Streszczenia prezentacji
Abstract of presentations

FDR Control via Sorted L-One Penalized Estimator (SLOPE) Method

The Sorted L-One Penalized Estimator (SLOPE) is defined as the solution to the convex optimization problem

$$\hat{b} = arg \min_{b} \left(-l(b) + \sum_{i=1}^{p} \lambda_i |b|_{(i)} \right) \ ,$$

where $l(b)$ denotes the log-likelihood function, $\{\lambda_i\}_{i=1}^{p}$ is a positive non-increasing sequence of tuning parameters, and $|b|_{(i)}$ denotes the i-th largest element of the vector $(|b_1|, \ldots, |b_p|)'$.

In the generalized linear model (GLM) framework, SLOPE performs variable selection by identifying predictors corresponding to non-zero components of $\hat{b}$.

In this talk, we present theoretical results showing that SLOPE (asymptotically) controls the false discovery rate (FDR) in GLMs under suitable regularity conditions. We consider both the low-dimensional setting, where the number of predictors (p) is fixed while the sample size (n) tends to infinity, and the high-dimensional setting, where (p) may grow substantially faster than (n). The theoretical findings are illustrated through simulation studies.

missKnockoffs: A Model-X Knockoff Framework for Variable Selection with Missing Data

Model selection in high-dimensional data has attracted substantial attention over the past two decades. However, existing methodologies rarely address the presence of missing data, despite its prevalence in many contemporary applications. We propose missKnockoffs, a novel extension of the Model-X knockoff framework for variable selection with incomplete observations. The procedure combines missing-value imputation with knockoff-based FDR control and incorporates multiple knockoff generations to improve stability. We also introduce a new aggregation method for knockoff statistics and investigate its theoretical properties. Extensive simulations demonstrate effective FDR control and favorable statistical power across a range of settings. The proposed methodology is further illustrated using real omics datasets, confirming its practical usefulness for high-dimensional inference in the presence of missing data.

Adaptive Bayesian SLOPE

Adaptive Bayesian SLOPE (ABSLOPE) method is a tool for variable selection in high-dimensional data, including data with missing values. The approach connects SLOPE regularization with Bayesian approach, which enables both model selection and estimation of model parameter. This talk focuses on theoretical basis of ABSLOPE and perspectives for future developments.

Network flow methods for alignment and deconvolution of multidimensional spectra

Multidimensional nuclear magnetic resonance (NMR) spectroscopy is an analytical technique widely used in proteomics, metabolomics, and structural biology. It reduces peak overlap and improves interpretability by spreading information across multiple chemical-shift axes. However, comparative or replicate experiments often introduce frequency shifts, hindering accurate matching of corresponding signals. In some experiments, such as variable-temperature NMR, the shifts are intentionally triggered, and frequency tracking provides important information. Accurate spectra alignment is crucial for comparative analysis, automated quantification, and structure elucidation. Another application of NMR spectroscopy is mixture analysis and reaction monitoring. In this context, deconvolution methods seek to identify individual compounds and estimate their relative abundances within complex mixtures. Reliable identification and quantification are frequently challenged by peak overlap, signal distortions, limited spectral resolution, missing signals, and discrepancies in peak positions between the observed mixture and reference compound libraries. Addressing these issues requires robust computational approaches.

In previous work, we addressed one-dimensional alignment and deconvolution using optimal transport theory. By representing spectra as discrete probability measures, we formulated both tasks as mass-transport problems and quantified spectral dissimilarity using the truncated Wasserstein distance. While effective in one dimension, these methods do not scale well to multidimensional spectra due to the high computational cost of evaluating Wasserstein distances. To overcome this limitation, we develop a network-flow-based framework for multidimensional spectral analysis. We show that the Wasserstein distance can be efficiently approximated by solving a minimum-cost flow problem on a network whose nodes represent spectral signals from replicate spectra in the alignment setting and from mixture and reference compounds in the deconvolution setting. This formulation substantially improves scalability while retaining the robustness of optimal-transport-based approaches.

Modeling Hydrogen-Deuterium Exchange Mass Spectrometry data via convolutions of exchange probabilities

Hydrogen-Deuterium Exchange Mass Spectrometry (HDX-MS) enables studying protein structure and dynamics by monitoring changes in masses of molecules due to exchange between hydrogens and their heavier isotope - hydrogen. A common approach to the analysis of data acquired in HDX-MS studies focuses on studying shifts in average masses molecules. However, this is insufficient to capture all types of exchange patterns. In this talk, I will present an alternative approach that uses isotopic distributions of peptides observed in mass spectra to describe complete exchange dynamics of both peptides observed directly, and their overlapping sub-sequences. Based on joint work with Jürgen Claesen (Amsterdam UMC) and Tomasz Burzykowski (UHAsselt).

Challenges of AP-MS data interpretation in biological relevance-driven studies

Investigating protein-protein interaction (PPI) networks helps determine the regulation underlying health and disease, hence it is crucial in successful therapy design. Most common approaches in elucidating PPIs encompass affinity-based techniques aimed at identifying proteins copurifying with the studied protein (AP or AE — affinity purification or enrichment). Due to its definability, mass spectrometry (MS) has been a method-of-choice for detection in such experiments, collectively termed AP-MS (or AE-MS). However, the biologically dynamic nature of PPIs, coupled to a kinetically dependent AP step makes AP-MS data highly variable, and its statistical quantitative interpretation challenging. This being explicit in studies employing biological relevance-driven design with no perturbance to protein expression level and/or absence of large fusion tags. In my talk I will try to emphasize the deemed sources of AP-MS variability from a biophysical perspective, and will be happy to discuss any suggestions for coping with them.

What Counts as Significant? Lessons from Applied Proteomics

Proteomics experiments generate increasingly rich datasets, but converting statistical output into biological insight remains a surprisingly challenging step in practice. Using examples from recent collaborative projects and the newly established Mass Spectrometry Facility within the IN-MOL-CELL infrastructure at IIMCB, Vanessa Linke will discuss selected challenges in the analysis and interpretation of affinity purification mass spectrometry (AP-MS) data, including significance testing, missing value handling, and the relationship between statistical evidence and biological relevance. Through a series of real-world examples, the seminar will explore how different analytical decisions can influence downstream biological interpretation and aim to stimulate discussion on how best to extract meaningful biological insights from complex datasets.

European Funds for Smart Economy
Republic of Poland
Co-funded by the European Union
FNP Foundation for Polish Science

Network flow methods for alignment and deconvolution of multidimensional spectra

Maria Bochenek[1], Michał Startek[1], Krzysztof Kaźmierczuk[2], and Anna Gambin[1]

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097, Warsaw, Poland
[2] Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland

Multidimensional nuclear magnetic resonance (NMR) spectroscopy is an analytical technique widely used in proteomics, metabolomics, and structural biology. It reduces peak overlap and improves interpretability by spreading information across multiple chemical-shift axes.

However, comparative or replicate experiments often introduce frequency shifts, hindering accurate matching of corresponding signals. In some experiments, such as variable-temperature NMR, the shifts are intentionally triggered, and frequency tracking provides important information. Accurate spectra alignment is crucial for comparative analysis, automated quantification, and structure elucidation.

Another application of NMR spectroscopy is mixture analysis and reaction monitoring. In this context, deconvolution methods seek to identify individual compounds and estimate their relative abundances within complex mixtures. Reliable identification and quantification are frequently challenged by peak overlap, signal distortions, limited spectral resolution, missing signals, and discrepancies in peak positions between the observed mixture and reference compound libraries. Addressing these issues requires robust computational approaches.

In previous work, we addressed one-dimensional alignment and deconvolution using optimal transport theory. By representing spectra as discrete probability measures, we formulated both tasks as mass-transport problems and quantified spectral dissimilarity using the truncated Wasserstein distance. While effective in one dimension, these methods do not scale well to multidimensional spectra due to the high computational cost of evaluating Wasserstein distances.

To overcome this limitation, we develop a network-flow-based framework for multidimensional spectral analysis. We show that the Wasserstein distance can be efficiently approximated by solving a minimum-cost flow problem on a network whose nodes represent spectral signals from replicate spectra in the alignment setting and from mixture and reference compounds in the deconvolution setting. This formulation substantially improves scalability while retaining the robustness of optimal-transport-based approaches.

From Structured Sparsity to Causal Effects: GSLOPE, PCGLASSO, and SLOPE-Based Doubly Robust Inference

This talk presents three applications of structured convex regularization in high-dimensional inference. GSLOPE estimates sparse precision matrices while identifying groups of equal or similar entries, whereas PCGLASSO provides scale-invariant estimation of sparse partial-correlation matrices. By recovering conditional-dependence structures, these methods can reduce the search space and provide useful building blocks for causal graph learning. I will also discuss the potential of SLOPE-based doubly robust methods for estimating causal treatment effects.