23w5030: Statistical Challenges for Complex Brain Signals and Images

Carolina Euan, (Lancaster University, UK)
Hernando Ombao, (King Abdullah University of Science and Technology, Saudi Arabia)
Mark Fiecas, (University of Minnesota, School of Public Health)

Apr 30 - May 5, 2023

This workshop aimed to bring together a diverse group of experts in the statistical analysis of brain signals, including experienced leaders in the field, early career researchers, and students with novel ideas. Thirty-two participants attended: four PhD students, seven early-career researchers, six mid-career researchers, and fifteen experienced professors. The workshop was held in a hybrid mode, with its own logistic challenges that we will describe later in this report.

1 Overview of the Field

Statistical analysis of large and complex structured data sets, such as high-frequency data or images, is a central challenge of this new era. Classical statistical methodologies can handle data that satisfy certain assumptions, such as stationarity or linear dependence. However, when applying to complex data, such assumptions contradict empirical observations. The statistical analysis of brain data is crucial to understanding normal brain function and alterations associated with neurological and mental diseases. Modern statistical methodologies developed for brain data analysis include (but are not limited):

- Non-stationary space-time models.
- Functional time series.
- Bayesian models.
- Networks.
- Clustering for brain data visualizations.

Brain signals reflect the complexity of unobserved brain processes. Thus, the primary considerations for developing statistical models are flexibility, generalizability, and incorporating known biology. These data are typically high dimensional; therefore, the statistical models that account for the dimensionality require substantial computations for inference and prediction. Furthermore, the studies that give rise to the data have features that the statistical models must consider. For instance, the model needs to account for subject-specific effects to reflect the between-subject variation. Therefore more complexity to the data appears by considering variability among different factors. The substantial rise in statistical sophistication and computational tools has opened new avenues in potential approaches to analyzing these data types. However, many challenges remain open problems.

The combined expertise of the participants of this workshop covered a wide range of topics related to statistical models for brain signals and expertise from colleges of the neurological sciences. This interaction helps to understand in depth the complexity of the problem and to increase the interdisciplinary impact of statistical research.

2 Recent Developments and Open Problems

2.1 Neuroscience Scientific Questions

Statistical research motivated by neuroscience applications strongly depends on the multidisciplinary approach. While statistical methodologies can move faster, assuming certain conditions, this might not necessarily hold for neuroscience applications. Or the scientific questions of interest are moving in an opposite direction. Therefore, during this workshop, we host two keynote speakers within the neuroscientist community: 1) Duygu Tosun-Turgut, PhD, a Professor of Radiology and Biomedical Imaging at UCSF and the Founding Director of Medical Imaging Informatics and Artificial Intelligence at the San Francisco Veterans Affairs Medical Center, and 2) Norbert Fortin, PhD, an Associate Professor at the Neurobiology and Behavior School of Biological Sciences at UCI.

Tosun-Turgut, who has expertise in investigating Alzheimer's disease, brought attention to biomarker selection and risk factors for cognitive decline. In specific, the nonlinear landscape of physiological biomarkers and the spatial patterns that have been empirically observed. Additionally, she emphasized how data has been collected. Data collected in clinical studies are based on multiple recruitment sites with possible different machines and technology updates for long-term longitudinal studies. Therefore, a naive statistical model will fail to capture this diversity in the data accurately if ignored. Therefore, new developments will require 1) robustness under diversity of data collection, 2) work under clustering structures of biomarkers, and 3) the ability to use complex domain variables to predict disease outcomes.

Fortin, who has expertise in investigating neural mechanisms to build (or disrupt) memories, stated the spatial and temporal structure of the memory process as key features. Although technologies are capable of capturing more and more complex neuroscience data, the methodologies to analyze all this information have not reached the neuroscientist community. *We are at a tipping point*- according to Fortin, few developments in data science could lead to transformative advances in neuroscience studies. As a specific example, we have discussed the case of rodent data, neural spiking and local field potential (LFP). Under this scenario, some questions of interest are:

1. Can we identify the flow of information among electrodes within trials? Thinking about moving away from the immediate connectivity only and develop a time-evolving (lead or lag) connectivity.

2. Can we use LFP activity across electrodes to decode the information present at different moments in time? Most recent studies focus on the spike data only, but much richer questions of interest might be answered by combining these two different data sources from the same individual.

3. Can we model neural states that are critical for task performance? To propose a model that accounts for the presence of latent sources and focuses on prediction or causality.

Both experts agreed that the communication between statisticians and neuroscientists is crucial to developing new methodologies aligned with scientific questions of interest. Also, the visualization of complex outcomes from the developed learning algorithms needs to be carefully designed to engage both communities in a common language and increase the impact of scientific developments.

2.2 Recent Statistical Methods for Neuroscience Applications

Motivated by the versatility of neuroscience applications, novel statistical methodologies have been developed in the literature. In this workshop, we cover a variety of different approaches to tackle complex challenges. For instance, within the context of functional magnetic resonance imaging (fMRI) group-level analysis, an important task is to create a common reference system for generalizations of the results. The ProMISes algorithm [2] developed recently tackles this problem by adopting a Bayesian approach, which incorporates

a stochastic component to the optimization problem. Based on different simulation studies, it has been shown that the proposed methodology improves model fitting and prediction accuracy after data alignment. In fMRI studies, detecting brain activation accurately is of great interest, and naive approaches will tackle this problem by multiple (voxel level) tests without adjustments to the power. Two novel approaches, using excursion sets [13] and accounting for spatial structures [11], were also discussed. Both approaches considered the spatial correlation between voxels or regions of interest and proposed novel computational strategies to improve the computational feasibility of the developed methods.

Another common feature of neuroscience data analysis is the presence of multi-subject and longitudinal studies; as pointed out by Tosun-Turgut, new developments should account for the diversity of this data and the dependent structures (time or space) involved. Relevant common structures in multi-subject studies can be detected using clustering strategies, e.g., spectral synchronicity [4]. A Bayesian mixture-based experts approach has been developed to identify fNIRS common patterns of infants' emotional reactions and recovery from stress [5]. Novel MCMC approaches to estimate the model parameters are described, and a comparison with the current methods shows that the developed methodology improves the identification of trajectory patterns. Now, clustering techniques can also account for multi-subjects and longitudinal data similarities simultaneously using biclustering [3]. Prof Harezlak introduced new developments on this line that combine a convex formulation of this problem with model-based features. The proposed formulation of the clustering model will account for correlation among individuals due to the longitudinal studies. An R implementation of this new methodology is available in Github [14].

Binary responses in the context of multi-subject studies was also discussed during this workshop. By considering a marginal approach, a penalized generalized estimating equation for relative risk regression (RR-PGEE) has been developed [7]. The main motivation was investigating brain lesion occurrence and understanding the relative risk in clinical applications. The developed methodology carefully addressed the inference problems of binary models for large data dimensions and longitudinal studies by introducing penalties in the likelihood components and Jeffrey priors.

Understanding underlying structures in correlated data is a crucial point in neuroscience applications. For instance, [15] developed a novel Bayesian framework to identify underlying structures in multi-subject event-related potentials (EEG-related) using a Gaussian processes approach. The main motivation was to develop a latent structure in a more accurate way than averaging across subjects, which is a naive approach. A similar problem for fMRI data is to characterize the collected data into ICA components (latent structure). By using a spatial template independent component analysis (stICA) and an SPDE approach, Amanda Mejía and its research group [1] proposed a novel approach to identify subject-level latent strictures in fMRI data.

The statistical analysis of fMRI data has captured the attention of the biostat community over the last decades, and new methodologies are continuously developing. Generalized linear models and time series models are among the most popular methodologies for analyzing fMRI data. However, few models incorporate spatial and temporal dependences simultaneously. Prof Timothy Johnson discussed the SATVAR model to analyze fMRI data series [6]. This model assumes four components: a stimuli-based regression (with spatially dependent priors for the coefficients), a non-parametric trend, a time-varying auto-regression component, and a spatially correlated error term (v-CAR). This model is fitted using a Bayesian approach where parameters are updated in parallel to speed computations. When applying this new methodology, the results are more appropriate in terms of model performance and interpretation of the results.

2.3 Alternative Methods to Analyse Brain Data

The interface between machine learning approaches and statistical methodologies is also a growing area of research for neuroscience applications. Novel techniques, including supervised learning, multilayer networks and deep kernel learning [9], are investigated as potential tools to improve the detection of tissue compartments and functional connectivity in fMRI studies. The developed methodologies will combine the machine learning tools with statistical techniques such as Monte Carlo studies, Bayesian inference and hypothesis testing on sub-graphs. In particular, the Multilayer Network Association Method (MOAT) extracts common sub-network structures between different subjects using a bipartite graph approach.

During the discussion, new potential applications of statistical approaches such as change point methods, count time series, and functional time series analysis were highlighted. The speakers introduced novel methodologies to identify change points in images [8] or covariance matrices [12], which can handle large

dimensions. This is a common scenario faced in neuroscience applications such as fMRI or EEG arrays. Additionally, novel statistical models and methods for functional time series analysis could be applied to neuroscience applications, taking advantage of the FDA's approach to handling large data sets in space and time [10].

3 Brain Storming Sessions

We encourage discussions and brainstorming sessions to increase interaction among participants. We host two case-study sessions led by neuroscience experts, followed by guided discussions. In addition, on Thursday afternoon, we had three brainstorming sessions focused on the following tracks:

1. Challenges in developing high-dimensional models for brain signals.
2. Computational challenges for pre-processing, model implementation, visualization, and software development.
3. Machine Learning algorithms and approaches to complement statistical techniques.

The main group was split into three according to the participant's preferences, and they reunited with the main group for feedback. Here are a few points that were reflected that afternoon.

Track 1 was a discussion on the topic "Challenges in developing high-dimensional models for brain signals." The discussion focused on four main subtopics: i) Spatio-temporal statistics, ii) Multiview analysis, iii) Dimension reduction, and iv) Graphical models and copulas. **Spatio-temporal statistics.** Modern fMRI data is on the cortical surface, leading to data on a cortical mesh. This is a very specific space for brain imaging, and a step away from volumetric fMRI data where spatial locations have Euclidean coordinates. Data on the cortical mesh better reflects the geometry of the human brain, but it introduces complexity in spatial modeling since the many of the past developments for spatial models for fMRI data used volumetric fMRI data. Spatial models for data on the cortical mesh exist, but is severely underdeveloped. To circumvent the high spatial dimensions of the data, one popular approach is to summarize the spatial data into different "regions of interest", which is a downsampling of the spatial dimension of the brain to a collection of regions or parcels. The definition of these regions and parcels, however, can be arbitrary and is often decided on by the practitioner. Data-driven creations of these regions or parcellations is an open area. **Multiview analysis.** We were in agreement that we often leave information behind because incorporating all possible data into a single model can be overwhelming, particularly when the data is so large. For instance, multi-modal imaging data are often analysed one modality at a time, but we all believe that borrowing information across modalities will improve the performance of the statistical models. One idea we discussed was to extract features from each modality as a way of summarising each modality, thereby reducing the dimensionality, but this is an open problem. Combining this topic with the parcellation discussion above, structural data from diffusion tensor imaging could be informative in creating data-driven parcels. **Dimension reduction.** Just as how the creation of regions or parcels can reduce the spatial dimension, an alternative approach is to carry out data-driven dimension reduction such as independent component analysis (ICA). ICA is a well-known approach in the fMRI literature, but this is not without its own challenges. Selecting the number of components, for instance, is arbitrary and not automatic, and requires hours of subjective assessments on whether the component is biologically plausible. This use of ICA has only been used for fMRI data, but has not been thoroughly explored for other types of imaging data, including multimodal data where it could have utility in a multiview analysis. **Graphical models and copulas.** Connectivity analyses rely on developments for multivariate models. A common assumption one makes about the data is that they are Gaussian, a simplifying assumption that lets us use many of the developments for graphical models (e.g., the graphical lasso for estimating the precision matrix). This is certainly a simplifying assumption, and we were aware that different distributional assumptions can lead to better model fits, though at the cost of added modeling complexity. There are works that use more complex modelling approaches, such as the use of copulas, but these more complicated models have not made a big impact in the literature due to the challenges with interpreting the results. Finally, we noted that statistical genetics have been using genetic pathways, but an analog of this for brain imaging has not yet been developed.

The discussion for track 2: "Computational challenges for pre-processing, model implementation, visualization, and software development," was developed around the concepts of data visualization and software development. **Data visualization** is often undervalued; however, to establish more impactful collaborations with neuroscientists, statisticians should put more emphasis on the pedagogical aspects of message transmission. Statisticians put a lot of effort into developing novel methodologies that can represent the true phenomena more realistically. This research usually underpins a complex inference methodology. However, this research will have a low impact if we cannot transfer the knowledge to the end user. There is a need to develop clear, simple interface pictures that can convey the main message effectively. This may require specialized training. **software development.** With multiple new methodological developments, the bridge between practitioners and statisticians is narrower if we do not develop software (or prototypes) for implementation. However, software development is not an easy task. It requires careful planning to produce an output that is robust to parameters for calibration, and it can run over a long period with low maintenance costs. The group believed that additional support was required to complete the 'end product' of a research project. For instance, contemplate the inclusion of software engineers in the research team (or at the Faculty level).

Track 3 was a discussion on the topic "Machine Learning algorithms and approaches to complement statistical techniques." Some points of interest in this topic are the computational efficiency of machine learning algorithms and explicable machine learning outputs. **Computational efficiency of machine learning algorithms.** The discussion brought to our attention a large number of new developments in statistics that are accounting for the power of machine learning algorithms, either to speed up computations or to improve prediction power. For instance, for spatio-temporal statistics, machine learning algorithms are used to find posterior estimates or maximizing complex likelihood functions. **Explicable machine learning**. One current limitation of machine learning-based models is the lack of model output interpretation. This limitation has been circumvented when combining statistical models with machine learning algorithms where partial control of the learning object is imposed. In general, the development of merged approaches that take into account the best of both sides is in its first stages, and there is a lot of potential to explore.

4 Interaction Between the Physical and Virtual Environments

After the life experience of the COVID pandemic, research collaborations have evolved. This event forced the research community to adapt to a new way of sharing knowledge via virtual talks using platforms such as Zoom. Even today, some events are running in hybrid mode since virtual talks open the door to a broader participation from audiences that, for external reasons, might not join an event in person. This workshop was no exception; it had five virtual talks, one being a case study talk from a neuroscientist. Opening the door to this type of contribution represented an added value to the workshop, and having good facilities to run this type of event was a key factor.

During the event, we offered a variety of activities such as scientific talks (both short and long), debates, and discussions. However, we were unable to host small breakout room discussions in virtual mode due to the limitations of the technology, which was only available in a single location and not accessible to the different physical locations of the three discussion groups. As a result, virtual participants were unable to join this activity. Nevertheless, we hope a better in-house plan will be implemented in future events.

Another limitation of this type of hybrid format is the time zone; most of our virtual participants could only be active during morning sessions as they were in a different time zone. Future planning on the agenda for hybrid workshops needs to consider this aspect. Overall, it was the first event for some of the organizers that ran in hybrid mode, and we acknowledge that many learning experiences happened. We are happy with the outcomes and the participant's experiences.

5 Outcomes of the Meeting

The workshop was fruitful in brainstorming and participation of each attendant. Networking was at the heart of each session. As part of the most tangible outcomes, we achieved two: video recordings of research talks and a future special issue on the discussed topic.

5.1 Video recordings

By agreeing to have their talk recorded, the majority of the participants have helped us to reach out to a wider audience and make the workshop even more impactful. Their contribution will enable many more people to benefit from this valuable experience. The workshop website will host these video recordings, and anyone interested in the topic can access them using the following link: https://www.birs.ca/events/2023/5-day-workshops/23w5030/videos.

5.2 Special Issue in Data Science in Science

After the event, we connected with Dr. David Matteson, editor-in-chief of Data Science in Science, a journal affiliated with the American Statistical Association, the leading professional organization for statisticians. In collaboration with Dr. Matteson, we created a special issue called “Data Science in the Brain Sciences.” Submissions for the special issue opened shortly after the event, and the special issue has since received 6 submissions. We expect to publish the special issue in 2024.

References

[1] Yu Ryan Yue Jiongran Wang Brian S. Caffo Amanda F. Mejia, David Bolin and Mary Beth Nebel. Template independent component analysis with spatial priors for accurate subject-level brain network estimation and inference. *Journal of Computational and Graphical Statistics*, 32(2):413–433, 2023.

[2] Angela Andreella, Livio Finos, and Martin A. Lindquist. Enhanced hyperalignment via spatial prior information. *Human Brain Mapping*, 44(4):1725–1740, 2023.

[3] Eric C. Chi, Genevera I. Allen, and Richard G. Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, 2017.

[4] Carolina Euán, Hernando Ombao, and Joaquín Ortega. Spectral synchronicity in brain signals. *Statistics in Medicine*, 37(19):2855–2873, 2018.

[5] Haoyi Fu, Lu Tang, Ori Rosen, Alison E. Hipwell, Theodore J. Huppert, and Robert T. Krafty. Covariate-guided bayesian mixture model for multivariate time series, 2023.

[6] Cui Guo, Jian Kang, and Timothy D. Johnson. A spatial bayesian latent factor model for image-on-image regression. *Biometrics*, 78(1):72–84, 2022.

[7] Petya Kindalova, Michele Veldsman, Thomas E. Nichols, and Ioannis Kosmidis. Penalized generalized estimating equations for relative risk regression with applications to brain lesion data. *bioRxiv*, 2021.

[8] Claudia Kirch, Philipp Klein, and Marco Meyer. Scan statistics for the detection of anomalies in m-dependent random fields with applications to image data, 2023.

[9] Guoxuan Ma, Bangyao Zhao, Hasan Abu-Amara, and Jian Kang. Bayesian image-on-image regression via deep kernel learning based gaussian processes, 2023.

[10] Israel Martínez-Hernández and Marc G. Genton. Surface time series models for large spatio-temporal datasets. *Spatial Statistics*, 53:100718, 2023.

[11] Ruyi Pan, Erin W. Dickie, Colin Hawco, Nancy Reid, Aristotle N. Voineskos, and Jun Young Park. Spatial-extent inference for testing variance components in reliability and heritability studies. *bioRxiv*, 2023.

[12] Sean Ryan and Rebecca Killick. Detecting changes in covariance via random matrix theory. *Technometrics*, 65(4):480–491, 2023.

[13] Fabian J E Telschow and Armin Schwartzman. Simultaneous confidence bands for functional data using the gaussian kinematic formula. *J Stat Plan Inference*, 216:70–94, 2022.

[14] Kalen Weaber and Jaroslaw Harezlak. longbc. https://github.com/clbwvr/longbc, 2023.

[15] Cheng-Han Yu, Meng Li, Colin Noe, Simon Fischer-Baum, and Marina Vannucci. Bayesian inference for stationary points in gaussian process regression models for event-related potentials analysis. *Biometrics*, 79(2):629–641, 2023.