Methods for Functional Connectivity Harmonization

Andrew Chen

The Log-Euclidean Framework

While applications to functional connectivity are sparse and quite recent (Dai et al., 2019; Ginestet et al., 2017; Li et al., 2019; Lin et al., 2019), research into how to best handle the space of symmetric positive semidefinite (SPD) matrices has been quite extensive. Among the innumerous methods developed around SPD matrices are regression on covariance (Chiu et al., 1996), local polynomial regression (Yuan et al., 2012), PCA (Horev et al., 2017), and more. Many of these articles stem from a seminal paper by Arsigny et al. (2007) which provides a convenient and computationally-efficient framework for working with SPD matrices. This framework, often dubbed the log-Euclidean framework, stems from the fact that the matrix exponential (with inverse function being the matrix logarithm) is a one-to-one mapping between the space of symmetric matrices and the space of SPD matrices. And furthermore, the geodesic distance between any two SPD matrices can be simply expressed as the Euclidean distance between the matrix logarithms of those matrices. That is, for any two matrices A and B in the space of $n \times n$ real SPD matrices $Sym_*^+(n)$, their geodesic distance $g^2(A, B) = \|\log(A) - \log(B)\|_F^2$. This finding can be utilized in conjunction with generalized notions of mean and variance (see Fr'echet mean and variance and an example by Horev et al. (2017))

The log-Euclidean framework is employed throughout recent work in the functional connectivity literature. Lin et al. (2019) model the matrix logarithm of a SPD matrix as a function of the matrix logarithm of a SPD mean matrix and SPD error term and derive a change-point detection method. Li et al. (2019) model dynamic connectivity by assuming that the matrix log of observations within each subject follow a latent factor Gaussian process.

PVD-Based Approach

Instead of using post-hoc methods to guarantee that harmonized matrices are SPD, we can incorporate the log-Euclidean framework into our existing method, which centers around the principal component analysis (PCA) method using Population Value Decomposition (Crainiceanu et al., 2011, PVD). In the functional connectivity setting, we are instead dealing with sample functional connectivity matrices Σ_{ij} for each of j subjects in i sites. First applying the matrix logarithm, the PVD model in this case assumes that the log covariance matrices can be expressed as

$$\log \Sigma_{ij} = \mathbf{P}\mathbf{V}_{ij}\mathbf{P}^T + \mathbf{E}_{ij}$$

where estimation of $\mathbf{P}$ proceeds according to the original paper. To obtain an analogous principal components decomposition of these functional connectivity matrices, we turn to the reduced-dimension $A \times A$ matrices $\mathbf{V}_{ij}$. Let $\mathbf{v}_{ij}$ be the vectorized versions of the $\mathbf{V}_{ij}$ and perform PCA on these vectors to obtain

$$\mathbf{V_{ij}} = \sum_{k=1}^{K} \Lambda_{ijk}\boldsymbol{\phi}_k + \boldsymbol{\eta}_{ij}$$

where Λ_{ijk} are uncorrelated random coefficients, $\boldsymbol{\phi}_k$ are the eigenvectors obtained from PCA analysis arranged into $A \times A$ matrices, and η_{ij} is a noise process. By multiplying on the left by $\mathbf{P}$ and right by $\mathbf{P}^T$ we obtain

$$\log \Sigma_{ij} = \sum_{k=1}^{K} \Lambda_{ijk}\mathbf{P}\boldsymbol{\phi}_k\mathbf{P}^T + \mathbf{P}\boldsymbol{\eta}_i\mathbf{P}^T + \mathbf{E}_{ij} = \sum_{k=1}^{K} \Lambda_{ijk}\boldsymbol{\Phi}_k + \mathbf{e}_{ij}$$

after some substitutions. From here, we can proceed to correct on the scores across sites via a ComBat-like procedure as detailed in the first section. In this case, we are instead correcting on eigenmatrices $\boldsymbol{\Phi}_k$ rather than eigenvectors, and performing the correction on scores obtained from the lower-dimensional representations $\mathbf{V}_{ij}$. Then we simply take the matrix exponential of these corrected observations to obtain Σ_{ij}^{CovBat}.

Location-Scale Model

Lin et al. (2019) propose a matrix-log mean model for SPD matrices Y_i as follows. Let $\mu_i \in Sym_*^+(m)$ be the subject-specific mean matrix and $\epsilon_i \in T_{\mu_i} Sym_*^+(m)$ be noise term where $T_{\mu_i} Sym_*^+(m)$ is the tangent space to $T_{\mu_i} Sym_*^+(m)$ at μ_i. Denoting log'_{μ_i} as the mapping from $T_{\mu_i} Sym_*^+(m)$ to $T_{\log \mu_i} Sym(m)$. Then their model is

$$\log Y_i = \log \mu_i + \log'_{\mu_i} \epsilon_i$$

This model is well-adapted for change-point detection problems, but not for population-level harmonization. Instead, we propose the following ComBat-like model. Let $\Sigma_{ij} \in Sym_*^+(a)$ be the $a \times a$ covariance matrices indexed by site $i = 1, \ldots, M$ and subjects $j = 1, \ldots, n_i$. Define the population mean across all sites as $\mu \in Sym_*^+(a)$ estimated as $\hat{\mu} = \exp\left(\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \log \Sigma_{ij}\right)$, which is the geometric mean proposed by Arsigny et al. (2007). Then a reasonable model is

$$\log \Sigma_{ij} = \log \mu + \log \gamma_i + \log'_{\mu} \epsilon_{ij}$$

where $\gamma_i \in Sym_*^+(a)$ are the site-specific location shift matrices. In the simplest case, this could just be the matrix exponential of the log site-specific mean minus the log population-level mean $\hat{\gamma}_i = \exp\left(\sum_{j=1}^{n_i} \log \Sigma_{ij} - \log \mu\right)$. If we wanted to incorporate a empirical Bayes estimation procedure, we could assume that the γ_i follow a common inverse Wishart prior distribution $\gamma_i \sim \mathcal{W}(\mathbf{\Psi}, \nu)$; however, this estimation procedure may not be straightforward. Following estimation of the γ_i, variance harmonization could procede by working on the ϵ_{ij} terms to harmonize the site-specific sample Fr'echet variances

$$\hat{V}_i = \frac{1}{n} \sum_{j=1}^{n_i} d^2(\hat{\mu}, \Sigma_i j)$$

which can be achieved by simply shrinking the $\log'_{\mu} \epsilon_{ij}$ terms toward zero by multiplication of scalars δ_i. The δ_i would be chosen to force equality of the $\hat{V}_i$ terms. How this estimation can be done more optimally is an open question, but it is also possible that a prior could be imposed for example $\delta_i \sim$ Inverse Gamma(α, β)

References

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2007). Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. *SIAM Journal on Matrix Analysis and Applications* **29**, 328–347.

Chiu, T. Y. M., Leonard, T. & Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association; Alexandria* **91**, 198.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M. & Punjabi, N. M. (2011). Population Value Decomposition, a Framework for the Analysis of Image Populations. *Journal of the American Statistical Association* **106**, 775–790.

Dai, M., Zhang, Z. & Srivastava, A. (2019). Analyzing Dynamical Brain Functional Connectivity As Trajectories on Space of Covariance Matrices. *arXiv:1904.05449 [cs]* .

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S. & Kolaczyk, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics* **11**, 725–750.

Horev, I., Yger, F. & Sugiyama, M. (2017). Geometry-aware principal component analysis for symmetric positive definite matrices. *Machine Learning* **106**, 493–522.

Li, L., Pluta, D., Shahbaba, B., Fortin, N., Ombao, H. & Baldi, P. (2019). Modeling Dynamic Functional Connectivity with Latent Factor Gaussian Processes. *arXiv:1905.10413 [stat]* .

Lin, Z., Kong, D. & Sun, Q. (2019). Modeling Symmetric Positive Definite Matrices with An Application to Functional Brain Connectivity. *arXiv:1907.03385 [stat]* .

Yuan, Y., Zhu, H., Lin, W. & Marron, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 697–719.