Final Project Report

Liza Kostina

November 2025

1 Motivation

Brain connectivity analysis often involves comparing large weighted networks across populations (e.g., patients versus healthy controls) or across time. Although such networks may contain tens of thousands of edges, many neuroimaging questions are more naturally addressed at the level of functional systems (groups of brain regions) rather than individual connections. The Graph-Aware Mixed-Effects model of Kim, Kessler, and Levina (2023) [1] provides a principled framework for this setting by enabling inference at the system (cell) level while accounting for correlations among edges within subjects. This leads to biologically interpretable and statistically valid conclusions that naive edge-wise methods cannot provide. In practice, however, the Graph-Aware Mixed-Effects model is computationally challenging to fit. The original formulation involves expensive covariance operations that become infeasible for realistic datasets such as COBRE (approximately 27,000 edges per subject). The existing R implementation is slow and difficult to adapt to modern workflows that involve multiple covariates, cross-validation, or longitudinal data.

This project aims to address these limitations by developing a scalable and reproducible Python implementation of the EM algorithm for the Graph Aware model. Motivation is driven by applications to large developmental studies such as ABCD (approximately 2,500 subjects with repeated measurements), where efficient computation and flexible model specification are essential. More broadly, this work provides an accessible reference implementation for researchers interested in graph-aware mixed-effects modeling in high-dimensional network data.

2 Project Description

This project implements a scalable and reproducible version of the EM algorithm for the **Graph-Aware Mixed-Effects** model of Kim, Kessler and Levina (2023). The primary contribution is a modular Python implementation that makes the model computationally feasible for high-dimensional brain connectivity data, while remaining easy to run and extend.

Model Overview

The Graph-Aware Mixed-Effects model represents edge-level connectivity measurements while targeting inference at the level of *cells*—pairs of functional systems. The model can be written as

$$y_{m,e} = x_m^\top \alpha_{c(e)} + x_m^\top \eta_e + \gamma_{m,c(e)} + \varepsilon_{m,e},$$

where α captures fixed cell-level effects of interest, η represents edge-level deviations, γ are cell-level random effects, and ε is residual noise. This structure enables interpretable system-level inference while accounting for correlation among edges within subjects.

2.1 Computational Challenge

For typical neuroimaging data (235 ROIs, 27,495 edges, 91 cells), a naive EM implementation would require inverting matrices of dimension on the order of $(55,000 \times 55,000)$ in each M-step, which is computationally infeasible. Although the original paper suggests using block coordinate descent, no implementation details are provided.

2.2 Implementation and Optimization

I implemented two versions of the EM algorithm:

1. 1. **Baseline EM implementation**, which closely follows the original formulation using dense linear algebra. This version serves as a correctness reference but is too slow for practical use.
2. 2. Optimized EM implementation, designed for scalability. Key improvements include:
    - sparse representations for the cell–edge design matrix,
    - block-wise updates that exploit the orthogonality between cell-level and edge-level effects induced by the sum-to-zero constraint,
    - efficient GLS variance computations using the Woodbury identity, reducing complexity from $O(E^3)$ to $O(C^3)$,
    - precomputation and caching of sufficient statistics to eliminate redundant computation,
    - vectorized NumPy/SciPy operations in place of Python-level loops.

Together, these techniques make model fitting feasible for datasets with tens of thousands of edges while preserving statistical correctness.

2.3 Software Structure

The project is organized into a clear and reproducible structure:

- src/model/ contains the EM algorithms and GLS utilities,

- src/design/ constructs cell assignments and covariate design matrices,
- src/design/ constructs cell assignments and covariate design matrices,
- demo/ provides a runnable end-to-end example (with COBRE or synthetic fallback),
- tests/ includes basic correctness checks,
- results/ stores generated outputs and benchmarks.

This modular structure allows users to run the demo with minimal setup and serves as a reference implementation for graph-aware mixed-effects modeling.

3 Lessons learned

Developing an efficient implementation of the Graph-Aware Mixed-Effects model highlighted several important technical and conceptual lessons.

Algorithmic structure matters. The original paper mentions block coordinate descent but does not provide implementation details. Deriving the orthogonality between cell-level and edge-level effects from the sum-to-zero constraint was essential for enabling block-wise updates and making the M-step computationally feasible.

Statistical details affect results. Early implementations used unweighted cell means, which led to discrepancies with published results. Correct GLS estimation required precision weighting of edges within cells. Additionally, differences in preprocessing—such as nuisance covariate regression for age, sex, and motion—substantially affected the number of detected significant cells, underscoring the importance of careful data preparation.

Numerical methods outperform naive implementations. Replacing dense matrix inversions with the Woodbury identity reduced the complexity of variance estimation from edge-level to cell-level dimensions. This change alone transformed an infeasible computation into a practical one.

Course-driven changes. As a result, the project evolved from a slow, single-threaded prototype into a structured and efficient implementation. Loop-based code was replaced by vectorized operations, dense inversions by matrix identities, and ad hoc convergence checks by explicit log-likelihood monitoring. The final design is also structured to support future parallelization.

4 Future Work

Several natural extensions remain for future development:

1. 1. Full longitudinal validation on ABCD data with repeated measurements.
2. 2. Parallelization of EM updates for large-scale studies (10,000+ subjects).

3. 3. Packaging the implementation as a documented, tested open-source library.

References

[1] Kessler D. Kim S. and Levina E. Graph-aware modeling of brain connectivity networks. *The Annals of Applied Statistics*, 17(3):1881–1903, 2023.