Book of Abstract

PRIMO Workshop 2023

PRIMO WORKSHOP 2023

Post-graduate Researchers in Inverse Problems, Machine Learning, and Optimization (PRIMO) research group

Department of Mathematics – University of Bari "Aldo Moro"
20 - 22 September 2023
https://primo2023.uniba.it
primoworkshop2023@gmail.com

Contents

About

The aim of the “PRIMO Workshop” series is to offer early-career researchers, that are working on topics within the broad research interests of the PRIMO Group, a stage to share their work and network with peers and future collaborators.

PRIMO Workshop 2023

The "PRIMO Workshop" offers early-career researchers working on topics within the broad research areas of imaging science, machine learning/deep learning and non linear optimization a stage to share their work and network with peers and future collaborators.

This event is strictly related to the recently founded PRIMO (Post-graduate Researchers in Inverse problems, Machine learning, and Optimization) Research Group. The group, which started its activities in Summer 2020, currently counts more than 60 members from different Italian and international istitutions.

Every young researcher in these areas is strongly encouraged to register and submit an abstract for an in-person presentation. The registration is free of charge (but mandatory) and includes coffee breaks and lunches.

The 3th edition of the "PRIMO Workshop", as part of the PRIMO ResearchGroup activities, is a three-day event taking in place in Bari (Italy) and organized by MIδAS (Mathematics In Data AnalysiS) Research Group. The event will take place on September 20-22, 2023 at the Department of Mathematics at UniBa.

Organizing committee

Nicoletta Del Buono
Flavia Esposito
Grazia Gargano
Laura Selicato
Gaetano Settembre

Supported by

Keynotes Speakers

Matrix Completion: Optimization Methods and Applications

Stefania Bellavia

Dipartimento di Ingegneria Industriale, Università di Firenze

stefania.bellavia@unifi.it

Matrix completion (MC) is an important technique which is aimed to recover a low-rank or nearly low-rank matrix from undersampled/incomplete data. Its application varies from wireless communications, recommendation systems, images inpainting, missing data imputation. In this talk we focus on the convex programming reformulation and discuss optimization approaches ranging from iterative methods based on the SVD decomposition to relaxed Interior Point methods. Computational results are reported with a special focus on the imputation of missing data in a real onshore wind farm. The data are organized into a matrix in a daily range and MC approaches are used to recover the missing data, showing that MC is a reliable and parameter-free tuning tool to impute missing data in these applications.

A Novel Computational Paradigm for approximation, data analysis and representation: the Scientific Machine Learning

Salvatore Cuomo

M.O.D.A.L. Laboratory, Università degli Studi di Napoli Federico II

salvatore.cuomo@unina.it

Nowadays in the age of the Internet of Data (IoD), methods and models for data analysis and representation play a key and crucial role in any application field. Numerical methods for data analysis and representation, have to be (re)designed and (re)thought by considering learning approaches. Classical Scientific Computing fruitfully intersects Machine learning to boost the accuracy and efficiency of numerical algorithms. In this talk, we present some applications related to the fascinating world of Artificial Intelligence research field oriented to Scientific Computing.

Presented results are in M.O.D.A.L.- Mathematical moedelling and Data Analysis Lobatory http://www.labdma.unina.it

References

[1] Cuomo, S., Cola, V. S. D., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F. Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next. Journal of Scientifc Computing, 2022

[2] Cuomo, S., Giampaolo, F., Izzo, S., Nitsch, C., Piccialli, F., Trombetti, C. A physics-informed learning approach to Bernoulli-type free boundary problems. Computers & Mathematics with Applications, 128, 34-43, 2022.

Nonnegative Matrix Factorization and Beyond

Nicolas Gillis

University of Mons

nicolas.gillis@umons.ac.be

Given a nonnegative matrix X and a factorization rank r, nonnegative matrix factorization (NMF) approximates the matrix X as the product of a nonnegative matrix W with r columns and a nonnegative matrix H with r rows; see [1]. NMF has become a standard linear dimensionality reduction technique in data mining and machine learning. In this talk, we first introduce NMF and show how it can be used as an interpretable unsupervised data analysis tool in various applications, including hyperspectral image unmixing, image feature extraction, and document classification. Then we discuss how NMFs can be computed, and also discuss the issue of non-uniqueness of NMF decompositions, also known as the identifiability issue, which is crucial in many applications. Finally, we present how we can go beyond NMF by considering non-linear and deep extensions which are useful in real-world applications and offer many venues for future research.

References

[1] Nicolas Gillis. *Nonnegative Matrix Factorization*, SIAM, Philadelphia, 2020.

Contributed Talks

Kalman filters for estimating hydraulic parameters

Marco Berardi

Consiglio Nazionale delle Ricerche

via F. De Blasio 5, 70132 Bari

marco.berardi@ba.irsa.cnr.it

F. Di Lena, R. Masciale, I. Portoghese, G. Passarella

Here we propose a simple approach for modelling the functioning of a real life cluster of interconnected infiltration basins: such a modelling is crucial for correctly managing artificial aquifer recharge (see [1]). We model and dynamically quantify the flow from thebottom of each basin, towards the outer walls of the cluster, and between adjacent basins: as a result of this modelling, saturated hydraulic conductivity parameters are estimated. This assessment is accomplished by an extended Kalman filter approach. Different versions of these filters are described (see for instance [2]), and their results in the case at hand are discussed.

References

[1] F. Di Lena, M. Berardi, R. Masciale, I. Portoghese, *Network dynamics for modelling artificial groundwater recharge by a cluster of infiltration basins*, Hydrological Processes, 37(5), e14876

[2] H. Medina, N. Romano, and G.B. Chirico, *Kalman filters for assimilating near-surface observations into the Richards equation–part 2: A dual filter approach for simultaneous retrieval of states and parameters*, Hydrology and Earth System Sciences, 18(7):2521–2541.

Dimensionality reduction techniques for the analysis of human microbiome data

Veronica Buttaro

Università degli Studi di Bari, Dipartimento di Informatica v.buttaro3@studenti.uniba.it

Gianvito Pio

Università degli Studi di Bari, Dipartimento di Informatica gianvito.pio@uniba.it

The human microbiome refers to the set of microorganisms living on the surface and within a human host. The analysis of human microbiome data is very important for diagnostics and preventive purposes. In fact, microbiome data is proving to be a rich source of promising biomarkers for several diseases, such as colerectal cancer, that is the focus of this study.

In recent years, the possible adoption of machine learning algorithms to study medical problems has become a trending topic. The availability of large amounts of data is a crucial factor that is bringing value, as well as optimizing time and cost of specific analyses. In this context, the goal of this work is to analyze data related to the human gut microbiome through machine learning methods, in order to build a predictive model for the diagnosis of colorectal cancer. One of the main issues raised by the analysis of microbiome data is that of the high number of features (high dimensionality), especially if compared to the relatively low amount of samples on which they are observed. This leads to the common problem known as the "curse of dimensionality".

In this study, we specifically focused on this issue, by constructing a supervised autoencoder for dimensionality reduction of gut microbiome data. The performed experiments demonstrated that the application of the proposed data preprocessing technique can bring benefits in terms of performances of the learned classifier. Moreover, the proposed technique proved to make the learned classification model less sensitive to data imbalance.

References

[1] Nielson T Baxter, Mack T 4th Ruffin, Mary Ann Rogers, and Patrick D Schloss. *Microbiota-based model improves the sensitivity of fecal im-munochemical test for detecting colonic lesions*, Genome Medicine,8(1):37, 2016

[2] Richa Bharti and Dominik G Grimm *Current challenges and best-practice protocols for microbiome analysis.*, Briefings in bioinformatics, 22(1):178–193, 2021

[3] Min Oh and Liqing Zhang *Deepgeni: Deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy.*, Scientific Reports, 13(1):4599, 2023

Nyström Kernel Quadratures

Antoine Chatalic

Via Dodecaneso 35, 16146 Genova GE antoine.chatalic@dibris.unige.it

Nicolas Schreuder, Ernesto De Vito, Lorenzo Rosasco

Via Dodecaneso 35, 16146 Genova GE

In this work we consider the problem of numerical integration, i.e. approximating integrals with respect to a target probability measure using only pointwise evaluations of the integrand. When the latter belongs to a reproducing kernel Hilbert space, the problem corresponds to building a sparse approximation of the kernel mean embedding of the target distribution. Direct applications include the efficient computation of maximum mean discrepancies between distributions and the design of efficient kernel-based tests. We focus on the setting where the target distribution is only known via a set of n i.i.d. observations, and propose an efficient procedure based on the Nyström method which exploits a small i.i.d. random subset of $M < n$ samples drawn either uniformly or using approximate leverage scores. Our main result is an upper bound on the approximation error of this procedure for both sampling strategies [1, 2]. It yields sufficient conditions on the subsample size to recover the standard (optimal) $n^{1/2}$ rate while reducing drastically the number of functions evaluations, and thus computational costs. We illustrate our theoretical findings with numerical experiments.

References

[1] Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, Alessandro Rudi. *Nyström Kernel Mean Embeddings.*, Proceedings of the 39th International Conference on Machine Learning (2022)

[2] Antoine Chatalic, Nicolas Schreuder, Ernesto De Vito, Lorenzo Rosasco *(Work in progress, not yet published)*

Improving Classification Trustworthiness in Random Forests

Maria Stella de Biase

Dipartimento di Matematica e Fisica, Università della Campania "L. Vanvitelli", Viale Lincoln 5,

Caserta, Italy mariastella.debiase@unicampania.it

Stefano Marrone, Fiammetta Marulli, Laura Verde

Dipartimento di Matematica e Fisica, Università della Campania "L. Vanvitelli", Viale Lincoln 5,

Caserta, Italy {stefano.marrone, fiammetta.marulli, laura.verde}@unicampania.it

Critical applications of Machine Learning (ML) techniques are more and more widespread in the scientific and industrial communities. This diffusion is starting to become a critical aspect of new software-intensive applications due to the need for rapid reactions to both temporary and permanent changes in data. Ensemble methods is a research line in Artificial Intelligence (AI) explored by researchers for more than twenty years: the Random Forest (RF) technique is one of the most famous ML classification and regression approaches [1]. Notwithstanding the progresses in ensemble methods, some problems still affect RFs [2, 3, 4]: the strong dependency on input data and the tendency of RFs to overfit. To cope with these scenarios, this work investigates the improvement of reliability in ML based classification by extending RF with Bayesian Network (BN) models. Concretely, the paper introduces the Reputation Oriented Random Forest (RORF) formalism, a modified version of RFs where some BN nodes are added to manage the reputation of each single DT. The domain chosen to demonstrate the feasibility of the approach is the healthcare domain. The example is taken from an open-source dataset hosted by the well-known Kaggle portal[1]. The dataset reports 11 clinical features that can be used to predict the probability of stroke in people of all ages and both genders [5].

References

[1] Breiman, L., *Random forests*, Machine Learning, 2001

[2] Feng, W. and Ma, C. and Zhao, G. and Zhang, R., *FSRF:An Improved Random Forest for Classification*, Geocarto International, 2015

[3] Adelabu, S., Mutanga, O., and Adam, E., *Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods*, Proceedings of 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, AEECA 2020

[4] Angluin, D. and Laird, P., *Learning From Noisy Examples*, Machine Learning

[5] *Stroke prediction dataset*, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[1] http://www.kaggle.com

Machine Learning Algorithms Comparison: An application on Amino Acids Volume Prediction.

Roberta De Fazio, Rosy Di Giovannantonio, Stefano Marrone

Dipartimento di Matematica e Fisica, Università degli Studi della Campania Luigi Vanvitelli

roberta.defazio,stefano.marrone@unicampania.it

Emanuele Bellini

Dipartimento di Studi Umanistici, Università degli Studi Roma Tre emanuele.bellini@uniroma3.it

The application of Machine Learning (ML) algorithms in real-life problems is gaining popularity, allowing its usage in several different contexts from healthcare to railway. However, it requires a well-known and structured knowledge of the domain to enhance the exploration of the results in terms of explainability. In particular, this work aims to predict the volume of the amino acids in proteins' primary structure, taking into account some features that characterize the protein itself. .

To pursue this goal, we propose an implementation of a mathematical model [1, 2], that describes the protein from a geometrical point of view, starting from raw data. The data are open-source and extracted from Protein Data Bank (PDB)[2], considering 446 myoglobins structures. The proposed methodology requires an initial qualitative analysis of the PDB to define the subset of features, that will be the subject of the study. Then, a great effort is spent in the extraction and population processes oriented to import data from PDB files into the new structured Database (DBR^2). In particular, the extracted data are manipulated to infer geometrical features defined by the mathematical model. Subsequently, the analysis is carried out by comparing two ML algorithms: Random Forest (RF) and Multi-layer Perceptrons (MLP). The two different amino acid volume predictors are trained on the feature set stored in DBR^2. Finally, a widespread tool, ELI-5 [3, 4, 5], is used to extract the most important features from the not explainable model trained - i.e. MLP -, in order to compare the results with those obtained from the explainable one - i.e. RF.

A more exhaustive description of the proposed work is published in [6].

References

[1] Vitale, Federica *On statistically meaningful geometric properties of digital three-dimensional structures of proteins* , Mathematical and Computer Modelling, 2008

[2] Vitale, Federica *A topology for the space of protein chains and a notion of local statistical stability for their three-dimensional structures* , Mathematical and Computer Modelling, 2008

[3] Kuzlu, Murat and Cali, Umit and Sharma, Vinayak and Güler, Özgür *Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools*, IEEE Access, 2020

[4] Sarp, Salih and Knzlu, Murat and Cali, Umit and Elma, Onur and Guler, Ozgur *An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool*, IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2021

[5] Gilpin, Leilani H. and Bau, David and Yuan, Ben Z. and Bajwa, Ayesha and Specter, Michael and Kagal, Lalana *Explaining explanations: An overview of interpretability of machine learning*, Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018

[6] De Fazio, Roberta and Di Giovannantonio, Rosy and Bellini, Emanuele and Marrone, Stefano *Explainabilty Comparison between Random Forests and Neural Networks: Case Study of Amino Acid Volume Prediction* , Information, 2023

[2] https://www.rcsb.org/

Training of stable neural ordinary differential equations

Arturo De Marinis

arturo.demarinis@gssi.it

Nicola Guglielmi, Francesco Tudisco, Anton Savostianov

GSSI - Gran Sasso Science Institute, Viale Francesco Crispi 7, L'Aquila, Italy

Neural ODEs [1] constitute a fascinating link between differential equations and neural networks. There is a wide well-consolidated theory both on qualitative properties and numerical stability properties for ODEs, while most of what is known about neural networks and machine learning is based on heuristics. An important aspect of this connection is the possibility to employ ODE theory to tackle a few challenging problems of neural networks, such as their vulnerability to adversarial attacks, i.e. imperceptible perturbations, added to the inputs of a neural network, designed in such a way that the output corresponding to the perturbed input is far away from the output corresponding to the original input [2, 3]. The analysis of stability and contractivity of a neural ODE can be efficiently used to the aim of making a neural ODE robust and stable against adversarial attacks.

Our contribution is in this direction. Specifically, given a neural ODE whose vector field is a one-layer neural network, we propose an additional step to the state-of-the-art training strategy that makes the neural ODE *contractive*, that means it does not amplify the error in the input data to the output data, and therefore robust and stable against adversarial attacks.

We consider the neural ODE

$$\dot{x}(t) = \sigma(Ax(t) + b), \qquad t \in [0, T],$$

where $x : [0, T] \to \mathbb{R}^n$ is the feature vector evolution function, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ are the parameters, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, assumed to be absolutely continuous and such that $\sigma'(\mathbb{R}) \subset [m, 1]$, with $0 < m \leq 1$.

Then, we notice that the neural ODE is contractive if

$$\sup_{D\in\Omega_m} \mu_2 (DA) \leq 0, \tag{1}$$

where $\Omega_m = \{D \in \mathbb{R}^{n\times n} \ : \ D \text{ is diagonal and } m \leq D_{ii} \leq 1, \ \forall i = 1, \dots, n\}$, and μ_2 denotes the logarithmic 2-norm of a matrix, and we perform a transformation on the matrix A, after each step of gradient descent, that assures condition (1) is satisfied.

To illustrate our methodology, we compare the performance of two neural ODEs for MNIST classification against the Fast Gradient Sign Method (FGSM) attack: the former trained according to the state-of-the-art training strategy, and the latter trained according to our proposed strategy. Our experiments indicate that the latter shows a significant improvement in robustness against the FGSM attack.

References

[1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt and David K. Duvenaud, *Neural Ordinary Differential Equations*, Advances in Neural Information Processing Systems 31, 2018.

[2] Joan Bruna, Christian Szegedy, Ilya Sutskever, Ian J. Goodfellow, Wojciech Zaremba, Rob Fergus and Dumitru Erhan, *Intriguing Properties of Neural Networks*, International Conference on Learning Representations, 2014.

[3] Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, International Conference on Learning Representations, 2015.

A phase unwrapping technique based on Approximated Iterative QLP decomposition

Antonella Falini

Computer Science Department, University of Bari, Via E. Orabona 4, 70125 Bari, Italy

antonella.falini@uniba.it

Francesca Mazzia

Computer Science Department, University of Bari, Via E. Orabona 4, 70125 Bari, Italy

francesca.mazzia@uniba.it

Using radar interferometry from Earth-orbiting satellites is possible to monitor geophysical-processes and urban deformations. The provided phase information is generally obtained in modulus 2π radians, therefore, standard *phase unwrapping* techniques look for jumps, within an established tolerance-parameter, in the given wrapped signal and add multiples of 2π between those consecutive elements of the time-series that are greater than the jump tolerance. However, the presence of high noise introduces spurious phase jumps which make this unwrapping technique unsatisfactory.

In the present talk we apply the algorithm presented in [3] in combination with a matrix factorization algorithm based on the iterative Stewart's QLP decomposition [1, 2] that constructs an approximate truncated SVD.

In particular, provided a given threshold, only an automatically selected subspace is used to approximate the input wrapped signal. Once the noise has been sufficiently removed, a classical regression model [3] can be applied to estimate the cleaned wrapped signal and therefore, standard techniques like e.g., [4] deliver the correct unwrapped phase.

References

[1] GW. Stewart, *The QLP approximation to the singular value decomposition*, SIAM, Journal on Scientific Computing, 20 (1999), pp. 1336–1348.

[2] D.A. Huckaby and T.F. Chan, *On the convergence of Stewart's QLP algorithm for approximating the SVD*, Numerical Algorithms, 32 (2003), pp. 287–316.

[3] H. Barkhuijsen, R. De Beer, & D. Van Ormondt, *Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals*, Journal of Magnetic Resonance 73 (1969), pp. 553–557.

[4] https://it.mathworks.com/help/matlab/ref/unwrap.html

Computing closest singular matrix-valued functions

Miryam Gnazzo

Gran Sasso Science Institute, L'Aquila (Italy) miryam.gnazzo@gssi.it

Nicola Guglielmi

Gran Sasso Science Institute, L'Aquila (Italy) nicola.guglielmi@gssi.it

Given a set of matrices $A_i \in \mathbb{C}^{n\times n}$ and a set of analytic functions $f_i : \mathbb{C} \mapsto \mathbb{C}$, we consider a regular matrix-valued function $\mathcal{F}(\lambda) = \sum_{i=0}^{d} f_i(\lambda)A_i$, that is $\det(\mathcal{F}(\lambda))$ is not identically zero for $\lambda \in \mathbb{C}$. An interesting problem consists in the computation of the nearest singular function $\widetilde{\mathcal{F}}(\lambda) = \sum_{i=0}^{d} f_i(\lambda)(A_i + \Delta A_i)$, with respect to the Frobenius norm. For example, this problem may arise when $\det(\mathcal{F}(\lambda))$ represents the characteristic equation of a system of differential algebraic equations, to robustly guarantee its well-posedness. This class of considered matrix-valued functions includes not only matrix polynomials, but also matrix functions arising in delay differential equations, such as $D(\lambda) = A_0 + \lambda A_1 + \sum_{i=2}^{d} e^{-\tau_i\lambda}A_i$. Therefore, in the case of more general entire functions $f_i(\lambda)$, the possible presence of an infinite number of roots of $\det(\mathcal{F}(\lambda))$ represents a delicate feature of the problem and requires an appropriate analysis in the construction of the numerical method for the computation of the distance to singularity.

The idea consists in rephrasing the matrix nearness problem for the matrix-valued function into an equivalent optimization problem. Nevertheless this problem turns out to be highly non-convex. We propose a two level procedure, following the idea introduced in [1], and impose that the determinant vanishes on a finite set of prescribed complex points $\{\mu_j\}_{j=1}^{m}$. This can be translated into the minimization of the functional

$$F_\varepsilon(\Delta A_0, \ldots, \Delta A_d) = \frac{1}{2}\sum_{i=0}^{d} \sigma_{\min}^2\left(\widetilde{\mathcal{F}}(\lambda)\right),$$

where $\sigma_{\min}$ denotes the smallest singular value and $\varepsilon = \|[\Delta A_0, \ldots, \Delta A_d]\|_F$ is the norm of the perturbation, and in finding the smallest value ε for which the functional F_ε vanishes.

This approach can be generalized to situations where the matrix-valued functions present a certain structure. For instance, we can include in the method additional constraints, such as a certain sparsity pattern determined by the original matrices or even structures involving the whole functions, like palindromic properties.

References

[1] M. Gnazzo, N. Guglielmi. *Computing the closest singular matrix polynomial*, arXiv preprint arXiv:2301.06335, (2023).

[2] M. Gnazzo, N. Guglielmi. *On the numerical approximation of the distance to singularity for matrix-valued functions*, In preparation, (2023).

Efficient kernel methods for statistical hypothesis testing

Marco Letizia[1]

[1]MaLGa Center - Università di Genova, Genova, Italy marco.letizia@edu.unige.it

Gaia Grosso[2], Maurizio Pierini[3], Lorenzo Rosasco[1], Andrea Wulzer[4], Marco Zanetti[5]

[2]IAIFI, Boston MA, USA [3] CERN, Geneva, Switzerland [4] IFAE-BIST, Barcelona, Spain

[5]Università di Padova, Padova, Italy

Statistical hypothesis testing is a fundamental tool in data analysis, especially in the context of natural and life sciences. Most strategies are however based on simple one dimensional tests, accompanied by various recipes for dealing with correlated multivariate data. In this talk I will present the New Physics Learning Machine approach to hypothesis testing, a machine learning-based multivariate strategy to detect data departures from a reference model, while remaining agnostic about the potential source of discrepancy. Initially developed for searching anomalies in particle physics data, this approach has a larger reach, including the validation of generative models and data quality monitoring. Our implementation is based on an efficient large-scale implementation of kernel methods that leverages several statistical and computational ideas combining optimization, numerical linear algebra and random projections. The resulting model is fast, scalable and statistically sound.

References

[1] M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini et al., *Learning new physics efficiently with nonparametric methods*, https://doi.org/10.1140/epjc/s10052-022-10830-y*Eur. Phys. J. C* **82** (2022) 879 [https://arxiv.org/abs/2204.023172204.02317].

[2] G. Grosso, N. Lai, M. Letizia, J. Pazzini, M. Rando, L. Rosasco, A. Wulzer and M. Zanetti, *Fast kernel methods for Data Quality Monitoring as a goodness-of-fit test*, Accepted in https://iopscience.iop.org/article/10.1088/2632-2153/acebb7 *Machine Learning: Science and Technology* [https://arxiv.org/abs/2303.054132303.05413].

[3] G. Grosso, M. Letizia, M. Pierini and A. Wulzer, *Goodness of fit by Neyman-Pearson testing*, https://arxiv.org/abs/2305.141372305.14137.

3d Nesting for Autologous Ear Reconstruction: a global optimization approach.

Greta Malaspina

Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze greta.malaspina@unifi.it

Stefania Bellavia, Michaela Servi

Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze

stefania.bellavia@unifi.it, michaela.servi@unifi.it

Sara Nozzoli

Dipartimento di Matematica e Informatica, Università degli Studi di Firenze

sara.nozzoli@stud.unifi.it

Autologous Ear Reconstruction, is a surgical procedure where the external ear of a patient is reconstructed by using tissue from the patient himself. Typically, the main elements of the ear are carved from a graft of rib cartilage, then used to create the framework of the new ear. The shape of the elements to be carved makes this procedure particularly complex, so several instruments have been developed to ease the work of the surgeon and try to partially automate the procedure. One of the main issues is to decide where, in the rib, each of the ear elements should be carved, taking into account both the fact that the size of the rib graft should be as small as possible, and that carving the elements along the border of the rib makes the procedure easier for the surgeon and increases the quality of the reconstruction. This gives rise to what is typically referred to as a nesting problem: given a container and a set of objects, we want to find the optimal location of the objects inside the container.

We formulate the problem as a constrained, finite-sum minimization problem. The problem is low dimensional, however the objective function is expensive to evaluate, and its derivatives are not easily available. Moreover, because of the form of the objective function and the topology of the container, the problem is highly non-convex and presents several local minima. To enforce the belonging of the object to the container we employ a penalty strategy. We solve the resulting unconstrained method using CBO, a particle-based method for global optimization, combined with a mini-batch strategy. We present numerical results that show the effectiveness of the algorithm, and we study the performance of the method for different choices of the parameters and different penalty and sampling strategies.

Learning from data with via overparametrization

Cesare Molinari

MaLGa, DIMA, Università degli Studi di Genova; cecio.molinari@gmail.com

Cristian Vega, Lorenzo Rosasco, Silvia Villa

The goal of machine learning is to achieve a good prediction exploiting training data and some a-priori information about the model. The most common methods to achieve the last objective are explicit and implicit regularization. In the first technique, a regularizer is explicitly introduced to find, among all the solutions, a good generalizing one. The second technique, i.e. implicit regularization, is based on the inductive bias intrinsically induced by the specific method used to optimize the parameters involved.

Recently, the success of learning is related to over- and re-parametrization, that are widely used - for instance - in neural networks applications and the optimization method used. However, there is still an open question of how to find systematically what is the inductive bias hidden behind the model for a particular optimization scheme. In this talk, we take a step in this direction, studying extensively many reparametrization used in the state of the art, providing a common structure to analyze the problem in a unified way. We show that, gradient descent on the empirical loss for many reparametrization is equivalent to mirror descent on the original problem. The mirror function depends on the reparametrization and introduces an inductive bias, which plays the role of the regularizer. Our theoretical results provide asymptotic behavior and convergence results in the simplified setting of linear models.

Quantification of granular sparkling at echocardiography in patients with transthyretin-related cardiac amyloidosis using radiomic and mathematical morphology features

Noemi Montobbio[1,*]

[1]Department of Health Sciences (DISSAL), University of Genoa noemi.montobbio@edu.unige.it

Sara Mori[2,*], Cristina Campi[3], Giulia Elena Mandoli[4], Matteo Cameli[4], Maria Pia Sormani[1], Marco Canepa[2]

[2]DIMI & [3]DIMA, University of Genoa [4]DBM, University of Siena [*]equal contribution

Granular sparkling is a well-known echocardiographic feature found in patients with transthyretin-related cardiac amyloidosis (ATTR-CA). However, there is no objective technique for quantifying this feature, which therefore remains a qualitative, elusive and ultimately unreliable imaging characteristic. The aim of this study is to statistically and geometrically characterise granular sparkling as a volume-independent texture property of the myocardium in patients with ATTR-CA. We collected echocardiogram video-clips in parasternal long axis and 4-chamber view of 58 patients with ATTR-CA, and 60 age- and gender-matched patients without any known cardiac disease except hypertension, followed at San Martino Hospital, Genoa, and Le Scotte Hospital, Siena. For each video-clip, one end-diastole frame was extracted and annotated by an expert to identify a region-of-interest (ROI) within the interventricular septum (IVS). Left ventricle chamber masks were also extracted, and used as a brightness reference to enforce invariance w.r.t. the settings of the specific ultrasound system. Many established radiomic textural features are also heavily volume-confounded. We analysed the ROI texture by extracting a subset of volume-invariant radiomic features, as well as morphological granulometry features. We then fitted a support vector machine (SVM) classifier to discriminate between ATTR-CA and controls based on the computed textural features. The texture-based classifier predicted the diagnosis with a cross-validated accuracy of 80%. Moreover, mathematical morphology analyses revealed a significantly smaller grain size in the IVS tissue of patients with ATTR-CA as compared to controls (mean±SEM: 0.152 ± 0.007 vs. 0.184 ± 0.007 cm, $p = 0.0012$). Our results confirm the presence of morphological differences in the IVS tissue between ATTR-CA patients and healthy controls detectable from echocardiography, with a more distinctly granular texture associated with cardiac amyloidosis.

Online identification and control of PDEs via Reinforcement Learning methods

Agnese Pacifico

Sapienza University of Rome agnese.pacifico@uniroma1.it

Alessandro Alla, Michele Palladino, Andrea Pesare

alessandro.alla@unive.it, michele.palladino@univaq.it, andreapesare1@gmail.com

We focus on the control of unknown Partial Differential Equations. We propose an algorithm based on the idea to control and identify on the fly the unknown system configuration. We assume that, for a given control input, we are able to observe the true system evolution without its knowledge as typically done in Reinforcement learning methods (see e.g. [2]). In this talk ([1]), the control is based on the State-Dependent Riccati approach ([3]), whereas the identification of the model on Bayesian linear regression (see e.g. [4]).

The workflow of the proposed method can be summarised as follows:

1. Pick a parameter configuration,
2. Compute the corresponding control,
3. Observe the trajectories,
4. Update the parameter configuration based on the observations,
5. Go to the second step.

At each iteration, we obtain an estimate of the *a-priori* unknown parameter configuration of the PDE based on the observed data and then we compute the control of the corresponding model. We show by numerical evidence the convergence of the method for infinite horizon control problems.

References

[1] A. Alla, A. Pacifico, M. Palladino, A. Pesare *Online identification and control of PDEs via reinforcement learning methods*, in preparation

[2] R. S. Sutton, A. G. Barto *Reinforcement Learning: An Introduction* The MIT Press, 2018

[3] H. T. Banks, B. M. Lewis, and H. T. Tran *Nonlinear feedback controllers and compensators: a state-dependent riccati equation approach* Comput. Optim. Appl., pp. 177–218, 2007

[4] C. Rasmussen and C. Williams *Gaussian Processes for Machine Learning* Adaptive Computation and Machine Learning, MIT Press, 2006

AutoencODEs: an extension of NeurODEs for width-varying Neural Networks

Alessandro Scagliotti

TU Munich & Munich Center for Machine Learning scag@ma.tum.de

Cristina Cipriani, Massimo Fornasier

TU Munich & Munich Center for Machine Learning cristina.cipriani@ma.tum.de

TU Munich & Munich Center for Machine Learning massimo.fornasier@cit.tum.de

In 2017, it was observed that Residual Neural Networks (ResNets) can be studied as discretization of continuous-time control systems, which are often called NeurODEs. In the last years, Control Theory has been fruitfully applied to study the properties of existing networks, and to develop new ones. Since the dimension of the phase-space of a NeurODE is constant, they have not been used so far to model Deep Learning architectures where the dimensions of the inputs and the outputs vary along the layers. In particular, this is the case for Autoencoders, where the dimension of the data is compressed during the encoding phase, and it increases during the decoding. In our work [2], we model a continuous-time Autoencoder, which we call AutoencODE, and we extend to this case the mean-field control framework already developed for classical NeurODEs. Moreover, we tackle the case of low Tikhonov regularization, resulting in possibly non-convex landscapes of the cost functional. It turns out that most of the results holding globally in case of high Tikhonov regularization (see [1]) can be recovered in regions where the loss is locally convex.

References

[1] B. Bonnet, C. Cipriani, M. Fornasier, H. Huang. *A measure theoretical approach to the mean-field maximum principle for training NeurODEs*, Nonlinear Analysis, 227: 113–161, (2023).

[2] C. Cipriani, M. Fornasier, A. Scagliotti. *From NeurODEs to AutoencODEs: a mean-field control framework for width-varying Neural Networks*, arXiv preprint: 2307.02279 (2023).

A low rank ODE for spectral clustering stability

Stefano Sicilia

Gran Sasso Science Institute, L'Aquila (Italy) stefano.sicilia@gssi.it

Nicola Guglielmi

Gran Sasso Science Institute, L'Aquila (Italy) nicola.guglielmi@gssi.it

Spectral clustering is a well-known technique which identifies k clusters in an undirected graph with weight matrix $W \in \mathbb{R}^{n\times n}$, by exploiting its graph Laplacian

$$L(W) = \operatorname{diag}(W\mathbf{1}) - W, \qquad \mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n,$$

whose eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and eigenvectors are related to the k clusters. Since the computation of λ_{k+1} and λ_k affects the reliability of this method, the k-th spectral gap $\lambda_{k+1} - \lambda_k$ is often considered as a stability indicator. This difference can be seen as an unstructured distance between $L(W)$ and an arbitrary symmetric matrix $L_\star$ with vanishing k-th spectral gap.

A more appropriate structured distance to ambiguity such that $L_\star$ represents the Laplacian of a graph has been proposed in [2]. Slightly differently, we consider the objective functional

$$F(\Delta) = \lambda_{k+1}\left(L(W + \Delta)\right) - \lambda_k\left(L(W + \Delta)\right),$$

where Δ is a perturbation such that $W + \Delta$ has non-negative entries and the same pattern of W. We look for an admissible perturbation $\Delta_\star$ of smallest Frobenius norm such that $F(\Delta_\star) = 0$.

In order to solve this optimization problem, we exploit its low rank underlying structure. Similarly to [1], we formulate a rank-4 symmetric matrix ODE whose stationary points are the optimizers sought. The integration of this equation benefits from the low rank structure with a moderate computational effort and memory requirement, as it is shown in some illustrative numerical examples.

References

[1] Nicola Guglielmi, Christian Lubich and Stefano Sicilia, *Rank-1 matrix differential equations for strutured eigenvalue optimization*, SIAM Journal on Numerical Analysis, 2023

[2] Eleonora Andreotti, Dominik Edelmann, Nicola Guglielmi and Christian Lubich *Measuring the stability of spectral clustering* , Linear Algebra and its Applications, 610: 673–697, 2021

[3] Nicola Guglielmi and Stefano Sicilia *A low rank ODE for spectral clustering stability*, arXiv preprint arXiv:2306.04596, 2023

Alternating Projections for Matrix Completion

Mattia Silei

Dipartimento di Matematica e Informatica, Università di Firenze, Viale Morgagni 67/a, 50134, Firenze

mattia.silei@unifi.it

Stefania Bellavia, Simone Rebegoldi

Dipartimento di Ingegneria Industriale, Università di Firenze, Viale Morgagni 40, 50134, Firenze

stefania.bellavia@unifi.it, simone.rebegoldi@unifi.it

Matrix Completion (MC) involves estimating missing entries in a partially observed matrix, by exploiting the inherently low-rank structure of the underlying data [2, 3]. Such a problem arises in various fields, including recommendation systems, image inpainting, and collaborative filtering.

In this talk, we consider the reformulation of the MC problem as a feasibility problem, under the assumption that the rank of the sought matrix, say r, is known. More precisely, we reformulate MC as the problem of finding a matrix belonging to the intersection of two matrix spaces, i.e., the set of matrices with the same observed entries as the original one and the set of the matrices with rank less or equal to r. We focus on the Alternating Projection Method (APM), i.e., an iterative procedure where we alternatively project onto the above two matrix spaces. We investigate the convergence of the APM method applied to the MC problem, by exploiting some analytical properties of the involved sets. Furthermore, we report computational results comparing the performance of the proposed approach with the widely used SVT method [1], both on synthetic and real-world data. The experiments show the reliability of the proposed approach and highlight its strengths and limitations.

References

[1] Cai J.F., Candès E.J., Shen Z. *A singular value thresholding algorithm for matrix completion* SIAM J. Optim. 20, 4 (2010), pp. 1956-1982

[2] Candès E. J. and Recht B., *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[3] Ramlatchan A., Yang M., Liu Q., Li M., Wang J., Li Y. *A survey of matrix completion methods for recommendation systems* Big Data Mining and Analytics, 1(4) (2018) pp. 308-323

Exploiting Copulas Families and empirical density estimation with Spline Quasi-Interpolation for Unsupervised Classification

Cristiano Tamborrino, Antonella Falini, Francesca Mazzia
Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Italy

cristiano.tamborrino@uniba.it, antonella.falini@uniba.it, francesca.mazzia@uniba.it

Data clustering is still an open research problem. Determining how the data should be grouped together, especially when their joint distribution does not fit the classical Gaussian distribution and the data size is too large for visual inspection, is difficult to achieve. In recent years, Copulas have gained popularity for addressing these challenges [1, 4]. Copulas enable capturing the dependence between data features independently of the choice of marginal distributions [2]. In this study, we introduce a flexible Copula mixture model with a non-parametric estimation of the marginals using spline quasi-interpolation [3]. Numerical experiments show the advantages of employing this strategy for non-parametric estimation of the marginals, leading to improved computational efficiency and component identification, particularly for larger datasets. We put our proposed method to the test using both artificial and real datasets.

Acknowledgements. This research was supported by the following grants: the research of Antonella Falini is founded by PON Ricerca e Innovazione 2014-202 FSE REACT-EU, Azione IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione" CUP H95F21001230006, the research of Cristiano Tamborrino is founded by "PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI" CUP H97G22000210007 under the NRRP MUR program funded by the NextGenerationEU.

References

[1] Nelsen, Roger B. (2006) *An Introduction to Copulas*. Springer Series in Statistics Berlin, Heidelberg.

[2] Harry Joe (2014). *Dependence Modeling with Copulas*. 2^{nd} edition, Chapman and Hall/CRC, New York.

[3] F. Mazzia, and A. Sestini (2009), The BS class of Hermite spline quasi-interpolants on nonuniform knot distributions. *BIT Numerical Mathematics*, **49**, 611–628.

[4] Tamborrino C.; Mazzia F. (2023) *Classification of hyperspectral images with copulas.* Journal of Computational Mathematics and Data Science.

Operational research and machine learning to optimize the performance of route optimization and customer satisfaction in logistics - A case study

Tea Tavanxhiu

University of Tirana, Albania ttavanxhiu@gmail.com

This study focuses in exploring the enhancement of the outputs from the integration of machine learning in operational research algorithms' parameters and thus the increasing of the customer satisfaction in the logistics sector. Presenting the case study of an Italian startup company, the research investigated different areas of improvement to the operational research technique that uses Vehicle Routing Problem by setting initially fixed parameters that will change in time from the outputs of the machine learning algorithms such as reinforcement learning and neural networks. By integrating the Vehicle Routing Problem with Machine Learning we are aiming to build an intelligent system that creates optimal delivery routes, minimizes transportation costs, improves delivery punctuality, and subsequently, enhances customer satisfaction.

The strength of this research study lies in applying hybrid methodologies in a real-world environment, analyzing both qualitative and quantitative data collected from the company's operations in the last year. The outcome of the study presents significant improvements in route efficiency, reduction in delivery times, and an increase in customer satisfaction ratings. It also brings to light potential challenges and areas for future research. This study contributes valuable insights for the three year doctoral study of the author and also in the data-driven decision making of the operations for the logistics businesses sector.

Keywords: *operational research, vehicle routing problem, machine learning, logistics*

Gaussian processes based data augmentation and expected signature for time series classification

Francesco Triggiano

Scuola Normale Superiore, Pisa francesco.triggiano@sns.it

Marco Romito

Università di Pisa marco.romito@unipi.it

The signature is a fundamental notion that describes a path in terms of algebraic objects (iterated integral) [3]. The various properties of the signature of a path, such as the capability of characterising a path or the existence of a universal approximation theorem, suggest that the signature transform can be extremely useful for analysing or classifying time series [1]. In this work we propose a new time series classification model based on the signature transform. The model merges two main ideas. The first is a stochastic data augmentation based on a Gaussian Processes regression model. The second idea is to capture the relevant features of paths by means of the expected signature, computed over the ensemble obtained in the phase of data augmentation. In particular, the model takes a time series $x = (x_t)_{t\in I}$, over a set of times I, as input, and generates a new set of time series, $\{y^i = (y^i_s)_{s\in T} : i \le K\}$ on a richer set of times T, i. e. $I \subset T$. The generation of the new set of series is performed by sampling from a Gaussian distribution with mean and variance learned by the model itself. The expected signature is estimated by averaging over a normalized version of the signature of each y^i and, then, used to get the prediction. Signature normalization has a strong theoretical explanation: In [2] has been proved that the expectation of a normalized version of the signature is able to characterize the law of a large family of stochastic processes. In our numerical context normalization turns out to be crucial to stabilize the estimate of the expected signature. Our model presents several advantages, such as its scalability and adaptability. It integrates seamlessly in more complex architectures and it can be easily modified in order to solve also time series regression problems.

References

[1] I. Chevyrev, A. Kormilitzin *A primer on the signature method in machine learning*, arXiv preprint arXiv:1603.03788, 2016

[2] I. Chevyrev, H. Oberhauser *Signature moments to characterize laws of stochastic processes*, Journal of Machine Learning Research, 23 (2022), pp. 1–42

[3] T.J. Lyons, M. Caruana, T. Lévy *Differential equations driven by rough paths* , Springer, 2007

Addressing data challenges in ML for Medicine

Rosanna Turrisi, Annalisa Barla

DIBRIS and MaLGa, University of Genova, Genova, Italy

rosanna.turrisi@edu.unige.it, annalisa.barla@unige.it

The integration of artificial intelligence (AI) into medicine has the potential to revolutionize healthcare by enhancing diagnosis, treatment, and overall patient care. In the last decades, Machine Learning (MD) and Deep Learning (DL) techniques have shown promising results in various medical domains, including Disease Detection [4, 5, 6], Tumor Segmentation [1, 2, 3], 3D reconstruction of organs and bones [7, 8, 9]. However, the limited availability and quality of medical data pose significant challenges. Indeed, medical datasets often suffer from small size and class imbalance, posing obstacles that must be overcome with specialized techniques, such as data augmentation [10], transfer learning [11]. This research focuses on applying these methodologies to ML models for diagnosing neurodegenerative diseases, using 3D Magnetic Resonance Imaging (MRI) of the brain. Specifically, two applications are investigated: Alzheimer's Disease (AD) detection, and Amyotrophic Lateral Sclerosis (ALS) differential diagnosis. The AD study analyzes a public dataset of approximately 500 images, exploring the impact of different data augmentation strategies and DL model depths on the diagnostic performance. The ALS study leverages transfer learning to extract DL-based features from the MRI scans, which are then fed into various ML classifiers. The classification performance is compared to that of standard radiomic features. The results of this work demonstrate the essential role of techniques such as data augmentation and transfer learning in improving the accuracy of neurodegenerative disease diagnosis. This research highlights the potential of ML in the medical field, paving the way for its practical application in real-world clinical settings to aid in accurate and timely disease diagnoses.

References

[1] Havaei M. *et al.* Brain tumor segmentation with deep neural networks. *Medical Image Analysis.* **35** pp. 18-31 (2017)

[2] Wang W. *et al.* Transbts: Multimodal brain tumor segmentation using transformer. *Medical Image Computing And Computer Assisted Intervention–MICCAI 2021: 24th Int. Conference Proceedings, Part I 24.* pp. 109-119 (2021)

[3] Chen W. *et al.* S3D-UNet: separable 3D U-Net for brain tumor segmentation. *Brainlesion 4th International Workshop, Held In Conjunction With MICCAI 2018, Revised Selected Papers, Part II 4.* pp. 358-368 (2019)

[4] Senturk Z. Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical Hypotheses.* **138** pp. 109603 (2020)

[5] Jo, T., Nho, K. & Saykin, A. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers In Aging Neuroscience.* **11** pp. 220 (2019)

[6] Faghri F. *et al.* Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. *The Lancet Digital Health.* **4**, e359-e369 (2022)

[7] Wang Y. *et al.* DeepOrganNet: on-the-fly reconstruction and visualization of 3D/4D lung models from single-view projections by deep deformation network. *IEEE Trans Vis Comput Graph* (2019)

[8] Widya A. *et al.* 3D reconstruction of whole stomach from endoscope video using structure-from-motion. *2019 41st Int. Conf Proc IEEE Eng Med Biol Soc (EMBC).* pp. 3900-3904 (2019)

[9] Kasten Y. *et al.* End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images. *Int. Workshop ML For Medical Image Reconstruction in MICCAI.* (2020)

[10] Chlap P. *et al.* A review of medical image data augmentation techniques for deep learning applications. *Journal Of Medical Imaging And Radiation Oncology.* **65**, 545-563 (2021)

[11] Kim H. *et al.* Transfer learning for medical image classification: A literature review. *BMC Medical Imaging.* **22**, 69 (2022)

Industry Talks

How to boost your Industry-Specific Application with A.I.

Nicola Procopio, Nicolò Marziale

Fincons Group AG

nicola.procopio@finconsgroup.com, nicolo.marziale@finconsgroup.com

Nowadays, LLMs (Large Language Models) are cited everywhere. New applications and use cases arise all the time, but what is the best way to leverage them in a specific market?

This presentation discusses two techniques, RAG and Hybrid Search, which combined can improve the performance of LLMs without impacting the resources used.

Virtualizing Tyre Design: A Neural Network Approach to Predicting Noise Emission

Elena Di Lascio, Nicola Melas

Pirelli & C. S.p.A.

elena.dilascio@pirelli.com, nicola.melas@pirelli.com

In the rapidly evolving landscape of tyre manufacturing, the push towards virtualization of the design phases using AI stands as a pivotal strategy for the future. This transformative approach promises significant reductions in development time, fostering quicker time-to-market, while also championing sustainability by minimizing environmental impact. At the heart of this initiative lies the challenge of accurately predicting tyre noise emission, a critical parameter governed by European regulations, especially pertinent in the realm of electric vehicles where tyre noise becomes more pronounced.

In this presentation, the Pirelli Data Science team introduces a innovative neural network algorithm that ingeniously leverages 3D tyre design and key physical attributes to predict noise emission levels. By circumventing the traditional need for physical tests in anechoic chambers, this algorithm not only streamlines the design process but also positions itself as a game-changer in addressing noise pollution challenges.

Posters

Efficient datasets distance measure

Paolo Didier Alfano[1]

paolo.alfano@iit.it

Vito Paolo Pastore[2], Antoine Chatalic[2], Lorenzo Rosasco[1,2], Francesca Odone[2],

vito.paolo.pastore@unige.it, antoine.chatalic@dibris.unige.it, lorenzo.rosasco@unige.it, francesca.odone@unige.it

[1]Italian Institute of Technology (IIT), [2]Genoa University

In the last decade, deep models have emerged as the de facto standard approach to learning problems. At the same time, there has been a growing focus on the data used to train them. However, it is not easy to establish a quantitative relationship between different datasets, even within the same task. Intuitively, how can we compute a *distance* between datasets or measure *how similar* they are? Even on simple tasks, the concept of dataset distance is elusive. This research line has been partially explored via optimal transport techniques [1], and divergence measures such as Kullbach-Leibler [2].
In this work, focusing on image datasets, we first map a collection of datasets into the same feature space via a state-of-the-art, pre-trained neural network. We then propose two strategies to compute similarity between datasets. The first one exploits histograms to represent the dataset feature distributions, which can be seen as an empirical approximation of the corresponding marginals. The second one relies on kernel methods and random Fourier features. We show that, with both approaches, it is possible to compute a coherent metric, when dealing with a limited number of real-world datasets. Moreover, our pipeline does not involve any training process, resulting in an efficient methodology.

References

[1] Alvarez-Melis, David and Fusi, Nicolo *Geometric dataset distances via optimal transport*, Advances in Neural Information Processing Systems

[2] Achille, Alessandro and Lam, Michael and Tewari, Rahul and Ravichandran, Avinash and Maji, Subhransu and Fowlkes, Charless C and Soatto, Stefano and Perona, Pietro *Task2vec: Task embedding for meta-learning* , Proceedings of the IEEE/CVF international conference on computer vision

Topological Machine Learning and industrial applications

Serena De Benedictis

Università degli Studi di Bari, Dipartimento di Matematica serenadebenedictis98@gmail.com

Topological data analysis (TDA) is a recently developed computational technique, that uses algebraic topology to study the "shape" of an object or a dataset. TDA can handle efficiently different kinds of input data and can aggregate all significant information of the problem under study, summarizing its most important properties.
The most prominent tools of TDA are Persistent Homology (PH) and Mapper. PH is a method for computing topological features of a space whereas Mapper allows a graph representation of the source data, enabling its condensed visualisation.
To analyse industrial processes, it is useful to combine TDA and Machine Learning (ML). The benefit of this technique integration, which is frequently referred to as Topological Machine Learning (TML), is that the interaction between TDA and ML succeeds in compensating for the limitations of both approaches, allowing ML to obtain a wide range of input data while simultaneously increasing the interpretability of the data obtained by TDA. TDA and ML can be combined in different ways:

- Information provided by applying PH to the data can be processed to become suitable inputs for ML;
- ML routines can be revised using a much more topological point of view;
- Data summaries can be created in graphs using Mapper.

Different representations have been designed to incorporate topological features into ML algorithms, examples are feature vectors and kernel-based methods. Given that industrial processes can produce a variety of data types, it is possible to conduct a comprehensive analysis that includes the product and its design (also at the engineering level), market research, analysis of production equipment, satisfaction of customer demand for specific features, analysis of product features, and analysis of production sustainability. In literature there are several examples of application to different contexts, suggesting a possible parallelism with inputs from industry.
In literature emerges a gap between theory and practical applications, what we want to do is understand if and how this gap could be filled, creating, if possible, a theoretical and computational framework that could adapt to all contexts.

An NMF-based approach to identify differentially expressed genes in microarray and RNA-seq data

Grazia Gargano

Dipartimento di Matematica, Università degli Studi di Bari Aldo Moro

grazia.gargano@uniba.it

Flavia Esposito, Nicoletta Del Buono

Dipartimento di Matematica, Università degli Studi di Bari Aldo Moro

{flavia.esposito, nicoletta.delbuono}@uniba.it

Microarray and RNA-sequencing (RNA-seq) technologies represent powerful tools for analyzing genome-wide gene expression, allowing for large-scale examination of transcriptional changes associated with biological conditions of interest. An important application of high-throughput technologies for gene expression analyses is the identification of differentially expressed genes (DEGs) between two or more groups of samples, tissues, or populations of cells to reveal the underlying molecular mechanisms that differentiate distinct biological conditions (*e.g.*, disease and normal, treatment and control). Several statistical methods and data analysis pipelines have already been developed to identify DEGs using different pre-defined statistical/filtering thresholds for gene selection [1]. DEGs analysis have often been restricted to two-group comparisons. Nevertheless, comparing only two groups of samples provides limited information on the underlying biology of the system, as it does not consider the possible differences or similarities with other conditions. With only two groups, there is also an increased risk of false positive results, particularly if multiple testing correction methods are not applied.

In this work, we propose a mathematical framework based on nonnegative matrix factorization (NMF)[2] for performing DEGs analysis in gene expression data.This is done, by applying a variant of NMF approach to extract relevant genes from the basis matrix. Nonnegativity of the factors is also used to obtain label of sample membership from the coefficient matrix, which is intersected with a gene-score assignment to extract different expressed genes. Our approach is also able to identify DEGs in an unsupervised manner, without any prior knowledge about the considered conditions. It is based on automatically identify biological features of the groups that behave differently. This kind of procedure is able to discover subtypes of diseases or characterize the molecular mechanisms underlying a particular phenotype, helping in the identification of potential therapeutic targets. We apply the proposed framework for DEGs analysis to two problems in oncology field.

References

[1] Fang, Z., Martin, J., Wang, Z. *Statistical methods for identifying differentially expressed genes in RNA-Seq experiments.* Cell Biosci 2, 26 (2012).

[2] Lee, D., Seung, H. *Learning the parts of objects by non-negative matrix factorization.* Nature 401, 788–791 (1999).

Hypermatrix Analysis of Hi-C Data

David Katz and Yaping Liu

Northwestern University, Chicago david.katz@northwestern.edu

The approximately four meters of linear DNA and its associated chromatin proteins are tightly packed into each cell's nucleus. Despite this considerable compaction, there remains a notable degree of coordination among genomic regions, as illustrated by the precise temporal sequence in which DNA replication origins initiate. In this paper, we demonstrate that low-rank hypermatrix approximations of Hi-C data can be used to study the clustering and coordination of genomic regions.

The folding of the genome in three-dimensional space is a key mechanism of the physical interaction between linearly distant genomic regions, and, thus, a key mechanism of genomic regulation [1]. Hi-C is a laboratory technique that produces information on the three-dimensional architecture of the whole genome. The Hi-C technique produces a symmetric, non-negative contact matrix $X = (x_{ij})$, such that the number x_{ij} measures the intensity of the protein-mediated interactions between genomic regions i and j [2]. One of the most striking characteristics of Hi-C heatmaps is the intrachromosomal checkerboard pattern. The checkerboard-like Hi-C heatmaps suggest that each chromosome is clustered into two components, referred to as A/B compartments in Lieberman-Aiden et al. (2009).

The current method for calculating A/B compartments is based on the Principal Component Analysis (PCA) of the normalized Hi-C contact matrix. The biological meaning of such an eigenvector is unclear [3], and consistency of the eigenvectors across Hi-C datasets is low. In this paper, we present an alternative method for calculating A/B compartments by using a non-negative rank two decomposition of a three-fold hypermatrix with first two components symmetric.

References

[1] Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M. & Ren, B. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*. **16**, 991-993 (2019)

[2] Lieberman-Aiden, E., Van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M. & Others Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326**, 289-293 (2009)

[3] Zheng, X. & Zheng, Y. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics*. **34**, 1568-1570 (2018)

Estimating Koopman operators with sketching to provably learn large scale dynamical systems

Giacomo Meanti

Istituto Italiano di Tecnologia, Genova giacomo.meanti@gmail.com

The theory of Koopman operators allows to deploy non-parametric machine learning algorithms to predict and analyze complex dynamical systems. Estimators such as principal component regression (PCR) or reduced rank regression (RRR) in kernel spaces can be shown to provably learn Koopman operators from finite empirical observations of the system's time evolution [1]. Scaling these approaches to very long trajectories is a challenge and requires introducing suitable approximations to make computations feasible. In this paper, we boost the efficiency of different kernel-based Koopman operator estimators using random projections (sketching). We derive, implement and test the new "sketched" estimators with extensive experiments on synthetic and large-scale molecular dynamics datasets. Further, we establish non asymptotic error bounds giving a sharp characterization of the trade-offs between statistical learning rates and computational efficiency. Our empirical and theoretical analysis shows that the proposed estimators provide a sound and efficient way to learn large scale dynamical systems from observational data. In particular our experiments indicate that the proposed approximate estimators retain the same accuracy of the full PCR or RRR models, while being much faster.

References

[1] Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L. & Pontil, M. Learning Dynamical Systems via Koopman Operator Regression in Reproducing Kernel Hilbert Spaces. *NeurIPS*. (2022)

Mathematical Modelling of Transmission Dynamics and Optimal Control of Meningitis Serogroup A and C in Ghana

Bernard Opoku

Department of Mathematics, Kwame Nkrumah University of Science and Technology (KNUST),

Kumasi-Accra Road, Kumasi, Ghana bopoku49@st.knust.edu.gh

Adu Sakyi, Reindorf Nartey Borkor

Department of Mathematics, Kwame Nkrumah University of Science and Technology (KNUST),

Kumasi-Accra Road, Kumasi, Ghana asakyi@knust.edu.gh, reinbork@knust.edu.gh

In this study, an epidemiological model called the Susceptible, Carrier (Men A), Carrier (Men C), Infected, Protected and Recovered (SCCIPR) is constructed to accurately estimate meningitis's occurrence and transmission. To validate this approach, reported data obtained by researchers on case studies in regions of Ghana that are part of the Sub-Saharan meningitis belt were used. Based on the model's prediction, infected individuals have an equal probability of transmitting the infection to others through contact. Simulation results from the equilibrium-state analysis of this methodology reveal that there is a specific threshold parameter ς such that the disease-free equilibrium is globally asymptotically stable when $\mathcal{R}_0 < \varsigma \leq 1$, whereas backward bifurcation occurs when $\varsigma < \mathcal{R}_0 \leq 1$, in which case the disease-free equilibrium is unstable when $\mathcal{R}_0 > 1$. A unique endemic equilibrium is present when $\mathcal{R}_0$ is near 1, with the possibility of this equilibrium being locally asymptotically stable. Furthermore, the proposed model effectively fits the data from the examples studied. It was deduced from the numerical simulations that high coverage of susceptible individuals receiving vaccinations, combined with the development of an efficient vaccine, is the best strategy for managing the bacteria.

References

[1] Asamoah, J. K. K., Oduro, F. T., Bonyah, E., and Seidu, B. *Modelling of rabies transmission dynamics using optimal control analysis* , Journal

[2] Yusuf, Tunde Tajudeen and Benyah, Francis *Optimal control of vaccination and treatment for an SIR epidemiological model*, Journal

Penalty Hyperparameters Optimization in Non-negative Matrix Factorization problems

Laura Selicato

Università degli Studi di Bari Aldo Moro, Italy laura.selicato@uniba.it

Nicoletta Del Buono

Università degli Studi di Bari Aldo Moro, Italy nicoletta.delbuono@uniba.it

Flavia Esposito

Università degli Studi di Bari Aldo Moro, Italy flavia.esposito@uniba.it

Rafal Zdunek

Politechnika Wroclawska, Poland rafal.zdunek@pwr.edu.pl

The hyperparameter optimization (HPO) problem in learning algorithms represents an open issue of great interest since it can strongly affect any real data analysis. Finding the optimal value associated with the lowest error approximation, a well-enforced constraint, the best clustering performance, or the combination of these criteria are some examples. Due to their nature, these approaches need to solve the optimization problem several times, reflecting a waste of time and resources. This scenario includes the well-known Matrix Decompositions (MDs), which are gaining attention in Data Science as mathematical techniques capable of capturing latent information embedded in large datasets. Among the low-rank MDs, Nonnegative Matrix Factorization (NMF) is one of the most effective methods for analyzing real-life nonnegative data. It can be seen as an optimization problem often penalized to emphasize useful properties of the data matrix, such as sparseness. How to automatically choose optimal penalty hyperparameters is an open question in this context. In this study, we suggested a dynamic choice for the penalty hyperparameter that includes its tuning directly in the optimization problem. We proposed to express the penalty hyperparameters problem in NMF in terms of a bi-level optimization. This problem is approached from two perspectives: the existence and convergence theorems of numerical solutions, under appropriate assumptions, are presented together with the design of a novel algorithm, named Alternating Bi-level (AltBi), which incorporates the hyperparameters tuning procedure into the updates of NMF factors. Results of the existence and convergence of numerical solutions, under appropriate assumptions, are studied, and numerical experiments are provided.

References

[1] Del Buono, N., Esposito, F., Selicato, L. & Zdunek, R. Bi-level algorithm for optimizing hyperparameters in penalized nonnegative matrix factorization. *Applied Mathematics And Computation.* **457** pp. 128184 (2023)