Innovations in Multimodal Data Fusion and Data-Driven Discovery in Population Neuroimaging

Lisa Nickerson, PhD Organizer
McLean Hospital/Harvard Medical School
Imaging Center
Belmont, MA
United States

Symposium

This timely symposium aims to provide a dive into multivariate data-driven unsupervised machine learning methods for multimodal data analysis that are of increasing importance in this exciting new era of population neuroimaging. Our presentations cover a range of topics that represent the latest developments in statistical methods for multimodal data fusion and how these methods can overcome the challenges and opportunties of large-scale population neuroimaging – from a survey of new state-of-the-art methods for data fusion and a deep investigation into two of the most popular methods for cross-modal data analysis, to data-driven identification of imaging confounds in large-scale datasets and a new scalable computational method for super big data fusion. Our symposium is very timely given the recent shift in our field to large-scale population neuroimaging to simultaneously map brain activation, functional connectivity, white matter microstructure and connectivity, and structural morphometry patterns that link to population health, behavior and function.

Objective

The learning objectives of this symposium are to gain understanding and insight into: 1) the diversity of multimodal data fusion statistical methods, including advantages and challenges of both popular cross-modal data fusion methods and brand new analytical approaches, 2) the range of applications for data fusion to address the challenges of population brain imaging, and 3) recent developments in super big data fusion for analysis of full voxel-wise multi-modal neuroimaging data in tens of thousands of participants.

Target Audience

Our target audience is any researcher who is working with multimodal brain imaging data and/or data from large-scale population studies such as HCP, ABCD, UKBB or other large-scale datasets constructed from smaller studies.

Presentations

A Diversity of Multimodal Data Fusion Approaches

Multimodal data fusion has been shown to be generally more informative and predictive than unimodal data analyses, often revealing novel information that is invisible to a single modality. There have been a wide variety of approaches presented, each with different advantages. More recently it has become clear that neuroimaging data may benefit from models that are constructed (or adaptive) to varying types of complexity or subspaces. Here we highlight three complementary approaches that highlight the flexibility and benefits that can be obtained with each. We first present a direct, principled approach for fusing multimodal data to identify complex multidimensional subspace structures that capture the underlying modes of shared and unique variability across and within datasets. A new method called multidataset independent subspace analysis (MISA) is designed to permit joint information and shared variability contained in multiple (ofttimes heterogeneous) datasets to be leveraged together in a flexible and synergistic fashion. We also show that this approach provides a robust generalization of many multivariate approaches including independent component analysis (ICA), independent vector analysis (IVA), and independent subspace analysis (ISA) models via a single unified model. Next, we present a model that adapts to varying levels of depth, or complexity, when fusing information flow among different combinations of brain networks, in the context of maximizing performance. And finally, we present an approach called parallel group ICA + ICA, which enables us to link together temporal information in fMRI and spatial information in brain structural images, in a single joint modeling approach. Results from several large datasets highlight the ability of these models to capture joint information that reflects the heterogeneity and complexity of the human brain. The presented approaches have broad utility including (but not limited to) multimodal information fusion, including sample poor regimes and low signal-to-noise ratio scenarios, promoting novel applications in both unimodal and multimodal brain imaging data.

Presenter

Vince Calhoun, GSU/GATech/Emory, TReNDS Center, Atlanta, GA, United States

---

Data Fusion for Data-Driven Identification of Scanner Effects for Denoising Imaging Confounds from Large-Scale Neuroimaging Data

Large-scale neuroimaging datasets are either designed to be collected on multiple different scanners located at different imaging sites, as with the Adolescent Brain Cognitive Development (ABCD) and UK Biobank studies, or they are constructed by pooling data from different existing research studies. These strategic approaches for increasing sample size and statistical power offer exceptional opportunities to enhance reliability and reproducibility of neuroscience research. However, scanner confounds hinder pooling data collected on different scanners or across software and hardware upgrades on the same scanner, even when all acquisition protocols are harmonized. While these confounds reduce power and can lead to spurious findings, methods to address this problem are scant. We present a novel multi-modal data fusion approach that implements a linked independent component analysis (LICA) for data-driven identification of scanner-related effects that can be used to denoise scanner effects from combined MRI data. We demonstrate this approach using existing study data from several small boutique studies of chronic cannabis use that were all collected on a single 3T scanner, but with different scanner operating systems, across a major hardware upgrade, and using different acquisition parameters. Our proposed denoising method shows a greater reduction of scanner-related variance compared with standard GLM confound regression or ICA-based single-modality denoising. Our LICA-based approach should prove even better at identifying scanner effects in multi-site data as between-scanner variability is generally much larger than within-scanner variability. As such, our method has great promise for identifying imaging confounds in large-scale multi-site population studies that are not identified with traditional approaches.

Presenter

Lisa Nickerson, PhD, McLean Hospital/Harvard Medical School, Imaging Center, Belmont, MA, United States

Stability of CCA/PLS for Cross-Modality Multivariate Associations in Population Neuroimaging

Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) are increasingly applied to population neuroimaging datasets to discover multivariate feature profiles that carry inter-individual associations across different modalities. To study the stability of CCA/PLS in the typical regimes of population neuroimaging, we developed a generative modeling framework to produce synthetic datasets parameterized by dimensionality, variance structure, and association strength. We found that CCA/PLS associations could be highly unstable and inaccurate when the number of samples per feature is relatively small. We confirmed these trends in two large neuroimaging datasets, Human Connectome Project (n~1000) and UK Biobank (n=20,000), linking functional and diffusion MRI (fMRI and dMRI) connectivity features to behavioral and demographic features. These model and empirical findings, in conjunction with a meta-analysis of estimated stability in the brain-behavior CCA literature, suggest that typical CCA/PLS studies in population neuroimaging are prone to instability. Finally, we provide a software package, GEMMR, for calculation of estimation errors and required sample sizes for CCA/PLS.

Presenter
John Murray, PhD, Yale University New Haven, CT, United States

---

Super Big Data Fusion for Phenotype Discovery in Large-Scale Population Neuroscience Studies

Data-driven discovery of patterns of population variability in the brain is complicated by computational challenges in analyzing many different structural and functional imaging modalities from thousands or tens of thousands of participants. We present a multimodal independent component analysis based approach that is scalable for data fusion of voxel-level data in the full UK Biobank (UKBB) dataset that will soon reach 100,000 imaged participants. This new computational approach can estimate modes of population variability that enhance the ability to predict thousands of phenotypic and behavioral variables using data from UKBB and the Human Connectome Project. A high-dimensional decomposition achieved improved predictive power compared with widely-used analysis strategies, single-modality decompositions and imaging-derived phenotypes (IDPs) constructed by experts. In UKBB data (14,503 subjects with 47 different data modalities), many interpretable associations with non-imaging phenotypes were identified, including multimodal spatial maps related to fluid intelligence, handedness and disease, in some cases where IDP-based approaches failed.

Presenter
Christian Beckmann, PhD, Radboud University, Cognitive Neuroscience, Nijmegen, Nijmegen, Netherlands