The Abdus Salam
International Centre
for Theoretical Physics

Workshop on Learning and Inference from Structured Data: Universality, Correlations and Beyond | (SMR 3850)

03 Jul 2023 - 07 Jul 2023
ICTP, Trieste, Italy

P01 - ABBOTT Charles Michael

Far from Asymptopia

P02 - ADEBAYO Segun

Inference from Meteorological Parameter and Key Performance Indicator Data for Quality of Service Prediction in Mobile Communication Network

P03 - BARDONE Lorenzo

Beyond Gaussian models: a mathematical framework for the analysis of real-world data structures

P04 - BARRIOS MORALES Guillermo

Quasi-universal scaling and near-critical behavior in neural systems

P05 - BEREUX Nicolas

Learning a restricted Boltzmann machine using biased Monte Carlo sampling

P06 - DUDEJA Rishabh

Spectral Universality of Regularized Linear Regression with Nearly Deterministic Sensing Matrices

P07 - GUEDDARI Mohammed-Younes

Approximate Message Passing for elliptic random matrices and its applications to theoretical ecology

P08 - HOSHISASHI Kentaro

Deep Learning with No-arbitrage constraints

P09 - JIMOH Abdulganiyu

APPLICATION OF ARTIFICIAL INTELLIGENCE (AI) AND COLLECTIVE INTELLIGENCE (CI) FOR DIAGNOSIS OF BREAST CANCER; CASE STUDY OF AFRICAN WOMEN

P10 - KAPLA Daniel

Generalized Linear Models: Multi-Array Valued Predictors

P11 - LAM Duy Khanh

UNIVERSAL INVESTMENT STRATEGIES BY ONLINE LEARNING UNDER LATENT INFORMATION-DEPENDENCE STRUCTURE

P12 - LIAO Zhenyu

Random Matrix Methods for Machine Learning: "Lossless" Compression of Large and Deep Neural Networks

P13 - MARCONDES Diego

A data-driven systematic, consistent ans non-exhaustive approach to Model Selection

P14 - MASSA Emanuele

Correction of overfitting bias in ML regression

P15 - MORETTIN Alberto Pedro

Wavelet-based Clustering and Classification

P16 - OZEL KADILAR Gamze

Machine Learning Techniques for Earthquake Size and Magnitude Prediction in Turkey

P17 - RENÉ Alexandre

A robust criterion for selecting among fitted models

P18 - ROQUE DOS SANTOS Edmilson

Ergodic Basis Pursuit for network reconstruction

P19 - ROSSETTI Riccardo

An approximate message passing algorithm for the matrix tensor product model

P20 - SZEKELY Eszter

Data-driven separation between feature and lazy learners for higher-order statistics

P21 - TOMASINI Maria Umberto

How deep convolutional neural networks lose spatial information with training

P22 - VALDORA Silvia Marina

Robust estimators for functional logistic regression

P23 - VELASQUEZ VARELA Diego

Estimation of Tumor Benignity Probability using the 4C Method, DBSCAN, and Data Imputation Techniques with Cross-Validation

P24 - ZHANG Ebony Yuwen

Tensor Network Message Passing

P01

Far from Asymptopia

Michael C. Abbott, Benjamin B. Machta

Department of Physics, Yale University, New Haven, CT 06520, USA

Inference from limited data requires a notion of measure on parameter space, which is most explicit in the Bayesian framework as a prior distribution. Jeffreys prior is the best-known uninformative choice, the invariant volume element from information geometry, but we demonstrate here that this leads to enormous bias in typical high-dimensional models. This is because models found in science typically have an effective dimensionality of accessible behaviors much smaller than the number of microscopic parameters. Any measure which treats all of these parameters equally is far from uniform when projected onto the sub-space of relevant parameters, due to variations in the local co-volume of irrelevant directions. We present results on a principled choice of measure which avoids this issue and leads to unbiased posteriors by focusing on relevant parameters. This optimal prior depends on the quantity of data to be gathered, and approaches Jeffreys prior in the asymptotic limit. However, for typical models, this limit cannot be justified without an impossibly large increase in the quantity of data, exponential in the number of microscopic parameters. [1]

[1] M. C. Abbott, B. B. Macta, Entropy **25**, 434 (2023).

P02

Inference from Meteorological Parameter and Key Performance Indicator Data for Quality of Service Prediction in Mobile Communication Network

The poor Quality of Services (QoS) experience by mobile network users may be as a result of the network providers relying on the troposphere for signal transmission without first evaluating and characterizing the area where the signals are conveyed. Knowing which weather variables have an impact on signal propagation is necessary to define network vulnerability. This study examines how four mobile networks—MTN, Airtel, Globacom, and 9mobile—in South Western States of Nigeria (Ikeja, Ibadan, Osogbo, Akure, Ado-Ekiti and Abeokuta) respond to changes in relative humidity, wind speed, and temperature. Seven years data of weather variables collected from the Modern-Era Retrospective Analysis for Research and Applications (MERRA) and seven years key performance indicators: Call Setup Success Rate (CSSR), (DCR), Stand-alone Dedicated Channel (SDDCH) congestion, Traffic Channel congestion (TCCH) data obtained from the telecommunications regulatory body, Nigerian Communication Commission (NCC), was used in this study, both spanning from January 2016 to December 2021. Curated dataset was analyzed using infographics such scatter graphs. The effect of meteorological variables on key performance indicators were evaluated using statistical inference such as correlation and universality.

P03

Abstract for poster

L.Bardone[1] **and S.Goldt**[1]

[1] *SISSA*

Title: Beyond Gaussian models: a mathematical framework for the analysis of real-world data structures

Neural Network algorithms are very good at extracting information from real-world data. But what makes them perform so well where other algorithms fail? Recent works have linked the success of Neural Networks to their ability to easily access properties of the data distribution that are more complex than the mean vector and the covariance matrix (the high order cumulants-HOC).

To investigate the role of HOC for inference tasks, we consider a hypothesis testing problem that compares Gaussian noise with a non-Gaussian whitened distribution. The aim is to bound the sample complexity (SC), i.e. the number of samples necessary to understand which of the two distributions data was drawn from. In this non-Gaussian model the first two cumulants of the data give no information, which is in fact concentrated on HOC. Hence random matrix results, usually employed for similar Gaussian models, cannot be readily applied and new ways to get SC bounds need to be found. In the poster, we provide an overview on possible methods to achieve these estimates, including techniques based on the Low-Degree Likelihood Ratio.

P04

Quasi-universal scaling and near-critical behavior in neural systems

Guillermo B. Morales[1] **, Serena di Santo**[1] **, and Miguel A. Muñoz**[1]

[1] *Department of Electromagnetism and Condensed Matter, University of Granada*

The brain is in a state of perpetual reverberant neural activity, even in the absence of specific tasks or stimuli. Shedding light on the origin and functional significance of such a dynamical state is essential to understanding how the brain transmits, processes, and stores information. An inspiring, albeit controversial, conjecture proposes that some statistical characteristics of empirically observed neuronal activity can be understood by assuming that brain networks operate in a dynamical regime with features, including the emergence of scale invariance, resembling those seen typically near phase transitions. Here, we present a data-driven analysis based on simultaneous high-throughput recordings of the activity of thousands of individual neurons in various regions of the mouse brain. To analyze these data, we construct a unified theoretical framework that synergistically combines a phenomenological renormalization group approach and techniques that infer the general dynamical state of a neural population, while designing complementary tools. This strategy allows us to uncover strong signatures of scale invariance that are "quasiuniversal" across brain regions and experiments, revealing that all the analyzed areas operate, to a greater or lesser extent, near the edge of instability.

P05

Learning a restricted Boltzmann machine using biased Monte Carlo sampling

Nicolas Béreux[1], **Aurélien Decelle**[1,2], **Cyril Furtlehner**[2] **and Beatriz Seoane**[1]

[1] Departamento de Física Teórica, Universidad Complutense de Madrid, 28040 Madrid, Spain

[2]Université Paris-Saclay, CNRS, INRIA Tau team, LISN, 91190, Gif-sur-Yvette, France.

Restricted Boltzmann Machines are a generative model able to learn any complex distribution from a dataset. Its simple structure makes it particularly useful for interpretability and pattern extraction applications. RBMs, like other energy-based generative models, struggle to describe highly structured data, mainly because their training relies on costly Markov Chain Monte Carlo (MCMC) processes and the cost of sampling multimodal distributions is prohibitive. In particular, we observe that RBMs perform dramatically poorly on artificial low-dimensional clustered datasets. In our work [2], we investigate a biased sampling method named Tethered Monte Carlo (TMC)[1] to overcome this limitation. This method allows to properly sample such low dimensional datasets in a significantly shorter time, leading to a more accurate likelihood gradient during training, allowing the RBM to accurately learn such datasets. This method can also be used to retrieve the distribution learned by the RBM after training, allowing to assess the quality of the training. However, this method breaks the intra-layer independance of the RBMs which forbids the parallelisation of the MCMC updates, limiting the size of the model we can use.

[1] Martin-Mayor, Seoane and Yllanes, Tethered monte carlo: Managing rugged free- energy landscapes with a helmholtz-potential formalism, Journal of Statistical Physics **144**(3), 554 (2011)

[2] Béreux, Decelle, Furtlehner, Seoane, Learning a restricted Boltzmann machine using biased Monte Carlo sampling. SciPost Physics, **14(3)**, 032 (2023).

P06

Spectral Universality of Regularized Linear Regression with Nearly Deterministic Sensing Matrices

Rishabh Dudeja[1], **Subhabrata Sen**[1], **and Yue M. Lu**[1]

[1] *Harvard University*

Spectral universality refers to the empirical observation that asymptotic properties of a high-dimensional stochastic system driven by a structured random matrix are often determined only by the spectrum (or singular values) of the underlying matrix - the singular vectors are irrelevant provided they are sufficiently “generic”. Consequently, the properties of the underlying system can be accurately predicted by analyzing the system under the mathematically convenient assumption that the singular vectors as uniformly random (or Haar distributed) orthogonal matrices. This general phenomenon has been observed in numerous contexts, including statistical physics, communication systems, signal processing, statistics, and randomized numerical linear algebra. We study this universality phenomenon in the context of high-dimensional linear regression, where the goal is to estimate an unknown signal vector from noisy linear measurements specified using a sensing matrix. We prove a spectral universality principle for the performance of convex regularized least squares (RLS) estimators for this problem. Our contributions are two-fold: (1) We introduce a notion of a universality class for sensing matrices, defined through nearly deterministic conditions that fix the spectrum of the matrix and formalize the heuristic notion of generic singular vectors; (2) We show that for all sensing matrices in the same universality class, the dynamics of the proximal gradient algorithm for the regression problem, and the performance of RLS estimators themselves (under additional strong convexity conditions) are asymptotically identical. In addition to including i.i.d. Gaussian and rotational invariant matrices as special cases, our universality class also contains highly structured, strongly dependent, and even nearly deterministic matrices. Examples include randomly signed incoherent tight frames and randomly subsampled Hadamard transforms. Due to this universality result, the performance of RLS estimators on many structured sensing matrices with limited randomness can be characterized using the rotationally invariant sensing model with uniformly random (or Haar distributed) singular vectors as an equivalent yet mathematically tractable surrogate.

P07

Approximate Message Passing for elliptic random matrices and its applications to theoretical ecology

Mohammed-Younes GUEDDARI[1], **Walid HACHEM**[1], **Jamal NAJIM**[1]

[1] *CNRS and Université Gustave Eiffel, France*
5 Bd Descartes, 77420 Champs-sur-Marne

The problem of finding the equilibrium point in a system of Lotka-Volterra type differential equations arises naturally when modeling the dynamics of interactions between species in an ecological environment. Motivated by this problem, this equilibrium point verifies a classical optimization problem called LCP (Linear Complementarity Problem) [1], but in an unusual setting, i.e. when the matrix considered is random and high-dimensional. To obtain the statistical properties of such an equilibrium point, we use a class of theoretically tractable iterative algorithms called AMP (Approximate Message Passing)[2]. These algorithms are already studied when the matrix is symmetric, in this work we are interested in the generalization of these algorithms for elliptical matrices, this choice making the species interaction model more realistic.

[1] Imane Akjouj et al. *Equilibria of large random Lotka-Volterra systems with vanishing species: a mathematical approach*. 2023. arXiv: `2302.07820 [q-bio.PE]`.

[2] Oliver Y. Feng et al. *A unifying tutorial on Approximate Message Passing*. 2021. arXiv: `2105.02180 [math.ST]`.

P08

Deep Learning with No-arbitrage Constraints

Kentaro Hoshisashi[1,2], **Carolyn E. Phelan**[1], **and Paolo Barucca**[1]

[1]*Department of Computer Science, University College London, United Kingdom*
[2]*Sumitomo Mitsui Banking Corporation, Tokyo, Japan*

This study investigates the calibration method of option pricing models using artificial neural networks to adapt to market data. In calibration from limited and structured market data, it is crucial to represent entire option prices under no-arbitrage conditions for the valuation of other financial instruments [1, 2].

In this study, we develop a deep learning architecture that satisfies no-arbitrage conditions and provides useful analytical functions through the constructed networks, such as sensitivity analysis and efficient calibration. The applied network is structured as a general feed-forward architecture that enforces no-arbitrage conditions via a loss function with penalty terms as soft constraints in learning, which are specified in the experimental design parameters for optimal performance. We also demonstrate that the developed architecture enhances computational efficiency with sufficient price accuracy and provides appropriate interpolation while considering no-arbitrage constraints in sparse data.

[1] P. Carr, D. Madan, B. Finance Res. Lett. **2(3)**, 125-130 (2005).
[2] D. Ackerer, N. Tagasovska, and T. Vatter. NeurIPS **33**, 11552-11563 (2020).

P09

APPLICATION OF ARTIFICIAL INTELLIGENCE (AI) AND COLLECTIVE INTELLIGENCE (CI) FOR DIAGNOSIS OF BREAST CANCER; CASE STUDY OF AFRICAN WOMEN

Jimoh Abdulganiyu[1]
Mohammed VI Polytechnic University
School of Collective Intelligence
Ben Guerir, Morocco

Adeniran Akeem[2]
Mohammed VI Polytechnic University
School of Collective Intelligence
Ben Guerir, Morocco

Klein Mark[3]
Massachusetts Institute of Technology
Center for Collective Intelligence
Cambridge, Massachusetts, United States

Abstract

Background: Breast cancer claimed the lives of 74,072 people in Africa in 2018, with an estimated 168,690 cases [1]. Between 2004 and 2008, the annual increase in breast cancer incidence rate among Moroccan women was 2.85%, causing the age-standardized rate to rise from 35.0 to 39.0 cases per 100,000 women worldwide [2]. Breast cancer accounts for 27.7% of total cancer cases in African nations, followed by cervical cancer, which accounts for 19.6% of total cases. When taken as a whole, this is the most prevalent in African women [3]. **Methods:** This study aims to investigate whether Collective Intelligence (CI) can outperform AI in classifying breast cancer cases as benign or malignant for early detection, hypothesizing that CI will demonstrate higher accuracy than AI. The adoption of convolutional neural networks (CNN) and DenseNet transfer learning feature extraction to classify mammogram images as either benign or malignant, with the aim of improving early detection of breast cancer. The dataset was preprocessed and augmented before being fed into the network, and the resulting classification was validated and tested for accuracy. The findings indicate that this approach can be an effective tool for detecting breast cancer and contributing to early diagnosis, potentially leading to better treatment outcomes. Additionally, an online survey was conducted using breast mammograms that had been classified by AI algorithms. Expert raters were recruited to further classify these images using their analytic skills in order to make effective decisions regarding the presence of malignancies. The findings also suggest that this approach can be a valuable tool in improving the accuracy of breast cancer diagnosis. **Results and conclusion:** Our initial experiment average metrics score results yielded accuracy 90%, precision 92%, recall 0.91%, and f1-score 91% prompting us to conduct a second laboratory experiment. Through this experiment, we aim to discover novel insights that can benefit radiologists, oncologists, pathologists, and breast cancer researchers, as well as healthcare stakeholders. By leveraging the power of AI and CI, our research aims to improve cancer diagnosis in Africa's at early stage detection, with the goal of reducing false positives and false negatives by 100%.

References

[1] Sharma, R. (2021). Breast cancer burden in Africa: evidence from GLOBOCAN 2018. Journal of Public Health, **43**(4), 763-771.

[2] Khalis, M., El Rhazi, K., Charaka, H., Chajès, V., Rinaldi, S., Nejjari, C., ... & Charbotel, B. (2016). Female breast cancer incidence and mortality in Morocco: comparison with other countries. Asian Pacific Journal of Cancer Prevention: APJCP, **17**(12), 5211.

[3] Bahnassy, A. A., Abdellateif, M. S., & Zekri, A. R. N. (2020). Cancer in Africa: is it a genetic or environmental health problem? Frontiers in Oncology, **10**, 604214.

P10

Generalized Linear Models: Multi-Array Valued Predictors

Kapla D.[1], **Bura E.**[1]

[1] *TU Wien, Vienna, Austria*

We consider regression and classification for *general* response Y and tensor-valued predictors $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ (multi dimensional arrays) and propose a *novel formulation* for sufficient dimension reduction. Assuming the distribution of the tensor-valued predictors given the response $\mathbf{X} \mid Y$ is in the quadratic exponential family, we model the natural parameter as a multi-linear function of the response. This allows per-axis reductions that drastically reduce the total number of parameters for higher order tensor-valued predictors. We derive maximum likelihood estimates for the sufficient dimension reduction and a computationally efficient estimation algorithm which leverages the tensor structure. The performance of the method is illustrated via simulations and real world examples are provided.

P11

UNIVERSAL INVESTMENT STRATEGIES BY ONLINE LEARNING UNDER LATENT INFORMATION-DEPENDENCE STRUCTURE

Duy Khanh Lam[1], Daniele Giachini[2], and Giulio Bottazzi[2]

[1] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
[2] Institute of Economics & EMbeDS Department, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56127 Pisa, Italy

This study investigates the growth of self-financing investment strategies that comprise portfolios of several stocks subject to short-selling constraints over a long-term horizon. We derive universal strategies in an online fashion that rely solely on gradually realized data in a stochastic stock market with an unknown information-dependence structure. Initially, we model the latent information-dependence structure of the market under the efficient market hypothesis, which posits that the dynamics of daily stock prices immediately reflect all observable and unobservable side information other than the prices of the stocks being considered. We subsequently relax the efficiency hypothesis to account for the persistence of historical side information and prices on future prices. We propose a universal strategy that can attain the optimal growth rate even without knowledge of the information-dependence structure of the market. Additionally, our analysis reveals that inadequate knowledge of the information-dependence structure and unobservable side information can undermine the growth rate of prediction-based universal strategies. Accordingly, we discuss several methods to improve the performance of these strategies.

P12

Random Matrix Methods for Machine Learning: "Lossless" Compression of Large and Deep Neural Networks

A data-driven systematic, consistent and non-exhaustive approach to Model Selection

Diego Marcondes[1]

[1] *Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Brazil*

Modern science consists on conceiving a set of hypotheses to explain observable phenomena, confronting them with reality, and keeping as possible explanations only hypotheses which have not yet been falsified. Such a set of hypotheses is called a model, hence an important step of the scientific method is to select a model. Under a Statistical Learning framework, this consists on selecting a model among candidates based on quantitative evidence, and then learning hypotheses on it by the minimization of an empirical risk function. The need to select a model, rather than considering the union of the candidates as the possible hypotheses, is the liability to *overfitting*, from which arises a complexity-bias trade-off. If we choose a highly complex model, then we may have in it hypotheses which explain the underlying process very well, but there may also be hypotheses which explain the empirical data very well, and it is not clear how to separate them, so we overfit the data. If we choose a simpler model, it may happen that the hypotheses which well fit empirical data are the same that better explain the process, but may not explain it very well, as there may be hypotheses not in the model which are better, so there is a bias when learning on this model. Therefore, properly choosing the model is an important part of the solution of a learning problem, and is performed via Model Selection. This work proposes a data-driven systematic, consistent and non-exhaustive approach to Model Selection. The main feature of the approach are the collections of candidate models, which we call Learning Spaces, that, when seen as a set partially ordered by inclusion, may have a rich structure which enhance the quality of learning. The approach is data-driven since the only features which are chosen are the Learning Space and risk function, so all other features are based on data. It is systematic since it is constituted of a formal system of two steps: select a model from the Learning Space and then learn hypotheses on it. From a statistical point of view, there is a target model among the candidates, which is that with the lowest bias and complexity, and the approach is consistent since, as the sample size increases, the selected model converges to the target with probability one, and the estimation errors related to the learning of hypotheses on it converge in probability to zero. We establish U-curve properties of the Learning Spaces which imply the existence of U-curve algorithms that can optimally estimate the target model without an exhaustive search, which can also be efficiently implemented to obtain suboptimal solutions. The main implication of the approach are instances in which the lack of data may be mitigated by high computational power, a property which may be behind the high-performance computing demanding modern learning methods. We illustrate the approach on simulated and real data to learn on the important Partition Lattice Learning Space, to forecast binary sequences under a Markov Chain framework, to learn multilayer W-operators, and to filter binary images via the learning of interval Boolean functions.

Correction of overfitting bias in ML regression

E. Massa[1], **M. Jonker**[2], **K. Roes**[2] **and A.C.C. Coolen**[1,3]

[1]*(Presenting author underlined) Radboud Universiteit, DCN, Neurophysics*
[2]*Affiliation 2* Radboud UMC, Health Evidence Department
[3] Saddle Point Science Europe

Regression analysis based on many covariates is becoming increasingly common. When the number of covariates p is of the same order as the number of observations n, statistical inference based on maximum likelihood estimation of regression and nuisance parameters of the assumed model becomes unreliable due to overfitting. This often leads to systematic biases in (some of) the estimators and their variance is larger than expected from classical results. It is then crucial to be able to correctly quantify these effects for inference and eventually prediction purposes. In the literature, several methods to overcome overfitting bias or to adjust estimates have been proposed. The vast majority of these focus on the regression parameters only. On the other hand, failure to correctly estimate also the nuisance parameters may lead to significant errors in outcome predictions, as these control the shape of the cumulative distribution function of the response.

In [1] we study the overfitting bias of maximum likelihood (ML) estimators for the regression and the nuisance parameters in parametric generalized linear regression models in the overfitting regime where $p = \zeta n$ ($\zeta < 1$). Via a jacknife methodology, we show that the overfitting biases of the ML estimators can be estimated by solving a small set of non-linear equations. We find that the nuisance parameters are generally affected by overfitting bias, even when the regression parameters are not (i.e for linear models), and we quantify the latter. This allows us to correct the ML estimators to obtain unbiased estimators, asymptotically. We show with simulations on some prototypical models that the resulting procedure is relatively straightforward for linear models, in particular for those with noise in the location scale scale family, but is in general more involved for non-linear models. When the true regression parameter vector $\boldsymbol{\beta}_0$ is diffuse, i.e. each component of order $1/\sqrt{p}$, previous studies agree that our methodology can be used to understand the variance of any component of the corresponding estimator ($\hat{\boldsymbol{\beta}}_n$). We find that when this is not true, i.e $\boldsymbol{\beta}_0$ is sparse, the variance of the component of $\hat{\boldsymbol{\beta}}_n$ along $\boldsymbol{\beta}_0$ is underestimated. We make sense of this empirical observation by means of the stochastic representation introduced by us [2]. On the other hand our theoretical results well describe the variance of the components that are orthogonal to $\boldsymbol{\beta}_0$, even in the sparse case. This gives an approximate asymptotic distribution for a Wald statistics which can be used to test the hypothesis that a component of $\boldsymbol{\beta}_0$ is actually zero.

[1] E.Massa, M.Jonker, K. Roes, A.C.C. Coolen, Arxiv 2204.05827, (2022) .

[2] E.Massa, M.Jonker, A.C.C. Coolen , J.Phys. A., 48, 55 , 485002, (2022) .

P15

Wavelet-based Clustering and Classification

Pedro A. Morettin[1], **Chang Chiann**[1], **João R. Sato**[2] **and Brani Vidakovic** [3]
[1] University of São Paulo, [2] Federal University of ABC and [3] Texas A & M University

In this work we consider several wavelet-based procedures for clustering and classification purposes. In some situations, the time domain approach may not lead to clear classification or discrimination. When we move to the wavelet domain, the multiresolution analysis leads to look at data in several levels of resolution (or scales) and then the separation may become better. Among the wavelet-based procedures, we mention:

(a) Multifractal Spectra (MFS) and associated descriptors.

(b) DWT-CEM procedure: discrete wavelet transform combined with classification expectation maximization algorithm.

(c) DWT-Schur measures: discrete wavelet transform followed by the use of some Schur monotone measure.

(d) Wavelet-based Bayesian discriminant function.

References

1. McQueen J (1967). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symposium on Math. Stat. and Prob. 1:281-296.

2. Sato, J.R., Fujita, A., Amaro Jr, E., Mouraão Miranda, J., Morettin, P.A. and Brammer, M.J. (2007). DWT-CEM: An algorithm for scale-temporal clustering in fMRI. Biological Cybernetics, 97, 33-45.

3. Vidakovic, B. (1999). Statistical Analysis by Wavelets. Wiley.

P16

Machine Learning Techniques for Earthquake Size and Magnitude Prediction in Turkey

P17

A robust criterion for selecting among fitted models

A. René[1,2,3], **and A. Longtin**[1,4]

[1]*Neurophysics and nonlinear dynamics group, University of Ottawa, Canada*
[2]*Department of Physics, RWTH Aachen, Germany*
[3]*Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany*
[4]*Brain and Mind Research institute, University of Ottawa, Canada*

As we learn to learn increasingly complex models from data, so too must the methods which deal with the inferred models become more sophisticated. In particular, for models with symmetries, near-degeneracies and almost-equivalent parameters, how does one distinguish a good fit from a bad one? Ultimately the goal of fitting a model to data is to learn something about those data – but to do that, we first need a reliable criterion for keeping only the most relevant models. Unfortunately standard model comparison methods cannot help us here: we need a criterion which can distinguish between models which all have the same form (i.e. equations), but differ in the values of their parameters. Our motivating use case are time series data, for which we have already demonstrated the ability to learn rich dynamical models by directly optimizing the parameters of the differential equation [1]. Our approach is not limited to data of this form, however they do make one key requirement for the criterion especially obvious: it must not be sensitive to the number of data points. After all, for such data the recording time and/or frequency are often arbitrary choices dictated by experimental convenience. And yet sensitivity to the data set size is a property of most methods which are part of current standard practice; using a simple example, we illustrate how this can lead to pathological *overconfidence* in one model over another. The usual solution – to use a fixed data set size – is unsatisfactory and in many cases untenable [2]. Instead, we propose a solution which is part ontological, part mathematical.

Ontologically, we argue that it is essential to distinguish the *physical model* – which is grounded in theory – from the *observation model.* We care less about the parameters of the observation model, since its role is to account for the inevitable mismatch between theory and data. In fact, we propose that the key to a robust comparison criterion is to forego parameterising the observation model altogether, and thus allow for "unknown unkowns". Instead, we compare parameter sets based on the expected likelihood on *unseen* data, averaged over the space of all possible observation models.

The mathematical part of our solution is a practical method for computing the expectation over unparameterised observation models. This is built upon three main ideas. The first is to cast expectations as path integrals in a space of 1d quantile functions. The second is then a method for sampling quantile paths, and thus computing the path integral numerically. The third is a calibration procedure, which allows to adjust the criterion's sensitivity to the specifics of the model and the data. We provide all necessary code as a Python library, making this we hope a practical and flexible tool for practitioners.

[1] A. René, A. Longtin, J. H. Macke, Neural. Comput. **32** 1448–1498 (2020).
[2] A. Gelman, Y. Yao, J. Phys. G: Nucl. Part. Phys. **48** 014002 (2020)

Ergodic Basis Pursuit for network reconstruction

Edmilson Roque dos Santos[1], **Sebastian van Strien**[2], **Tiago Pereira**[1,2]

[1] *Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, Brazil*

[2] *Department of Mathematics, Imperial College London, London, UK*

Reconstruction of the interaction structure from multivariate time series of networks is an important problem in multiple fields of science. This problem is ill-posed for large networks leading to the reconstruction of false interactions. We put forward the Ergodic Basis Pursuit (EBP) method that uses the network dynamics' statistical properties to ensure the exact reconstruction of sparse networks when a minimum length of time series is attained. We show that this minimum time series length scales quadratically with the node degree being probed and logarithmic with the network size. Our approach is robust against noise and allows us to treat the noise level as a parameter. We show the recovery power of the EBP in experimental multivariate time series from optoelectronic networks.

P19

An approximate message passing algorithm for the matrix tensor product model

Riccardo Rossetti[1], **Galen Reeves**[1,2]

[1] *Department of Statistical Science, Duke University*
[2] *Department of Electrical and Computer Engineering, Duke University*

We propose and analyze an approximate message passing (AMP) algorithm for the matrix tensor product model, which is a generalization of the standard spiked matrix models that allows for multiple types of pairwise observations over a collection of latent variables. A key innovation for this algorithm is a method for optimally weighing and combining multiple estimates in each iteration. Building upon the approach of Berthier et al. (2020) [1], we prove a state evolution for non-separable functions that provides an asymptotically exact description of its performance in the high-dimensional limit. We leverage this state evolution result to provide necessary and sufficient conditions for recovery of the signal of interest. Such conditions depend on the singular values of a linear operator derived from an appropriate generalization of a signal-to-noise ratio for our model. Our results recover as special cases a number of recently proposed methods for contextual models (e.g., covariate assisted clustering) as well as inhomogeneous noise models.

[1] Berthier R., Montanari A., Nguyen P, Inf. Inference J. IMA **9** 1, 33-79 (2020).

P20

Data-driven separation between feature and lazy learners for higher-order statistics

Eszter Székely[1], Federica Gerace[1], Lorenzo Bardone[1] and Sebastian Goldt[1]

[1] *Scuola Internazionale Superiore di Studi Avanzati, SISSA*

We compare shallow two-layer networks in the fully-trained and lazy learning regime on tasks that rely on the information contained in the higher-order statistics of the inputs. We design synthetic models of data where we control the relative importance of higher-order cumulants and study in which settings end-to-end trained networks achieve better performance than random features. We further study the features of data that neural networks fit.

P21

How deep convolutional neural networks lose spatial information with training

Umberto M. Tomasini[1,*], Leonardo Petrini[1,*], Francesco Cagnetta[1] and Matthieu Wyart[1]

[1] *Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL).*

A central question of machine learning is how deep nets manage to learn tasks in high dimensions. An appealing hypothesis is that they achieve this feat by building a representation of the data where information irrelevant to the task is lost. For image datasets, this view is supported by the observation that after (and not before) training, the neural representation becomes less and less sensitive to diffeomorphisms acting on images as the signal propagates through the net. This loss of sensitivity correlates with performance, and surprisingly correlates with a *gain* of sensitivity to white noise acquired during training. These facts are unexplained, and as we demonstrate still hold when white noise is added to the images of the training set. Here, we *(i)* show empirically for various architectures that stability to image diffeomorphisms is achieved by both spatial and channel pooling, *(ii)* introduce a model scale-detection task which reproduces our empirical observations on spatial pooling and *(iii)* compute analitically how the sensitivity to diffeomorphisms and noise scales with depth due to spatial pooling. The scalings are found to depend on the presence of strides in the net architecture. We find that the increased sensitivity to noise is due to the perturbing noise piling up during pooling, after being rectified by ReLU units.

[1] Tomasini, Petrini, Cagnetta, Wyart, ”How deep convolutional neural networks lose spatial information with training”, arXiv:2210.01506, accepted in the Workshop ”Physics4AI”, ICLR 2023.

P22

Robust estimators for functional logistic regression

Graciela Boente[1,2], **Marina Valdora**[1,2]

[1]*(Presenting author underlined) Universidad de Buenos Aires*
[2] *CONICET*

Functional data analysis provides tools for analysing data collected in the form of functions or curves which appear in fields such as chemometrics, image recognition and spectroscopy. Functional data are intrinsically infinite–dimensional and, as mentioned for instance in [5], this infinite–dimensional structure is indeed a source of information. For that reason, even when recorded at a finite grid of points, functional observations should be considered as random elements of some functional space rather than multivariate observations. In this way, some of the theoretical and numerical challenges posed by the high dimensionality may be solved. This framework has led to the extension of some classical multivariate analysis concepts, such as linear regression and logistic regression, to the context of functional data, usually through some regularization tool. An overview of different tools for analysing this type of data may be found in [4], see also [3].

The functional logistic regression model assumes that (y_i, X_i), $1 \leq i \leq n$ are independent observations such that $y_i \in \{0, 1\}$ and $X_i \in L_2(\mathscr{I})$ with $\mathscr{I}$ a compact interval and that the model relating the responses to the covariates is given by

$$P(y_i = 1|X_i) = \frac{\exp\{\alpha_0 + \langle X_i, \beta_0\rangle\}}{1 + \exp\{\alpha_0 + \langle X_i, \beta_0\rangle\}},$$

where $\alpha_0 \in R$, $\beta_0 \in L^2(\mathscr{I})$ and $\langle u, v\rangle = \int_{\mathscr{I}} u(t)v(t)dt$ is the usual inner product in $L_2(\mathscr{I})$.

As mentioned in [5], one of the challenges in functional regression is the inverse nature of the problem, which causes estimation problems mainly generated by the compactness of the covariance operator of X. The usual practice to solve this problem is regularization, which can be achieved in several ways, either reducing the set of candidates for estimating β_0 to those belonging to a finite–dimensional space spanned by some basis, such as spline, Fourier, or wavelet bases or by adding a penalty term as when considering $P-$splines or smoothing splines.

Taking into account the sensitivity of these estimators to atypical observations and based on the ideas given for euclidean covariates by [1] and [2], we define robust estimators of the intercept α_0 and the slope β_0 following a sieve approach combined with weighted $M-$estimators. Theoretical assurances regarding the consistency and convergence rates of our proposal will be presented. Besides, through the results of a numerical study, we will illustrate the sensitivity of the classical estimator and the stability of the proposed method.

[1] Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics*, **109**, 17-34.

[2] Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, **44**, 273-295.

[3] Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer, New York.

[4] Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, Springer, Berlin.

[5] Wang, J. L., Chiou, J. & Müller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

P23

Estimation of Tumor Benignity Probability using the 4C Method, DBSCAN, and Data Imputation Techniques with Cross-Validation

Diego Velásquez Varela [1]

[1] *(dvelasquev@eafit.edu.co)*

Missing data is a common challenge in various fields, impacting the accuracy and reliability of analyses. Data imputation methods aim to address this issue by estimating missing values based on available information. This project presents a novel approach to data imputation, leveraging the Correlation Connected Clusters (4C) method [1,2,3], which combines Principal Component Analysis (PCA) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The 4C method identifies patterns and correlations within datasets, making it a promising foundation for a new data imputation methodology [4,5].

Specifically, this study applies the proposed method to estimate the probability of a tumor being benign or malignant, a crucial task for early cancer detection and treatment. The proposed method's main advantage is its robustness against noise, preserving meaningful patterns and correlations while mitigating the impact of noise on imputation results [6]. The long-term objective is to develop a comprehensive data imputation methodology based on the 4C method, potentially outperforming traditional methods and advancing the field of high-dimensional data analysis.

[1] C. Böhm, K. Kailing, P. Kröger, A. Zimek, Computing Clusters of Correlation Connected Objects, Proc. 2004 ACM SIGMOD Int. Conf. Management of Data, 455-466 (2004).
[2] M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, LOF: Identifying Density-Based Local Outliers, SIGMOD Rec. 29, 93 (2000).
[3] E. P. Kastampolidis, P. G. Sarigiannidis, I. D. Moscholios, G. I. Papadimitriou, the 4C Method: Clustering Based on Correlations and Similarities, IEEE Trans. Knowl. Data Eng. 31, 1900 (2019).
[4] J. C. Bezdek, R. J. Hathaway, VAT: A Tool for Visual Assessment of (Cluster) Tendency, Proc. IEEE Intl. Joint Conf. Neural Netw. 3, 2225 (2002).
[5] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD Rec. 25, 103 (1996).
[6] S. Ramaswamy, R. Rastogi, K. Shim, Efficient Algorithms for Mining Outliers from Large Data Sets, SIGMOD Rec. 29, 427 (2000).

P24

Tensor Network Message Passing

Yijia Wang [1,2] **, Yuwen Zhang** [3] **, Feng Pan**[1] **and Pan Zhang** [1,4,5]

[1]CAS Key Laboratory for Theoretical Physics, Institute of Theoretical Physic,
Chinese Academy of Sciences, Beijing 100190, China
[2] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Physics and Astronomy, University College London, Gower Street ,London, WC1E6BT, UK
[4]School of Fundamental Physics and Mathematical Sciences, Hangzhou Institute for Advanced Study, UCAS,
Hangzhou 310024, China
[5]International Center for Theoretical Physics Asia-Pacific, Beijing/Hangzhou, China

When studying interacting systems, computing their statistical properties is a fundamental problem in various fields such as physics, applied mathematics, and machine learning. However, this task can be quite challenging due to the exponential growth of the state space as the system size increases. Many standard methods have significant weaknesses. For instance, message-passing algorithms can be inaccurate and even fail to converge due to short loops, while tensor network methods can have exponential computational complexity in large graphs due to long loops. In this work, we propose a new method called ``tensor network message passing." This approach allows us to compute local observables like marginal probabilities and correlations by combining the strengths of tensor networks in contracting small sub-graphs with many short loops and the strengths of message-passing methods in globally sparse graphs, thus addressing the crucial weaknesses of both approaches.