Union of Intersections (UoI) for interpretable data driven discovery and prediction in neuroscience

Kristofer Bouchard, Pratik Sachdeva, Mahesh Balasubramanian, Sharmodeep Bhattachrayya, Shashanka Ubaru

Summary: Interpretable and predictive statistical data analysis methods can provide insight into the biological processes that generate neuroscience data. Neuroscience applications include network formation, receptive field estimation, determination of the genetic control of behaviors, and the calculation of meaningful low-dimensional representations. However, commonly used statistical inference procedures generally fail to identify the correct features, and further introduce consequential bias in the estimates. To address these issues, we developed Union of Intersections (UoI)[1], a flexible, modular, and scalable framework for enhanced statistical feature selection and estimation. Methods (e.g., regression, classification, dimensionality reduction) based on UoI perform feature selection and feature estimation through intersection and union operations, respectively. In the context of linear regression (specifically $\text{UoI}_{\text{LASSO}}$), we summarize formal statistical proofs and extensive numerical investigation on synthetic data to demonstrate tight control of false-positives and false-negatives in feature selection with low-bias and low-variance estimates of selected parameters, while maintaining high-quality prediction accuracy (in the sense of cross-validation). In neuroscience data, we demonstrate: (i) the extraction of sparse, predictive, and interpretable functional networks from human ECoG during speech production and non-human primate single-unit recordings; (ii) increased model parsimony of spatio-temporal receptive fields from retinal ganglion cells; (iii) sparse genetic control of complex behavioral phenotypes; and (iv) parts-based decomposition (NMF) of neural spectrograms. Through optimization of single-node performance as well as multi-node scaling, we demonstrate the application of UoI to 4TB sized data sets in 45 minutes on a supercomputer[2]. To highlight the generality of the UoI framework, we show (with $\text{UoI}_{\text{Logistic}}$, UoI_{CUR}, and UoI_{VAR} variants of the framework) improved prediction parsimony for classification, accurate matrix factorization, and dynamic modeling on diverse neuroscientific datasets[1,3]. These results demonstrate that UoI improves interpretation and prediction across diverse neuroscience applications.

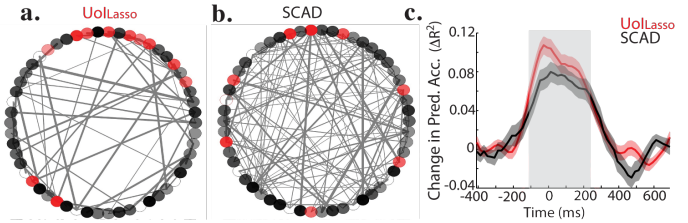

Figure 1

Significance: The increasing size and complexity of neuroscientific data could dramatically enhance basic discovery. Realizing this potential requires novel statistical analysis methods that are both interpretable and predictive. By *interpretable*, we mean that one can interpret the output of the method in terms of processes generating the data. This typically requires identification of a small number of elements of the actual data (sparsity) and accurate estimation of their contribution (low-bias and low-variance). By *predictive*, we mean optimizing the performance of some machine learning measure such as precision, recall, etc. However, there is often a trade-off between interpretability and predictive power, and methods that satisfy both are lacking. This tradeoff is particularly acute for neuroscientific applications, where the output of the model is used to provide insight into neurobiological functions. For example, in estimating networks from data (i.e., 'functional connectomics'), a core problem is determining the adjacency matrix; statistically, this is equivalent to feature selection (i.e., which edges are non-zero). While *ad-hoc* thresholding can be performed, such approaches are often set by hand and defy rigorous understanding. Likewise, statistically biased estimates can result in incorrectly inferred dynamics or under-estimated spike rates. These issues are fundamental to many neuroscientific data analyses and the interpretation thereof. The UoI framework, a novel statistical approach we have recently introduced, addresses these issues. In several real neuroscience data sets solving diverse real neuroscience analysis problems, we show qualitatively and quantitatively different results when using UoI-based methods compared to 'standard' methods. The algorithms enhanced by UoI (e.g., regression, classification, dimensionality reduction) are ubiquitous in systems neuroscience. Thus, the foundational statistical improvements of UoI-based methods will potentially have significant impacts across neuroscience.

Methods Summary: UoI is not a single method or algorithm but a *flexible* statistical framework into which other algorithms (regression: LASSO; classification: Logistic; dimensionality reduction: CUR/NMF; dynamics: Vector Autoregressive models) can be inserted. UoI-based methods leverage stochastic data resampling and a range of sparsity-inducing regularization parameters/dimensions to build families of potential feature sets robust to resamples (i.e., perturbations) of the data, and then average nearly unbiased parameter estimates of selected features to maximize predictive accuracy. UoI separates model selection with intersection operations from

model estimation with union operations: thus UoI is *modular*. The limitations of selection by intersection are counteracted by the union of estimates, and vice versa. Together, these innovations lead to state of the art statistical selection, estimation, and prediction accuracy. Importantly, this is done without explicitly imposing a prior on the parameters, and without introducing a non-convex optimization problem. In addition, we leveraged UoI's *scalable* nature along with a new random data distribution strategy on distributed computing systems to solve $\text{UoI}_{\text{LASSO}}$ with an improved implementation of the alternating directions method of multipliers (ADMM). Our UoI algorithms will be open-source and available through the Python Package Index (PyPI).

Results: Due to space constraints, we highlight a subset of our applications to neuroscience datasets. *Network formation*- We constructed sparse, predictive, and neuroscientifically-meaningful graphs from multi-electrode (86 electrodes) neural recordings taken directly from the surface of the human brain during speech production (45 trials)[4]. To estimate functional coupling, we calculated partial correlation graphs independently for each electrode, and compared the results of graphs estimated by $\text{UoI}_{\text{LASSO}}$ to those estimated by SCAD (a more advanced method than LASSO). $\text{UoI}_{\text{LASSO}}$ graphs were sparser (Fig. 1a, b) and more predictive (Fig. 1c) than graphs estimated with SCAD. Furthermore, the community structure (red nodes in each graph) extracted from the $\text{UoI}_{\text{LASSO}}$ graph correctly identified groups of electrodes in the speech production area, while the SCAD network did not contain this structure (data not shown for space). We created connectivity graphs from spiking data (after a variance-stabilizing square root transform) in monkey V1 (static grating stimuli) and monkey M1 (reaches along a grid). Functional coupling models fit by $\text{UoI}_{\text{LASSO}}$ maintained predictive quality (Fig 2A,D), using fewer parameters when beneficial to the model (Fig 2B,E), resulting in more parsimonious models of the network (BIC: Fig. 2C,F, lower is better). *Receptive fields*- We fit spatio-temporal receptive fields (STRFs) to spiking activity from retinal ganglion cells in response to 1-d white bar stimuli[6] using both $\text{UoI}_{\text{LASSO}}$ and LASSO. Example STRFs are depicted in Figure 3A; due to the improved selection and reduced bias, we observed that $\text{UoI}_{\text{LASSO}}$ provides sharper ON-OFF receptive fields compared to LASSO. This is quantified by lower selection ratios with no cost to predictive performance (Fig. 3B). *Genetic control of complex phenotypes*- We analyzed data from 365 mice in the genetically diverse Collaborative Cross cohort[7]. We regressed single-nucleotide polymorphisms (SNPs) from the entire genome of each mouse (11,563 SNPs) against a behavioral phenotype (rotorod performance). Compared to LASSO, $\text{UoI}_{\text{LASSO}}$ explained slightly more variance(=0.66 vs 0.65) but with 10^2–10^3 fewer parameters. *Dimensionality reduction*- We utilized UoI_{NMF} to obtain a parts-based decomposition of multi-unit triggered, average neural spectrograms (MuTANS) from rat auditory cortex. MuTANS describe the average spectral structure of cortical surface electrical potentials recorded at individual electrodes, triggered on the time of a multi-unit event. UoI_{NMF} demonstrated that all MuTANS (1195 total) could be reconstructed from bases that resembled actual MuTANS, results not observed by PCA, ICA, or base NMF (Fig. 4). *Theory:* Statistical proofs were derived showing superior asymptotic convergence rates of $\text{UoI}_{\text{LASSO}}$ relative to LASSO. In simulations, we compared $\text{UoI}_{\text{LASSO}}$ to Ridge, LASSO, SCAD, and debiased LASSO. Across a variety of model conditions $\text{UoI}_{\text{LASSO}}$ generally resulted in the highest selection accuracy (low false positives, low false negatives) and parameter estimates with lowest error (low bias, low variance), leading to the best prediction accuracy (cross-validated) and prediction parsimony (cross-validated BIC).

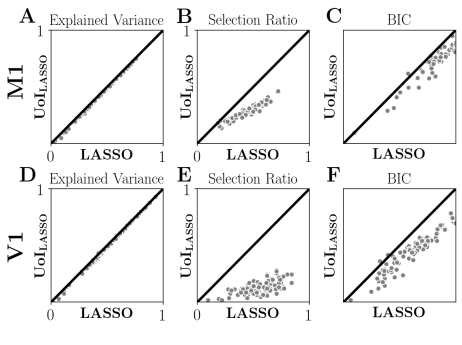

Figure 2

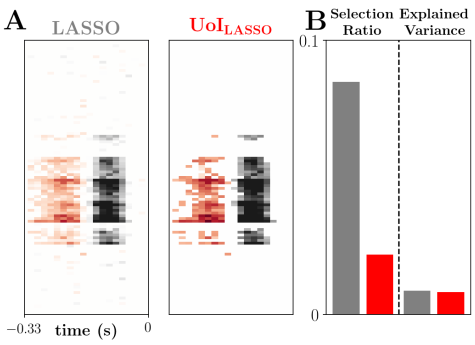

Figure 3

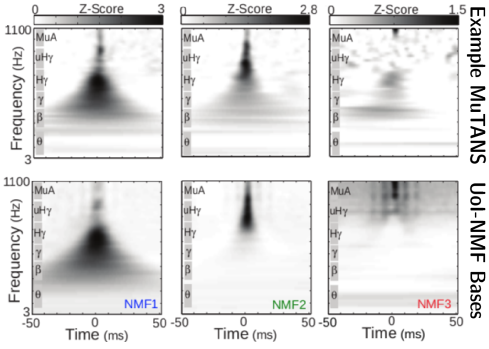

Figure 4

[1]Bouchard et al., *NIPS*, 2017; [2]Ubaru et al., *ICML-A*, 2017; [3]Balasubramanian et al., *arXiv*, 2018; [4]Bouchard et al., *Nature*, 2013; [5]O'Doherty et al., *Zenodo*, 2017; [6]Zhang et al., *CRCNS*, 2014; [7]Mao et al., *Sci. Rep.*, 2015.