broadening the model's scope to scenarios where edges exhibit heterogeneous weight distributions. We propose a nonparametric weighted (W)RDPG model that assigns a sequence of latent positions to each node. Inner products of these nodal vectors specify the moments of their incident edge weights' distribution via moment-generating functions. In this way, and unlike prior art, the WRDPG can discriminate between weight distributions that share the same mean but differ in other higher-order moments. We derive statistical guarantees for an estimator of the nodal's latent positions adapted from the workhorse adjacency spectral embedding, establishing its consistency and asymptotic normality. We also contribute a generative framework that enables sampling of graphs that adhere to a (prescribed or data-fitted) WRDPG, facilitating, e.g., the analysis and testing of observed graph metrics using judicious reference distributions. The paper is organized to formalize the model's definition, the estimation (or nodal embedding) process and its guarantees, as well as the methodologies for generating weighted graphs, all complemented by illustrative and reproducible examples showcasing the WRDPG's effectiveness in various network analytic applications.*

## A theoretical framework for M-posteriors: frequentist guarantees and robustness properties

Juraj Marusic (Columbia University)

*We provide a theoretical framework for a wide class of generalized posteriors that can be viewed as the natural Bayesian posterior counterpart of the class of M-estimators in the frequentist world. We call the members of this class M-posteriors and show that they are asymptotically normally distributed under mild conditions on the M-estimation loss and the prior. In particular, an M-posterior contracts in probability around a normal distribution centered at an M-estimator, showing frequentist consistency and suggesting some degree of robustness depending on the reference M-estimator. We formalize the robustness properties of the M-posteriors by a new characterization of the posterior influence function and a novel definition of breakdown point adapted for posterior distributions. We illustrate the wide applicability of our theory in various popular models and illustrate their empirical relevance in some numerical examples.*

## Spectral Methods for Polynomial Optimization

Elvira Moreno (Caltech)

*We present a hierarchy of tractable relaxations to obtain lower bounds on the minimum value of a polynomial over a constraint set defined by polynomial equations. In contrast to previous convex relaxation techniques for this problem, our method is based on computing the smallest generalized eigenvalue of a pair of matrices derived from the problem data, which can be accomplished for large problem instances using off-the-shelf software. We characterize the algebraic structure in a problem that facilitates the application of our framework, and we observe that our method is applicable for all polynomial optimization problems with bounded constraint sets. Our construction also yields a nested sequence of structured convex outer approximations of a bounded algebraic variety with the property that linear optimization over each approximation reduces to an eigenvalue computation. Finally, we present numerical experiments on representative problems in which we demonstrate the scalability of our approach compared to convex relaxation methods derived from sums-of-squares certificates of nonnegativity.*

## Identification of Laguerre-Gauss ($LG_{p,l}$) Modes in Near and Far-Field Using a Deep Learning Approach

Dudbil Olvasada Pabon Riaño (Universidad Autónoma de Bucaramanga)

*Optical vortices, beams of light with helical wavefronts, are fundamental in emerging fields such as high-capacity optical communications and quantum manipulation. A key property of these beams is their topological charge ($l$) and radial index ($p$), which define the beam's state. However, the diagnosis and identification of these modes in an experimental setting is a significant challenge. The 2D intensity patterns captured by a detector degrade and vary drastically with propagation distance ($z$) and, more critically, are severely distorted by atmospheric turbulence. This work presents the design and implementation of a robust classifier based on Deep Learning to overcome these challenges. A large-scale synthetic dataset was generated by simulating the propagation of Laguerre-Gauss ($LG_{p,l}$) beams through a Kolmogorov atmospheric turbulence model. The dataset includes intensity pattern images in both the near-field (Fresnel) and far-field (Fraunhofer) regimes to ensure model robustness. A Convolutional Neural Network (CNN) with a flexible architecture (based on GlobalAveragePooling2D) was trained to classify the beam's state ($p$, $l$) from a single intensity image. The results demonstrate that the model can identify optical modes with high accuracy in real-time, showing remarkable resilience to noise, turbulence distortion, and variability in propagation distance. This project serves as a successful case study on how data science techniques can solve complex metrology and diagnostic problems in applied physics.*

## Neural Network Estimation of Counterfactual Bayesian Nash Equilibria

Ana María Patrón (Universidad de Los Andes)

*This paper investigates optimal auction design using neural networks to learn counterfactual equilibria in incomplete-information (Bayesian) games. Equilibrium specifications often lead to partial differential equation systems without closed-form solutions. To address the challenges of computing Bayesian Nash Equilibria, I employ the Neural Pseudogradient Ascent (NPGA) method by [Bic+21] to approximate equilibrium bid functions in symmetric Bayesian auction games with continuous type and action spaces. Player bid functions are represented by neural networks trained via self-play, with input layers sampling valuations (unknown but potentially estimated in practice) and output layers producing bid schedules. Ex-post utility functions define each network's objective function, and, due to their discontinuity, evolutionary strategy gradients are used instead of standard backpropagation. I apply this framework to energy auctions, recovering unknown valuation functions from observed bid data and combining firms' profit-maximization conditions with bootstrapping procedures inspired by [HM10].*

## Critical thresholds in stochastic rumors on trees

Jhon Franklin Puerres (Universidade Federal de Pernambuco)

*The vertices of a tree represent individuals in one of three states: ignorant, spreader, or stifler. A spreader transmits the rumor to any of its nearest ignorant neighbors at rate one. At the same rate, a spreader becomes a stifler after contacting nearest neighbor spreaders or stiflers. The rumor survives if, at all times, there exists at least one spreader. We consider two extensions and prove phase transition results for rumor survival. First, we consider the infinite Cayley tree of coordination number $d+1$, with $d \geq 2$, and assume that as soon as an ignorant hears the rumor, the individual becomes spreader with probability $p$, or stifler with probability $1-p$. Using coupling with branching processes we prove that for any $d$ there is a phase transition in $p$ and localize the critical parameter. By refining this approach, we extend the study to an inhomogeneous tree with hubs of degree $d+1$ and other vertices of degree at most $k = o(d)$. The purpose of this extension is to illustrate the impact of the distance between hubs on the dissemination of rumors in a network. To this end, we assume that each hub is, on average, connected to $\alpha(d+1)$ hubs, with $\alpha \in (0,1]$, via paths of length $h$. We obtain a phase transition result in $\alpha$ in terms of $d$, $k$, and $h$, and we show that in the case of $k = \Theta(\log d)$ phase transition occurs iff $h \lesssim \Theta(\log d/(\log\log d))$.*

## Hierarchical Risk Parity with Autoencoder-based Covariance Estimation for Portfolio Allocation in the Latin American Market

Gonzalo Ramírez (Universidad del Rosario)

*This research presents an extension of the Hierarchical Risk Parity (HRP) portfolio allocation framework through the integration of Autoencoder-based covariance estimation. Traditional covariance matrices, often affected by noise and instability in emerging markets, are replaced by representations learned through unsupervised neural networks that capture non-linear dependencies among asset returns. The Autoencoder reconstructs denoised return series, producing a positive semi-definite covariance matrix that serves as input for the HRP quasi-diagonalization and hierarchical clustering stages. Using equities from the Latin American NUAM market, the proposed method improves portfolio diversification and out-of-sample risk control relative to the sample and Ledoit–Wolf covariance estimators.*

## Optimal Neural Network Search in Supervised Learning Contexts

Mateo Rodriguez (Universidad de los Andes)

*No matter the type of Supervised Learning problem a practitioner faces, a Neural Network could be a suitable candidate for solving it. Therefore, the standard procedure consists of fixing the network's graph structure and training the associated weights function to minimize a selected loss function. This approach forces practitioners to use a pre-determined network structure and restricts their capacity to modify it to give a better solution to the problem. This fact typically leads to practitioners overestimating the network size and assuming unnecessary costs. Motivated by this fact, in this work we propose a novel way to modify the network's structure during the training. So, exploiting the idea of searching over the network ´s structure space, we can use the Metropolis-Hastings Algorithm (MH) to explore a graph that describes an efficient way of visiting the different architectures. In more general contexts, it is a common issue to face a complexity problem, as complex models are associated with higher storage costs, longer training times, and harder computational challenges, among other difficulties. So, we provide Statistical Criteria based on Vapnik-Chervonenkis' Minimal Structural Risk ideas and some generalizations of it, to back up a low loss function and simple structure model selection. We also explain how to include these criteria as part of the MH Algorithm. Finally, we generalize the statistical consistency results from the Minimal Structural Risk theory to the best-fit function context. Hence, we get similar results in the Regression context and a more general theory.*

## Differentially Private Longitudinal Linear Regression

Getoar Sopa (Columbia University)

*We study user-level differentially private estimation and inference for longitudinal (panel) linear regression under Gaussian Differential Privacy ($\mu$-GDP). Our core idea is to calculate user-level local linear regression estimates and privately aggregate them using a novel user-level private mean estimation algorithm that automatically adapts to properties of the data. Concretely, each user forms a local OLS estimator; a DP trimming algorithm starts by projecting local estimates into a ball of very large radius that with high probability contains all local estimates and iteratively and privately shrinks and recenters the ball until it reaches an 'optimal' radius, yielding an aggregate estimator whose privacy error decays with both the number of users and the panel length. We provide finite-sample error bounds and show asymptotic normality; if the second moments of the users' covariates agree, our private estimate attains the asymptotic efficiency of the non-private OLS estimator. We further develop a user-level private feasible GLS procedure by privately estimating a the Toeplitz error covariance via DP autocovariances, and we analyze a two-group (e.g., treatment–control) extension that recovers DP differences in coefficients whilst still ensuring the privacy error decays with both the size and length of the dataset. Finally, we propose a user-level differentially private method for heteroskedasticity- and autocorrelation-robust covariance estimation in this setting, and corresponding asymptotically valid Wald tests that remain private by post-processing.*

## Singularities and Next-Generation Scientific Machine Learning

Juan Esteban Suarez (LMU)

*Solving partial differential equations (PDEs) with singularities remains a central challenge in scientific computing, as classical numerical schemes often fail to capture discontinuities and sharp transitions. We present a variational framework for approximating PDEs with singular solutions through hybrid surrogate models that combine the distributional structure of the problem with the PDE learning paradigm. These surrogates accurately approximate and parameterize discontinuity manifolds, achieving (in some settings) geometric convergence rates that outperform both traditional solvers and state-of-the-art machine learning approaches. Beyond numerical performance, the framework offers fundamental insights into the computability and complexity theory of PDEs, and enables data-driven morphometrics for biological shape analysis. In this context, modeling singularities uncovers the topological structure of the medial axis, providing a deeper geometric understanding of shape evolution and variability across developmental stages. Overall, this work bridges variational formulations, approximation theory, and scientific computing, leveraging singularities as structural features to advance the next generation of scientific machine learning methods.*

## Nonparametric Two-Sample Testing for Random Graphs via Bootstrap Subsampling

Luis Tejon (Universidad de Los Andes)

*This work addresses the problem of two-sample hypothesis testing for graphs. A discussion is provided on different random graph models and their applicability to the two-sample problem. A nonparametric procedure is proposed that combines latent position models on the sphere (Random Spherical Graph, RSG) with bootstrap-style resampling techniques, specifically through induced sub-graph subsampling, to approximate the null distribution without relying on concentration inequalities. Controlled simulations are presented to evaluate the method's performance in terms of Type I and Type II errors under various scenarios of structural complexity. The results show that the approach is flexible and robust across different community configurations and connection scales, providing a generalizable tool for statistical inference in complex, high-dimensional networks with limited sample sizes.*

## Local Differentially Private Bayesian Networks

Maria Gabriela Urango Llorente (Universidad de Los Andes)

*As Machine Learning (ML) models increasingly influence decisions in different domains such as hiring, lending and public resource allocation, concerns about privacy have become central. However, introducing privacy-preserving mechanisms often degrades model utility. This works proposes an Expectation-Maximization approach to estimate the parameters of a Naive Bayes model, explicitly accounting for the private variables. Using the restructured Adult dataset, we train an NB model that predicts income based on both public attributes (average weekly hours worked and marital status) and private ones (race and education level), which are locally perturbed using randomized response. Since the data aggregator only observes privatized versions of these attributes, we show how the proposed method is able to account for partial observability for parameter estimation under local privacy.*

## Differences in length of Days in Intensive Care Units Between Teaching Hospitals and Centers of Excellence Among Medicare Adults with Parkinson Disease in 2019

Sarah Valencia (UTHealth Houston School of Public Health)

*It is unknown where there are differences in the length of days in the intensive care unit (ICU) among Parkinson's disease (PD) 2019 beneficiaries when they first visit a teaching Hospital versus a Center of Excellence. Secondary data analyses of inpatient Medicare part A beneficiary's 65+ years old with a PD diagnosis and hospitalized between 01/01/2019-12/31/2019 were conducted to evaluate differences in the length of days in ICU between individuals first visiting a teaching Hospital versus first visiting a center of excellence in the USA. Challenges with the classification of centers of excellence and teaching hospitals, as well as the data management process of electronic medical records from Medicare from claims, to hospitals to summarizing patients, as well as the results of the statistical comparisons, adjusting for Bonferroni correction, are reported. Sensitivity analyses allow statisticians to reach the same conclusions with heavily skewed data.*

## Robust Estimation under Outcome Dependent Right Censoring in Huntington Disease: Estimators for Low and High Censoring Rates

Jesus Vazquez (Johns Hopkins University)

*Regression models with censored covariates are common in clinical and health research, yet standard methods often yield biased results when the censoring mechanism depends on the outcome. We address this gap by developing three consistent estimators for regression with a right-censored covariate under outcome-dependent censoring: two augmented inverse probability weighted (AIPW) estimators and the maximum likelihood estimator (MLE). We establish their theoretical properties, derive corrected sandwich variance estimators that account for the additional variability introduced by weight estimation, and demonstrate their implementation using Weibull accelerated failure time models for nuisance distributions. Through simulations, we show that the MLE is preferred when the censoring rate is low, while weighted estimators are preferred under high censoring rates. We applied all estimators to a Huntington disease observational study to analyze the composite Unified Huntington's Disease Rating Scale as a function of time to mild cognitive impairment.*

## Geometric Factor Analysis

Congwei Yang (University of Wisconsin-Madison)

*In statistics literature, the Wasserstein distance has been employed as a theoretical device. On the other hand, factor analysis has long been a cornerstone of statistics and data science. Our paper introduces a novel framework for factor rotations via optimization on the Stiefel manifold using a Wasserstein distance-based energy, offering a practical approach for statistical inference. Given the broad application of factor analysis, we demonstrate our proposed method achieves near-parametric convergence rate, and a practical online-SGD algorithm for estimation. We further present its effectiveness in identifying latent factors for image feature extraction, climate modeling, and network analysis.*