Population shrinkage of covariance (PoSCE) for better individual brain functional-connectivity estimation

Mehdi Rahim, Bertrand Thirion, Gaël Varoquaux

Parietal Team. INRIA/CEA. Paris-Saclay University. France

Abstract

Estimating covariances from functional Magnetic Resonance Imaging at rest (r-fMRI) can quantify interactions between brain regions. Also known as brain functional connectivity, it reflects inter-subject variations in behavior and cognition, and characterizes neuropathologies. Yet, with noisy and short time-series, as in r-fMRI, covariance estimation is challenging and calls for penalization, as with shrinkage approaches. We introduce population shrinkage of covariance estimator (PoSCE) : a covariance estimator that integrates prior knowledge of covariance distribution over a large population, leading to a non-isotropic shrinkage. The shrinkage is tailored to the Riemannian geometry of symmetric positive definite matrices. It is coupled with a probabilistic modeling of the individual and population covariance distributions. Experiments on two large r-fMRI datasets (HCP n=815, Cam-CAN n=626) show that PoSCE has a better bias-variance trade-off than existing covariance estimates: this estimator relates better functional-connectivity measures to cognition while capturing well intra-subject functional connectivity.

Keywords: Covariance, functional connectivity, shrinkage, population models

1. Introduction

Functional magnetic resonance imaging (fMRI) reflects neural activity in the brain through the blood-oxygen-level-dependent (BOLD) signal. Task-free or resting-state experiments (r-fMRI) are used to estimate brain functional connectivity between brain structures or regions. These pairwise interactions capture patterns that can be linked to the cognitive, psychiatric, or neurological status of individuals. With the advent of large cohort studies, functional connectivity has been used in neuroimaging population analyses to study cognitive differences between individuals (Smith et al., 2015; Finn et al., 2015). In clinical studies, functional connectivity can extract biomarkers related to neurological (Varoquaux et al., 2010b; Richiardi et al., 2012), neurodegenerative (Challis et al., 2015), or neuropsychiatric disorders (Zeng et al., 2012; Abraham et al., 2017).

These studies describe individuals by a *connectome*: a matrix of interactions between pairs of brain regions. Connectome matrices are typically based on the covariance of the signal: empirical covariance or Pearson correlation (i.e. normalized covariance). For estimated covariances to capture brain connectivity, they have to overcome the limitations of r-fMRI: low signal-to-noise ratio and short time-series. These limitations can easily lead to unreliable estimates with high within-subject variance, in particular when the number of regions is large. Regularized covariance estimators are used to reduce the variance of the estimates (Smith et al., 2011; Varoquaux et al., 2010a). They rely on injecting a prior on the covariance. For example, sparse inverse covariance models enforce zeros on some partial-correlation coefficients. Sparsity, while very useful for interpretation purpose, entails costly optimizations. Additionally, it has been criticized as an oversimplification of brain connectivity (Markov et al., 2012; Ercsey-Ravasz et al., 2013).

Covariance shrinkage is another class of biased estimators that have appealing theoretical properties in high dimension (Ledoit and Wolf, 2004; Chen et al., 2010). These estimators rely on a convex combination between the empirical covariance and a target matrix –usually the identity. The resulting well-conditioned estimators come with reduced variance at little computational cost. Shrinkage-based covariance estimators have wide applications, such as genomics (Schäfer and Strimmer, 2005) or signal processing (Chen et al., 2010). Current solutions for functional-connectivity estimation shrink the covariance towards the identity matrix (Brier et al., 2015). This prior seems modest and over-biased compared to the information provided by the large cohorts of modern population neuroimaging. Indeed, as suggested in Crimi et al. (2011); Mejia et al. (2016), shrinkage towards population average yields more stable and more accurate functional-connectivity estimates. However, taking into account the variability of the population, beyond its mean, is likely to inject pertinent information and could lead to more useful priors.

In this paper, we introduce Population Shrinkage Covariance Estimator (PoSCE), a covariance shrinkage that uses as a prior a probabilistic distribution of the covariances calculated on a population. The resulting estimator shrinks toward the population mean, but additionally it accounts for the population dispersion, hence uses a non-isotropic shrinkage.

PoSCE uses a Riemannian parametrization of covariances,

Email address: rahim.mehdi@gmail.com (Mehdi Rahim)

tailored to the information geometry and the positive-definite constraint of covariance matrices (Lenglet et al., 2006; Fletcher and Joshi, 2007). Riemannian frameworks for covariances have been successfully applied for testing differences (Varoquaux et al., 2010b), classification (Ng et al., 2014), or regression Qiu et al. (2015) with functional connectivity. Mathematically, we rely on the fact that an information-geometric Riemannian metric gives local Euclidean approximations of the maximum-likelihood risk, and makes it possible to use an efficient minimum-mean-squared-error estimation.

We demonstrate the efficiency of PoSCE through extensive experimental validations on r-fMRI data from large cohorts: 815 healthy subjects from the Human Connectome Project (HCP) dataset (Van Essen et al., 2013); and 626 subjects from the Cambridge Center (CamCAN) dataset (Taylor et al., 2017). Results show that PoSCE offers better bias-variance trade-off compared to state-of-the-art estimators. PoSCE leads to very efficient shrinkage algorithms that makes it usable routinely in any functional-connectivity analysis with a small computation cost. In particular, it does only one pass over the population, avoiding any iterative optimization.

The paper is organized as follows: section 2 gives the mathematical background of the shrinkage and the Riemannian manifold of covariances. Section 3 introduces PoSCE and its implementation. Experimental results are discussed in section 4. Section 5 concludes the paper.

2. Background: shrinkage and manifolds for covariances

Notations. n and p denote the number of samples and variables, respectively. In our case, r-fMRI time-points are the samples and ROIs' signals are the variables. We use boldface uppercase letters for matrices, $\mathbf{A}$. We work on two spaces: the ambient space of covariances $\mathbb{R}^{p\times p}$ and a tangent space. To distinguish from the ambient space, we write $\vec{a}$ for vectors in tangent space and $\vec{\vec{A}}$ for matrices in tangent space.

2.1. Prior art: shrinkage and covariances

2.1.1. Shrinkage and James-Stein estimators

A *risk* –or a loss– function characterizes the efficiency of an estimator. An estimator strives to minimize a given risk between the estimator and the true parameter. A typical risk is the mean squared error (MSE) in regression. Unbiased estimators, such as the empirical mean and variance, are not the best estimators in term of MSE: while their *empirical* risk –the MSE on the observed data– is minimum, their *expected* risk of the population can be improved. Shrinkage estimators such as James-Stein (JS) estimators (James and Stein, 1961) are biased estimators of the mean. This bias improves the estimation of the mean by giving a lower expected MSE risk. The fact that the empirical mean is not the best estimator of the population mean is sometimes known as Stein's paradox (Stein, 1956). Yet, decision making when faced with small data intuitively leads to slightly conservative choices that match Stein's paradox (Efron and Morris, 1977) 1. [1]

From a Bayesian stand-point, the shrinkage can be seen as an empirical Bayes estimation. The overall distribution of the data is the prior used to estimate individual parameters (Efron and Morris, 1973). The posterior mean –conditional on the data– gives the minimum MSE estimator (Lehmann and Casella, 2006, Corollary 4.1.2.). We use this result in the proposed covariance shrinkage model.

2.1.2. Application of shrinkage to covariances

Maximum-likelihood estimates for covariances –empirical covariances– are unstable when the number of variables is large and the sample size is small. Many prior works have used shrinkage as a regularization of the covariance estimation (Dey and Srinivasan, 1985; Daniels and Kass, 2001; Chen et al., 2010). Ledoit and Wolf (2003, 2004) propose a shrinkage estimator of covariances that is optimal in high-dimensional asymptotics with regards to the Frobenius risk –MSE for matrices. The Ledoit-Wolf estimator $\hat{\mathbf{\Sigma}}_{\text{LW}}$ shrinks the empirical covariance $\mathbf{S}$ towards a target $\mathbf{T}$ with a convex combination

$$\hat{\mathbf{\Sigma}}_{\text{LW}} = (1-\lambda)\mathbf{S} + \lambda\mathbf{T}. \quad (1)$$

Ledoit and Wolf (2003) use the identity matrix as a uniform shrinkage target. They propose an analytical calculation of the amount of shrinkage λ such that the Frobenius risk is minimized. As a JS-type estimator, the Ledoit-Wolf estimator is simple and fast to compute. It yields biased estimates that are more stable than the empirical covariance and often recommended for functional connectivity (Varoquaux and Craddock, 2013; Brier et al., 2015). However, when the covariance has a reproducible strong off-diagonal structure, as with functional brain networks, identity-based shrinkage is arguably suboptimal, i.e. induces high bias.

Beyond JS-type estimators, another approach to covariance regularization uses sparsity on the inverse covariance as zeros in the inverse covariance denote conditional independence. Several works rely on the graphical Lasso (Friedman et al., 2007) to estimate a sparse covariance for r-fMRI (Smith et al., 2011). For example, Varoquaux et al. (2010a) use a group-level estimator of sparse covariance. Yet, such models may not yield stable covariance coefficients: the ℓ_1 penalty used for sparsity often leads to unstable selection. In addition, they lead to costly optimization and are not tractable on large scale datasets. Different formulations of a sparsity prior on conditional dependencies, e.g. using Bayesian approaches (Hinne et al., 2014), give more stable estimates, but come with an even greater cost. Our approach to covariance estimation with good bias-variance and stability properties relies not on sparsity, but on leveraging the structure of the data.

[1] Efron and Morris (1977) showed a simple, yet famous, baseball-related example that highlights JS estimator efficiency over empirical average. They estimated batting abilities of players over a season from few observations (45 attempts of 18 players). Results shows that JS shrinkage estimator yields smaller mean squared errors w.r.t the season batting average of each player, compared to simple averages.

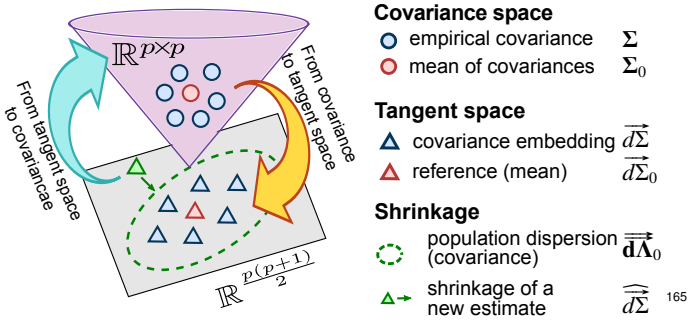

Figure 1: **Tangent embedding and population prior modeling.** $\boldsymbol{\Sigma}_0$ is the mean covariance from a train set of covariances. It is the reference point in the tangent space. The population prior is defined as a Gaussian multivariate distribution centered on $\overrightarrow{\mathbf{d\Sigma}}_0$. $\overrightarrow{\overrightarrow{\boldsymbol{\Lambda}}}_0$ is the covariance dispersion over the population. The arrows depict the mapping between the non-Euclidean covariance space and the tangent space.

2.2. Statistics on Riemannian manifolds for covariances

Information geometry (Amari and Nagaoka, 2007) analyzes families of distributions with Riemannian geometry. When used to parametrize multivariate normal distributions, it describes a manifold structure for covariance matrices. It defines from the Fisher information matrix a Riemannian metric that coincides locally with the KL-divergence. Pennec et al. (2006) and Lenglet et al. (2006) introduce theoretically these geometric concepts on the manifold of normal distributions, with applications to diffusion-MRI tensors.

The space of valid covariances matrices $\mathbb{R}^{p\times p}$ is the set of symmetric positive definite (SPD) matrices. Thus, covariance matrices live on the positive definite cone. Endowed with the Fisher-Rao metric, it defines Riemannian manifold that is well-suited for statistical modeling of covariances. In our case, the r-fMRI time-series for a given subject s are drawn from a Gaussian distribution: $\mathbf{X}_s \sim \mathcal{N}(\mu_s, \boldsymbol{\Sigma}_s)$. For centered data the mean μ_s is 0. The covariance $\boldsymbol{\Sigma}$ captures functional connectivity[2].

As shown in Fig.1, the cone of SPD matrices, endowed with an affine-invariant metric –as the Fisher-Rao metric–, is a Riemannian manifold. This metric is well suited to invariances of the Gaussian model (Pennec et al., 2006; Varoquaux et al., 2010b). The manifold can be projected onto a vector space where Euclidean distances approximate locally Riemannian distances on the manifold. The covariance matrices of a group of subjects can be modeled as drawn from a generalized Gaussian distribution on the manifold, centered on a representative covariance $\boldsymbol{\Sigma}_0$. Statistical model of a covariance $\boldsymbol{\Sigma}$ is then naturally performed by projecting it to $\mathbb{R}^{p\times p}$, using the tangent space representation at $\boldsymbol{\Sigma}_0$ (Varoquaux et al., 2010b). The tangent-space vector $\mathbf{d\Sigma} \in \mathbb{R}^{p\times p}$ is then:

$$\mathbf{d\Sigma} = \operatorname{logm}(\boldsymbol{\Sigma}_0^{-\frac{1}{2}}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}_0^{-\frac{1}{2}}), \qquad (2)$$

where $^{\frac{1}{2}}$ denotes the matrix square root[3] and logm is the matrix logarithm. $\mathbf{d\Sigma}$ stands for the tangent space of $\boldsymbol{\Sigma}$ at the reference covariance $\boldsymbol{\Sigma}_0$. A convenient parametrization $\overrightarrow{d\Sigma} \in \mathbb{R}^d$ with $d = p(p+1)/2$ removes repeated entries while preserving the norm:

$$\overrightarrow{d\Sigma} = \operatorname{vec}(\mathbf{d\Sigma}) = \{\sqrt{2}\,\mathbf{d\Sigma}_{i,j},\ j<i,\ \mathbf{d\Sigma}_{i,i},\ i=1...p\}. \qquad (3)$$

– From Kullback-Leibler divergence to squared error

The Kullback-Leibler (KL) divergence is a probabilistic measure of the discrepancy between two distributions. On the information geometric manifold, the ℓ_2 distance in the tangent space approximates the KL divergence:

Lemma 1. The squared Euclidean distance between the tangent-space embeddings of two covariance matrices is a second-order approximation of the KL divergence between the two corresponding Gaussian distributions.

Proof: The divergence between two centered Gaussian distributions $\mathcal{N}_1, \mathcal{N}_2$ characterized by their covariance $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively, is

$$D_{\mathrm{KL}}(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{2}\left(\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \log|\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1| - p\right)$$

Then, writing $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 + \mathbf{d\Sigma}$:

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{N}_1, \mathcal{N}_2) &= \frac{1}{2}\left(\operatorname{tr}(\mathbf{I} + \boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}) - \log|\mathbf{I} + \boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}| - p\right) \\ &= \frac{1}{2}\left(\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}) - \log|\mathbf{I} + \boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}|\right). \end{aligned}$$

The second-order expansion of the determinant is written:

$$|\mathbf{I} + \mathbf{dA}| = 1 + \operatorname{tr}(\mathbf{dA}) + \frac{\operatorname{tr}^2(\mathbf{dA}) - \operatorname{tr}(\mathbf{dA}^2)}{2} + o(\|\mathbf{dA}\|^2).$$

An important intermediate step is the second-order expansion of $\log(|\mathbf{I} + \mathbf{dA}|)$:

$$\log(|\mathbf{I} + \mathbf{dA}|) = \operatorname{tr}(\mathbf{dA}) - \frac{1}{2}\operatorname{tr}(\mathbf{dA}^2) + o(\|\mathbf{dA}\|^2).$$

Using these two expressions in the KL formula:

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{N}_1, \mathcal{N}_2) &= \frac{1}{4}\operatorname{tr}\left((\boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma})^2\right) + o(\|\boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}\|^2) \\ &= \frac{1}{4}\operatorname{tr}\left((\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\mathbf{d\Sigma}\boldsymbol{\Sigma}_2^{-\frac{1}{2}})^2\right) + o(\|\boldsymbol{\Sigma}_2^{-1}\mathbf{d\Sigma}\|^2). \qquad (4) \end{aligned}$$

This second order expansion corresponds to a squared error of the tangent space embedding that will be used for the covariance shrinkage.

[2] Note that this approach does not model the auto-correlations of time series. This could be handled in future work as a generalization of the model discussed here.

[3] Computed with an eigenvalues decomposition, taking the square roots of the eigenvalues.

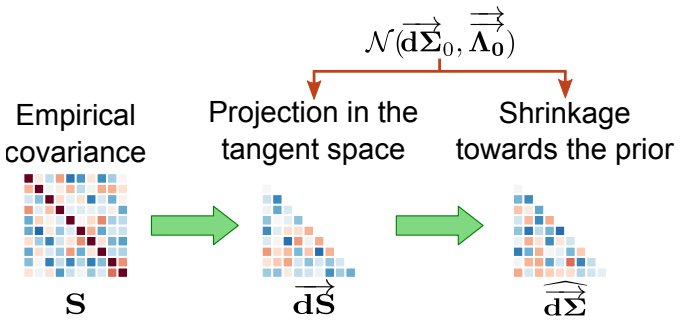

Figure 2: **PoSCE estimation workflow.** The empirical covariance is projected in the tangent space at $\mathbf{\Sigma}_0$ defined previously. This projected covariance is shrunk towards the population prior.

3. Population shrinkage covariance embedding (PoSCE)

Our contribution relies on adapting James-Stein shrinkage theory to the geometry of covariances, using the Riemannian formulation to turn a KL-divergence risk to an MSE risk. We propose a population shrinkage estimator of the tangent parametrization of covariance matrices. This estimator shrinks the covariance estimates toward a prior distribution of covariances, rather than using a constant target as in Ledoit and Wolf (2003). The prior can be seen as a generative model of covariances from which each individual covariance is drawn.

Fig.2 gives an overview of the method : *i)* a prior distribution for covariances is estimated over a training dataset; *ii)* each subject covariance is shrunk according to the prior in the tangent space at $\mathbf{\Sigma}_0$.

3.1. Prior construction from population distribution

We consider the prior as a generative model of the embeddings of the observed covariance models. To define the population prior, we use as inputs a set of covariances $\mathbf{S}_i$ from an r-fMRI dataset. We use the tangent space parameterization of these covariances, by applying formulas (2) and (3). The generative model is then a multivariate Gaussian distribution $\overrightarrow{d\Sigma} \sim \mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$. The prior is centered at the reference point of the tangent space $\overrightarrow{d\Sigma}_0 = \overrightarrow{0}$. The prior covariance $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0$ measures the element-wise dispersion of connectivity matrices in the tangent space. The dispersion with respect to the reference point is given by the mean outer product of the tangent embedding over the train set (Pennec et al., 2006):

$$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \frac{1}{N_{\text{train}} - 1} \sum_{i=1}^{N_{\text{train}}} \overrightarrow{dS}_i \otimes \overrightarrow{dS}_i, \tag{5}$$

where $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 \in \mathbb{R}^{d \times d}$ with $d = p\,(p+1)/2$. $\overrightarrow{dS}_i = \text{vec}(\mathbf{dS}_i)$ is the *parametrized* tangent space transform of the empirical covariance $\mathbf{S}_i$.

3.2. Population prior-based shrinkage

We use the prior distribution $\mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$ for optimal shrinkage of $\overrightarrow{d\Sigma}$ in the tangent space. As mentioned in section 2.1.1, JS-type shrinkage can be expressed in a Bayesian framework:

$$p(\overrightarrow{d\Sigma}|\overrightarrow{DS}) \propto p(\overrightarrow{DS}|\overrightarrow{d\Sigma})p(\overrightarrow{d\Sigma}), \tag{6}$$

where $p(\overrightarrow{d\Sigma})$ is the population prior and $\overrightarrow{DS}$ is the embedded empirical covariance (the embedding of the observed sample covariance). $p(\overrightarrow{DS}|\overrightarrow{d\Sigma})$ is the likelihood of the observed data $\overrightarrow{DS}$ given $\overrightarrow{d\Sigma}$. The log-likelihood is also the cross-entropy between the prior and the data, a key component of the KL divergence between both models: For $p(\overrightarrow{DS}|\overrightarrow{d\Sigma})$, the KL divergence is a natural loss on covariances. We then use the second order expansion to approximate the KL by a quadratic loss in the tangent space, as mentioned in *Lemma 1.* The posterior mean $p(\overrightarrow{d\Sigma}|\overrightarrow{DS})$ –conditional on the data– gives the minimum MSE shrinkage estimator of $\overrightarrow{d\Sigma}$ (see 2.1.1 and Lehmann and Casella 2006, Corollary 4.1.2.)). Thus,

$$\widehat{\overrightarrow{d\Sigma}} = \mathbb{E}[\overrightarrow{d\Sigma}|\overrightarrow{DS}] = \underset{\overrightarrow{d\Sigma}}{\text{argmin}}\, \text{MSE}(\overrightarrow{d\Sigma}, \overrightarrow{DS}). \tag{7}$$

To compute this expectancy we use in Equation 6:

- *Population prior:* $p(\overrightarrow{d\Sigma}) = \mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$ where $(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$ are estimated as in section 3.1.
- *Data likelihood:* $p(\overrightarrow{DS}|\overrightarrow{d\Sigma}) = \mathcal{N}(\overrightarrow{dS}, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}})$, the likelihood of the observed data $\overrightarrow{DS}$ given $\overrightarrow{d\Sigma}$. It is a Gaussian distribution centered on $\overrightarrow{dS}$, the tangent-space projection of the empirical covariance $\mathbf{S}$, and a covariance $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}$ which is a hyper-parameter of the algorithm discussed below.
  Figure 1 depicts both likelihood and prior distributions.

Population prior-based shrinkage (PoSCE) relies on Bayes rule for multivariate Gaussian distributions (Bishop, 2006, Section 2.3.6.). The posterior is $p(\overrightarrow{d\Sigma}|\overrightarrow{DS}) = \mathcal{N}(\widehat{\overrightarrow{d\Sigma}}, \overrightarrow{\overrightarrow{\mathbf{C}}})$, where the inverse posterior covariance is $\overrightarrow{\overrightarrow{\mathbf{C}}}^{-1} = \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} + \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0^{-1}$, and the posterior mean $\widehat{\overrightarrow{d\Sigma}}$ is:

$$\widehat{\overrightarrow{d\Sigma}} = \left(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} + \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0^{-1}\right)^{-1} \left(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} \overrightarrow{dS} + \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0^{-1} \overrightarrow{d\Sigma}_0\right). \tag{8}$$

As the prior mean $\overrightarrow{d\Sigma}_0$ is null, we have:

$$\text{PoSCE:} \qquad \widehat{\overrightarrow{d\Sigma}} = \left(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} + \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0^{-1}\right)^{-1} \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} \overrightarrow{dS}. \tag{9}$$

A fully-fledged estimate of the covariance can be computed by inverting formulas (3) and (2) to go from the tangent space back to the covariance manifold:

$$\hat{\mathbf{\Sigma}}_{\text{PoSCE}} = \mathbf{\Sigma}_0^{\frac{1}{2}} \,\text{expm}\left(\widehat{\mathbf{d\Sigma}}\right) \mathbf{\Sigma}_0^{\frac{1}{2}} \tag{10}$$

3.3. Implementation details

We detail in this section the steps to compute PoSCE. As depicted in figure 2, PoSCE takes as input the empirical covariance $\mathbf{S}$. Then, it uses multivariate Gaussian distributions on the tangent space parametrization of the covariance.

Algorithm 1 summarizes the tangent space parametrization used for population-based covariance shrinkage. It relies on the population reference $\mathbf{\Sigma}_0$. There are several ways to calculate the population reference $\mathbf{\Sigma}_0$. We use the Euclidean mean, since it yields more stable estimations than Fréchet mean, as mentioned in Ng et al. (2014). For a given set of N covariances $\{\mathbf{\Sigma}_i\}$,

$$\mathbf{\Sigma}_0 = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Sigma}_i \tag{11}$$

As depicted in algorithm 3, the population prior is modeled as a Gaussian distribution from a set of covariances. Since the population reference $\mathbf{\Sigma}_0$ is the population mean, the prior is centered at $\overrightarrow{d\Sigma}_0 = \overrightarrow{0}$. $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0$ is the dispersion (covariance) of the prior distribution. In practice, $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0$ is very high dimensional and is learned from a finite population of subjects. Rather than the scatter matrix $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_*$, that corresponds to the maximum likelihood estimate, we use for $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0$ a low-rank approximation which corresponds to regularizing with a PCA decomposition (see algorithm 2). Let $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_* = \mathbf{\Delta} L \mathbf{\Delta}^T$ and $\mathbf{D} = \mathbf{\Delta}_{[1..r]} \sqrt{\mathbf{L}_{[1..r]}}$, where $_{[1..r]}$ denotes the selection of the first r components; in practice r is set such that the captured variance ratio is above 70%. Finally $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \alpha\mathbf{I} + \mathbf{D}\mathbf{D}^T$, where α is set such that $Tr(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0) = Tr(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_*)$.

PoSCE is described in algorithm 4. For a given subject covariance, it uses the prior distribution from algorithm 3. $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}$ cannot be fully estimated from limited data, hence we take $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}} = \lambda\mathbf{I}$, where λ acts as a shrinkage control parameter. In our experiments, we set λ with a cross-validation on a subset of the train dataset. The optimal λ is chosen to maximize the log-likelihood of the test set data, for an estimator calculated on the train set.

Algorithm 1: Tangent embedding parametrization

Input: Covariance $\mathbf{\Sigma}$, reference $\mathbf{\Sigma}_0$

`/* Project covariance in tangent space */`

$\mathbf{d\Sigma} = \text{logm}(\mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{\Sigma}\, \mathbf{\Sigma}_0^{-\frac{1}{2}})$

$\overrightarrow{d\Sigma} = \text{vec}(\mathbf{d\Sigma}) = \{\sqrt{2}\,\mathbf{d\Sigma}_{i,j}, j < i, \mathbf{d\Sigma}_{i,i}, i = 1...p\}$

Output: Covariance embedding $\overrightarrow{d\Sigma}$

Algorithm 2: Low-rank approximation

Input: Matrix $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_*$

$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_* = \mathbf{\Delta L \Delta}^T$ `// eigenvalue decomposition`

Select r components such that variance ratio >= 70%

$\mathbf{D} = \mathbf{\Delta}_{[1..r]} \sqrt{\mathbf{L}_{[1..r]}}$

$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \alpha\mathbf{I} + \mathbf{D}\mathbf{D}^T$, α is set such that $Tr(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0) = Tr(\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_*)$

Output: Approximated matrix $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0$

Algorithm 3: Population prior estimation

Input: Set of N covariances $\{\mathbf{S}_i\}$, reference $\mathbf{\Sigma}_0$

`/* Set embedding parametrization */`

forall *covariance* $\mathbf{S}_i$ **do**

  $\overrightarrow{dS}_i \leftarrow$ Embedding of $\mathbf{S}_i$ according to algorithm 1

end

`/* Build prior distribution` $\mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$ `*/`

$\overrightarrow{d\Sigma}_0 = \overrightarrow{0}$ `// prior mean`

$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \frac{1}{N-1} \sum_{i=1}^{N} \overrightarrow{dS}_i \otimes \overrightarrow{dS}_i$ `// prior dispersion`

$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 \leftarrow$ low-rank approximation according to algorithm 2

Output: Population prior distribution $\mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$

Algorithm 4: Population shrinkage covariance embedding (PoSCE)

Input: covariance $\mathbf{S}$, population prior $\mathcal{N}(\overrightarrow{d\Sigma}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$, shrinkage parameter λ

`/* Build covariance distribution` $\mathcal{N}\left(\overrightarrow{dS}, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}\right)$ `*/`

$\overrightarrow{dS} \leftarrow$ Embedding of $\mathbf{S}$ according to algorithm 1

$\overrightarrow{\overrightarrow{\mathbf{\Lambda}}} = \lambda\mathbf{I}$ `// likelihood covariance`

`/* Shrink towards population distribution */`

$\widehat{\overrightarrow{d\Sigma}} = (\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} + \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0^{-1})^{-1} \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}^{-1} \overrightarrow{dS}$

Output: Shrunk covariance embedding $\widehat{\overrightarrow{d\Sigma}}$

3.4. Relating PoSCE to standard linear shrinkage

We observe that Equation 9 is a generalization of classic shrinkage estimators (Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005) that relies on a convex combination of a prior with the empirical covariance matrix. Below we make that link explicit and show that with $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \lambda_0\mathbf{I}$ and $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}} = \lambda\mathbf{I}$ we recover classic shrinkage.

In PoSCE, the shrinkage is in the tangent space and the amount of shrinkage is controlled by the likelihood covariance parameter λ. Equation 9 summarizes the shrinkage. It can be seen as a generalization of covariance shrinkage as in Ledoit and Wolf (2004) where the target is the average covariance. Indeed, considering uniform prior covariance $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 = \lambda_0\mathbf{I}$ and $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}} = \lambda\mathbf{I}$ in Eq.(9) leads to:

$$\begin{aligned}\widehat{\overrightarrow{d\Sigma}} &= (\frac{1}{\lambda}\mathbf{I} + \frac{1}{\lambda_0}\mathbf{I})^{-1}(\frac{1}{\lambda}\mathbf{I}\overrightarrow{dS} + \frac{1}{\lambda_0}\mathbf{I}\overrightarrow{d\Sigma}_0),\\ &= (1 - \frac{\lambda}{\lambda_0 + \lambda})\overrightarrow{dS} + \frac{\lambda}{\lambda_0 + \lambda}\overrightarrow{d\Sigma}_0.\end{aligned}$$

By taking $\lambda' = \frac{\lambda}{\lambda_0 + \lambda}$, we have:

$$\widehat{\overrightarrow{d\Sigma}} = (1 - \lambda')\overrightarrow{dS} + \lambda'\overrightarrow{d\Sigma}_0. \tag{12}$$

To *back-project* $\widehat{\mathbf{d\Sigma}}$ –with $\widehat{\overrightarrow{d\Sigma}} = \text{vec}(\widehat{\mathbf{d\Sigma}})$– into the ambient space, we choose any reference point $\mathbf{R}$, such that $\overrightarrow{dS}$ and $\overrightarrow{d\Sigma}_0$ are small. For any matrix $\mathbf{A}$ in the vicinity, we have $\mathbf{dA} = \text{logm}(\mathbf{R}^{-\frac{1}{2}}\mathbf{A}\mathbf{R}^{-\frac{1}{2}})$.
Hence, $\mathbf{A} = \mathbf{R}^{\frac{1}{2}}\text{expm}(\mathbf{dA})\mathbf{R}^{\frac{1}{2}} \simeq \mathbf{R}^{\frac{1}{2}}(\mathbf{dA} + \mathbf{I})\mathbf{R}^{\frac{1}{2}}$.
In particular, $\mathbf{S} \simeq \mathbf{R}^{\frac{1}{2}}(\mathbf{dS}+\mathbf{I})\mathbf{R}^{\frac{1}{2}}$ and $\mathbf{\Sigma}_0 \simeq \mathbf{R}^{\frac{1}{2}}(\mathbf{d\Sigma}_0+\mathbf{I})\mathbf{R}^{\frac{1}{2}}$, using first-order approximations.

We apply this to $\widehat{\mathbf{d\Sigma}}$:

$$\begin{aligned}\widehat{\mathbf{\Sigma}} &\simeq \mathbf{R}^{\frac{1}{2}}(\widehat{\mathbf{d\Sigma}} + \mathbf{I})\mathbf{R}^{\frac{1}{2}},\\ &\simeq \mathbf{R}^{\frac{1}{2}}((1 - \lambda')\mathbf{dS} + (\lambda'\mathbf{d\Sigma}_0 + \mathbf{I}))\mathbf{R}^{\frac{1}{2}},\\ &\simeq \mathbf{R}^{\frac{1}{2}}((1 - \lambda')\mathbf{dS} + (1 - \lambda')\mathbf{I} + \lambda'\mathbf{d\Sigma}_0 + \lambda'\mathbf{I})\mathbf{R}^{\frac{1}{2}},\\ &\simeq (1 - \lambda')\mathbf{S} + \lambda'\mathbf{\Sigma}_0,\end{aligned} \tag{13}$$

which corresponds to a linear shrinkage of covariance $\mathbf{S}$ towards the target $\mathbf{\Sigma}_0$ (formula (1) and Ledoit and Wolf 2003).

4. Experimental validation

In this section, we compare PoSCE to state-of-the-art estimators and show that it achieves a better bias-variance trade-off than current alternatives for various applications. Ideally, an estimator should be stable to sampling noise, e.g. have high test-retest reproducibility, yet it should retain key inter-subject differences. Hence, we specifically probe these differences as they are of direct interest for applications. With four different experiments, we assess: *i)* predicting subject age from its connectivity profile; *ii)* capturing functional connectivity similarities within twins, siblings and unrelated subjects; *iii)* characterizing multi-dimensional cognitive and behavioral phenotypes based on functinal connectivity; *iv)* reproducibility of connectivity estimates across two sessions of the same subject. In all experiments, covariances are estimated on signals fom 64 brain regions from bootstrap analysis of stable clusters (BASC) atlas (Bellec et al., 2010).

We use two distinct datasets in our experiments, namely the Human Connectome Project (HCP) dataset and the Cambridge Center for Ageing and Neuroscience (CamCAN) dataset. In each experiment, the dataset is split in two subsets. The first subset –200 randomly selected subjects– is used for population prior estimation. The second subset is used for the validation.

4.1. Performance in across-subject prediction

We assess the capacity of PoSCE to provide relevant connectivity features for out-of-sample prediction. Predicting subject phenotype or clinical status from neuroimaging data helps to identify brain biomarkers of a physiological or mental state. Several studies have highlighted the impact of the age on the brain. They rely on a regression model to predict subject age from r-fMRI (Liem et al., 2017) or anatomical MRI (Franke et al., 2010; Cole et al., 2017).

In this experiment, we consider the accuracy of age prediction from functional connectivity as a measure to benchmark different connectivity estimators. We use the CamCAN dataset, a publicly available dataset for brain ageing. It includes around 700 healthy subjects aged from 18 to 88 years. The dataset is presented in Taylor et al. (2017). Selected subjects and preprocessing steps are detailed in the appendix. We compare three age-prediction models based on three functional connectivity estimators as subject descriptors : correlation matrix, the tangent space embedding of the covariance, and the proposed PoSCE. We use a linear support vector machine regressor with the same parameters ($C = 1$). Predictive models are compared through 100 randomized cross-validations on left out data (10% of the dataset). At each iteration, the prediction accuracy is measured with the mean absolute error (MAE). It measures the average discrepancy in years of the predicted age compared to the true age. Learning curves in Figure 3 correspond to the three age-prediction models. The learning curves are obtained by varying the number of subjects included in the train set while keeping the test set fixed.

The results show that the PoSCE estimator systematically outperforms other functional connectivity estimators. Shrinking towards the population prior improve age-prediction accuracy, including in small sample-size setting. The results also demonstrate the benefits of using the tangent space parametrization rather than correlation matrices. This corroborates previous studies comparing different functional connectivity estimators for disease prediction (Abraham et al., 2017; Dadi et al., 2016).

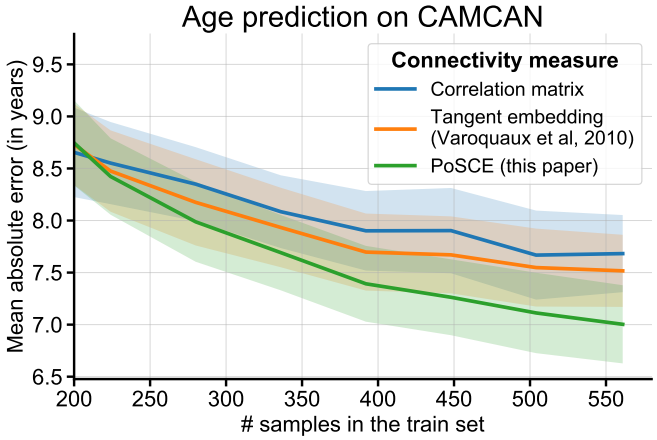

Figure 3: **Estimator performance for age prediction.** Learning curves shows that tangent space-based estimators decrease age overall error. The population-shrinkage estimator yields better age predictions.

4.2. Agreement with phenotype similarities

We evaluate the extent to which PoSCE estimator captures phenotype similarities between subjects. Recent studies investigate the link between phenotype similarity and brain connectivity. Colclough et al. (2017) measure the heritability of the functional connectivity on the Human Connectome Project dataset (HCP) (Van Essen et al., 2012). These studies show that twins have closer functional-connectivity profiles compared to other subjects. We perform a similar experiment on the same dataset, by comparing connectivity profiles of 815 subjects from HCP. First, we compute the connectivity profile for each subject using either PoSCE or the correlation matrix. Then, we compare the pairwise euclidean distances between the connectivity profiles of twins, siblings, and unrelated subjects.

Figure 4 shows the distribution of the distances for the three groups with PoSCE and the correlation matrix. PoSCE-based distances of twins are significantly smaller than siblings and unrelated subjects. On the opposite, distances based on correlation matrices do not capture significant differences between group: they show a wider spread and a smaller group effect. Figure B.7 in the appendix shows that PoSCE also significantly improves upon standard shrinkage or sparse covariance estimators. Overall, the results reproduce the findings of Colclough et al. (2017). They show the benefit of using a good covariance estimator using the population prior to characterize phenotypical differences.

4.3. Association with behavioral scores

We measure the correlation of PoSCE with multi-dimensional description of subjects demographics, psychomet-

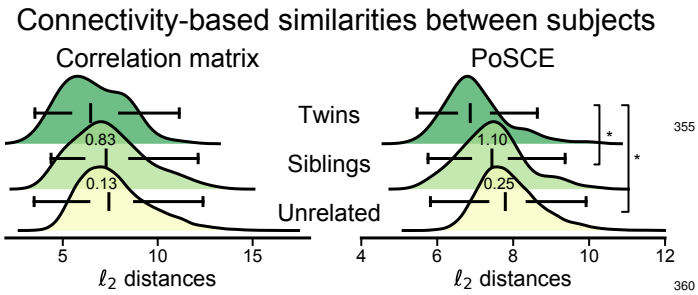

Figure 4: **Similarities between HCP subjects based on functional connectivity profiles.** Plots represent distributions of pairwise distances between connectivity profiles of each two twins, siblings and unrelated subjects. Digits represent differences between each two distributions. Group comparisons show higher differences of twins compared to siblings and unrelated subjects. Such differences are highlighted by the PoSCE estimator (*: $p < 0.001$, 10 000 permutations), whereas Pearson correlation does not capture twins specific similarities: the distributions overlap and there a no significant differences.

rics, and lifestyle. This type of analysis is popular in population neuroimaging (Smith et al., 2015; Miller et al., 2016; Xia et al., 2017). Smith et al. (2015) investigate the relationship between behavior and brain connectivity on the Human Connectome Project dataset. They apply a canonical correlation analysis (CCA) with 100 components on 158 selected behavioral scores and subjects functional connectivity profiles. We replicate this analysis on 615 subjects from HCP (we use 200 separate subjects to build the population prior). We compare PoSCE and the correlation matrix as functional connectivity features from r-fMRI scans. We first measure canonical correlation modes within sample, then –more importantly– the correlation of this mode on unseen data.

Figure 5 shows the principal CCA mode scatter plot. This mode relates functional connectivity to behavioral assessments. As expected from Smith et al. (2015), both PoSCE and correlation matrix estimators yield significant canonical correlation (*: $p < 10^{-3}$, 10 000 permutations). Yes, the PoSCE estimator yields more significant co-variations between functional connectivity and behavioral assessments. PoSCE also gives better accuracy on out-of-sample data. To illustrate correlations between connectivity and behavior, we represent the fluid intelligence (IQ) of each subject on the scatter plot in Figure 5 by a *purple-to-yellow* colormap. This shows consistent correlation between IQ and the first functional connectivity mode.

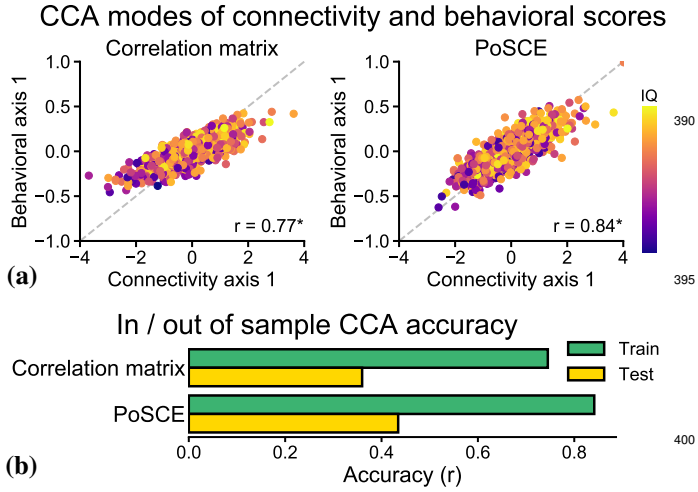

Figure 5: **Relating functional connectivity with behavior on HCP.** Subjects distribution over the first CCA connectivity-behavior mode shows higher correlation when using PoSCE as a connectivity descriptor **(a)**. First CCA mode with PoSCE has a better out-of-sample generalization compared to Pearson correlation **(b)**.

4.4. Reproducibility within subjects

Finally, we study the reproducibility of PoSCE within each subject across sessions. For each subject of the HCP dataset (815 subjects), we use two r-fMRI scans acquired at different sessions (rest1, rest2). We measure the fidelity of an estimator by the log-likelihood of the data from a r-fMRI session –rest2– in a model estimated on a previous session –rest1– from the same subject. For a covariance model $\boldsymbol{\Sigma}$ estimated on rest1, and observed data from rest2 characterized by the correlation matrix $\mathbf{S}$, the log-likelihood of the data observed from rest2 is: $1/2(-\mathrm{tr}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1}) + \det(\boldsymbol{\Sigma}^{-1}) - p\log(2\pi))$.

We compare six covariance models for each subject: *1)* the correlation matrix without any regularization; *2)* the graphical Lasso –GraphLassoCV from scikit-learn (Pedregosa et al., 2011)–, where a sparse precision matrix is estimated with an ℓ_1 penalty (Friedman et al., 2007; Smith et al., 2013); *3)* the Ledoit-Wolf estimator, that shrinks to the identity with an analytically set shrinkage parameter (Ledoit and Wolf, 2004); *4)* shrinkage to the identity in the tangent space with a shrinkage set by cross-validation (Identity shrinkage CV); *5)* the estimator with isotropic shrinkage towards the population mean (Prior shrinkage CV); *6)* the proposed PoSCE, with non-isotropic shrinkage towards the population mean, controlled by the population distribution (PoSCE). For the estimators with cross-validation, the optimal shrinkage is set such that it maximizes the log-likelihood on a separate subject session rest3.

Figure 6 summarizes the subject-wise dispersion of the log-likelihoods of each estimator, relative to the mean. The results demonstrate that shrinking the covariance towards the prior produces the highest likelihood values, and outperforms shrinkage to identity. There is a systematic gain with PoSCE compared to only using the mean covariance as target of the shrinkage model, as in Crimi et al. (2011). This suggests that the population distribution is useful to regularize connections that exhibit more variability across subjects. We also observe that the optimal shrinkage is better estimated with cross-validation than with the Ledoit-Wolf method. Indeed, the Ledoit-Wolf estimator strives to minimize a squared-error risk, and not a likelihood risk. Further comparisons show that the results are consistent across different brain atlases (see appendix, Figure C.8).

5. Conclusion

We introduced PoSCE, a covariance model that integrates the knowledge of population distribution for an optimal shrinkage of the covariance.

PoSCE belongs to James-Stein estimators family ensuring lower error than maximum-likelihood estimators. It relies on a parametrization of covariances that enables approximating KL divergences between covariances as Euclidean distances. Hence, it can use shrinkage results for squared error, namely the equivalence between minimum mean squared error and Bayesian analysis of Gaussian models. Our Bayesian formulation integrates not only a shrinkage target, but also the variability of covariance over a reference population for non isotropic shrinkage. It yields straightforward closed-form equations, and is thus computational cheap even on very large cohorts. Yet, PoSCE scalability can be improved to better approximate population-level variability matrix, since it depends on the number of brain regions used in the analysis.

Empirically, PoSCE shows an excellent bias-variance trade-off for brain functional connectivity: it reduces estimator variance (intra-subject variability) while highlighting more accurately co-variations between connectivity profiles and subjects behavioral assessments. Further work should study the transfer of the population prior, for instance across distinct brain-imaging datasets. Ideally, the definition of a *universal* prior to compute connectivity matrices could be used in all functional connectivity analyses, provided they rely on the same initial region definition.

Our extensive experimental results show that PoSCE captures better connectivity-phenotype covariation than all alternative estimators. Hence, it can be used to learn better biomarkers based on functional connectivity. Indeed, there is important ongoing research that builds prediction models of various neurological or psychiatric disorders as well as health outcomes from clinical r-fMRI data.

Acknowledgements. This work is funded by the NiConnect project (ANR-11-BINF-0004_NiConnect).

This project has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement No. 785907 (HBP SGA2).

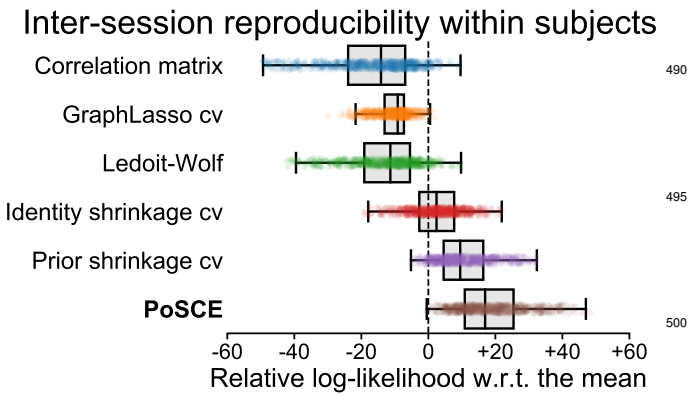

Figure 6: **Fidelity to subject data.** Boxplots represent the average log-likelihood of the data in a second session, which quantifies the relative similarities between respective sessions of 615 subjects from HCP. Shrinking the covariance towards population prior improves reproducibility compared with Pearson correlation, Ledoit-Wolf, and ℓ_1 penalized estimators. PoSCE outperforms all other estimators.

References

Abraham, A., Milham, M.P., Martino, A.D., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage 147, 736–745. URL: https://doi.org/10.1016%2Fj.neuroimage.2016.10.045, doi:10.1016/j.neuroimage.2016.10.045.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics 8. URL: https://doi.org/10.3389/fninf.2014.00014, doi:10.3389/fninf.2014.00014.

Amari, S., Nagaoka, H., 2007. Methods of information geometry. Amer Mathematical Society.

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage 37, 90–101.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51, 1126–1139. URL: https://doi.org/10.1016/j.neuroimage.2010.02.082, doi:10.1016/j.neuroimage.2010.02.082.

Bishop, C.M., 2006. Pattern recognition and machine learning.

Brier, M.R., Mitra, A., McCarthy, J.E., Ances, B.M., Snyder, A.Z., 2015. Partial covariance based functional connectivity computation using ledoit–wolf covariance regularization. NeuroImage 121, 29–38. URL: https://doi.org/10.1016/j.neuroimage.2015.07.039, doi:10.1016/j.neuroimage.2015.07.039.

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., Cercignani, M., 2015. Gaussian process classification of alzheimer's disease and mild cognitive impairment from resting-state fMRI. NeuroImage 112, 232–243.

Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O., 2010. Shrinkage algorithms for MMSE covariance estimation. IEEE Transactions on Signal Processing 58, 5016.

Colclough, G.L., Smith, S.M., Nichols, T.E., Winkler, A.M., Sotiropoulos, S.N., Glasser, M.F., Essen, D.C.V., Woolrich, M.W., 2017. The heritability of multi-modal connectivity in human brain activity. eLife 6. URL: https://doi.org/10.7554%2Felife.20178, doi:10.7554/elife.20178.

Cole, J.H., Ritchie, S.J., Bastin, M.E., Hernández, M.C.V., Maniega, S.M., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2017. Brain age predicts mortality. Molecular Psychiatry URL: https://doi.org/10.1038%2Fmp.2017.62, doi:10.1038/mp.2017.62.

Crimi, A., et al., 2011. Maximum a posteriori estimation of linear shape variation with application to vertebra and cartilage modeling. IEEE Trans on Med Imag .

Dadi, K., Abraham, A., Rahim, M., Thirion, B., Varoquaux, G., 2016. Comparing functional connectivity based predictive models across datasets, in: 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI), IEEE. URL: https://doi.org/10.1109%2Fprni.2016.7552359, doi:10.1109/prni.2016.7552359.

Daniels, M.J., Kass, R.E., 2001. Shrinkage estimators for covariance matrices. Biometrics 57, 1173–1184. URL: https://doi.org/10.1111%2Fj.0006-341x.2001.01173.x, doi:10.1111/j.0006-341x.2001.01173.x.

Dey, D.K., Srinivasan, C., 1985. Estimation of a covariance matrix under stein loss. The Annals of Statistics 13, 1581–1591. URL: https://doi.org/10.1214%2Faos%2F1176349756, doi:10.1214/aos/1176349756.

Efron, B., Morris, C., 1973. Stein Estimation Rule and Its Competitors. An Empirical Bayes Approach. Journal of the American Statistical Association 68, 117. URL: https://doi.org/10.2307%2F2284155, doi:10.2307/2284155.

Efron, B., Morris, C., 1977. Stein paradox in statistics. Scientific American 236, 119–127. URL: https://doi.org/10.1038%2Fscientificamerican0577-119, doi:10.1038/scientificamerican0577-119.

Ercsey-Ravasz, M., Markov, N.T., Lamy, C., Essen, D.C.V., Knoblauch, K., Toroczkai, Z., Kennedy, H., 2013. A predictive network model of cerebral cortical connectivity based on a distance rule. Neuron 80, 184–197. URL: https://doi.org/10.1016/j.neuron.2013.07.036, doi:10.1016/j.neuron.2013.07.036.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neuroscience 18, 1664–1671. URL: https://doi.org/10.1038/nn.4135, doi:10.1038/nn.4135.

Fletcher, P.T., Joshi, S., 2007. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87, 250–262.

Franke, K., Ziegler, G., Klppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from t1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50, 883–892. URL: https://doi.org/10.1016%2Fj.neuroimage.2010.01.005, doi:10.1016/j.neuroimage.2010.01.005.

Friedman, J., Hastie, T., Tibshirani, R., 2007. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441. URL: https://doi.org/10.1093%2Fbiostatistics%2Fkxm045, doi:10.1093/biostatistics/kxm045.

Hinne, M., Ambrogioni, L., Janssen, R.J., Heskes, T., van Gerven, M.A., 2014. Structurally-informed bayesian functional connectivity analysis. NeuroImage 86, 294–305.

James, W., Stein, C., 1961. Estimation with Quadratic Loss, in: 4th Berkeley Symposium on Maths Statistics and Probability.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790. URL: https://doi.org/10.1016/j.neuroimage.2011.09.015, doi:10.1016/j.neuroimage.2011.09.015.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. of Empirical Finance 10.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

Lehmann, E.L., Casella, G., 2006. Theory of point estimation.

Lenglet, C., Rousson, M., Deriche, R., Faugeras, O., 2006. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. Journal of Mathematical Imaging and Vision 25, 423–444.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188. URL: https://doi.org/10.1016%2Fj.neuroimage.2016.11.005, doi:10.1016/j.neuroimage.2016.11.005.

Markov, N.T., Ercsey-Ravasz, M., Ribeiro Gomes, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M., et al., 2012. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cerebral cortex 24, 17–36.

Mejia, A.F., Nebel, M.B., Barber, A.D., Choe, A.S., Lindquist, M.A., 2016. Effects of Scan Length and Shrinkage on Reliability of Resting-State Functional Connectivity in the Human Connectome Project. ArXiv e-prints arXiv:1606.06284.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nature Neuroscience 19, 1523–1536. URL: https://doi.org/10.1038%2Fnn.4393, doi:10.1038/nn.4393.

Ng, B., Dressler, M., et al., 2014. Transport on riemannian manifold for functional connectivity-based classification, in: MICCAI, pp. 405–412.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825.

Pennec, X., Fillard, P., Ayache, N., 2006. A riemannian framework for tensor computing. International Journal of Computer Vision 66, 41–66.

Qiu, A., Lee, A., Tan, M., Chung, M.K., 2015. Manifold learning on brain functional networks in aging. Medical Image Analysis 20, 52 – 60. URL: http://www.sciencedirect.com/science/article/pii/S1361841514001522, doi:https://doi.org/10.1016/j.media.2014.10.006.

Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.M., Greco, B., Hagmann, P., Schluep, M., Vuilleumier, P., Van De Ville, D., 2012. Classifying minimally disabled multiple sclerosis patients from resting state functional connectivity. Neuroimage 62, 2021–2033.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4.

Smith, S.M., Miller, K.L., et al., 2011. Network modelling methods for FMRI. Neuroimage 54, 875.

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E.J., Glasser, M.F., Ugurbil, K., Barch, D.M., Essen, D.C.V., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nature Neuroscience 18, 1565–1567. URL: https://doi.org/10.1038/nn.4125, doi:10.1038/nn.4125.

Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., Barch, D.M., Uğurbil, K., Essen, D.C.V., 2013. Functional connectomics from resting-state fMRI. Trends in Cognitive Sciences 17, 666–682. URL: https://doi.org/10.1016%2Fj.tics.2013.09.016, doi:10.1016/j.tics.2013.09.016.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: 3rd Berkeley Symposium on Maths Statistics and Probability.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-CAN, Henson, R.N., 2017. The cambridge centre for ageing and neuroscience (cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269. URL: https://doi.org/10.1016%2Fj.neuroimage.2015.09.018, doi:10.1016/j.neuroimage.2015.09.018.

Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Penna, S.D., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., Yacoub, E., 2012. The human connectome project: A data acquisition perspective. NeuroImage 62, 2222–2231. URL: https://doi.org/10.1016%2Fj.neuroimage.2012.02.018, doi:10.1016/j.neuroimage.2012.02.018.

Van Essen, D.C., Smith, S.M., et al., 2013. The WU-minn human connectome project: An overview. NeuroImage 80, 62–79.

Varoquaux, G., Craddock, R.C., 2013. Learning and comparing functional connectomes across subjects. NeuroImage 80, 405–415.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B., 2011. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity, in: Biennial International Conference on Information Processing in Medical Imaging, Springer. pp. 562–573.

Varoquaux, G., Gramfort, A., et al., 2010a. Brain covariance selection: better individual functional connectivity models using population prior, in: NIPS, p. 2334.

Varoquaux, G., et al., 2010b. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, in: MICCAI.

Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkin, M.E., Cook, P.A., de la Garza, A.G., Vandekar, S., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C., Gur, R.E., Shinohara, R.T., Bassett, D.S., Satterthwaite, T.D., 2017. Linked dimensions of psychopathology and connectivity in functional brain networks URL: https://doi.org/10.1101%2F199406, doi:10.1101/199406.

Zeng, L.L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., Hu, D., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135, 1498–1507.

Appendix A. R-fMRI datasets

We detail in this section the r-fMRI datasets used in our experiments. For each dataset, r-fMRI timeseries are extracted from a set of ROIs corresponding to a brain atlas. We use the `Nilearn` library (Abraham et al., 2014) for the temporal preprocessing. It includes linear detrending, motion confounds re-

gression, CompCor (Behzadi et al., 2007), and band filtering (0.01–0.1Hz).

– The Human Connectome Project dataset

We use r-fMRI scans from the 900-subjects release of the Human Connectome Project dataset (HCP), to assess estimator reproducibility and phenotype characterization. As summarized in table A.1, 815 quality-checked subjects are selected, including twins, siblings, and unrelated subjects. Each subject has four r-fMRI scans that have been already spatially preprocessed and normalized to the MNI space. Each r-fMRI scan is around 15 min-long comprising 1 200 time-points.

– The Cambridge Center for Ageing Neuroscience dataset

This dataset is used to highlight the interest of using PoSCE to predict subject age. The dataset comprises 626 healthy subjects equally distributed over age ranges. We apply a standard spatial preprocessing pipeline using `SPM12`. (motion correction, coregistration to T1-MRI, normalization to the MNI space).

Table A.1: **Datasets description.**

– The HCP dataset

N	Gender	Zygosity	Age
815	F : 462 M : 353	Twins : 181 Siblings : 100 Unrelated : 214	22 – 37

– The CamCAN dataset

N	Gender	Age
626	F : 318 M : 308	18 – 88

Appendix B. Comparing PoSCE with other estimators

We provide in this section full comparisons between PoSCE and other covariance estimators, by comparing experimental results of different regularizations and shrinkage targets.

– Age prediction

Table B.2 shows accuracies of CamCAN age prediction from brain connectivity extended to other estimators. We use the same setting as explained in section 4.1 : 100 randomized cross-validations on 10% left out data using support vector regression. Cross-validation mean absolute deviation and r-squared values suggest that PoSCE-based connectivity features better capture age-related variations. While shrinkage-based estimators results are similar to those from correlation matrix, ℓ_1 penalized connectivity has the lowest accuracy. It is less suited for inter-individual functional connectivity characterization.

Table B.2: **Estimator performance for age prediction.** Model accuracies (mean ± standard deviation) over 100 randomized train-test splits show better age prediction with PoSCE.

Connectivity measure	MAD	r-squared
Correlation matrix	7.63 ± 0.73	0.73 ± 0.05
GraphLasso cv	8.23 ± 0.68	0.69 ± 0.05
Ledoit-Wolf	7.54 ± 0.73	0.73 ± 0.05
Identity shrinkage cv	7.77 ± 0.73	0.72 ± 0.05
Prior shrinkage cv	7.77 ± 0.73	0.72 ± 0.05
PoSCE	**6.88 ± 0.69**	**0.76 ± 0.06**

Table B.3: **CCA performances** on train and test set.

Connectivity measure	r (train set)	r (test set)
Correlation matrix	0.78 ± 0.01	0.36 ± 0.09
GraphLasso cv	0.77 ± 0.01	0.10 ± 0.08
Ledoit-Wolf	0.78 ± 0.01	0.31 ± 0.07
Identity shrinkage cv	0.78 ± 0.01	0.32 ± 0.07
Prior shrinkage cv	0.77 ± 0.01	0.35 ± 0.08
PoSCE	**0.82 ± 0.01**	**0.43 ± 0.05**

– Phenotype similarities

Figure B.7 shows additional functional connectivity comparisons within similar phenotype groups on HCP. Results sustain the fact that PoSCE highlights better phenotype similarities, while graphical Lasso estimates yield more overlapping distributions of pairwise distances within twins and non-twins.

– Canonical brain-behavior correlations

Table B.3 summarizes the CCA performances on HCP from all connectivity estimators on in-sample (train) and out-of-sample (test) settings. Similarly to age prediction results, PoSCE connectivity features give better CCA generalization whereas graphical Lasso features have the lowest accuracies.

Appendix C. Impact of the brain atlas on PoSCE

In all previous experiments, covariances are estimated on 64 brain regions from bootstrap analysis of stable clusters (BASC) atlas (Bellec et al., 2010). To assess brain atlas impact on connectivity estimation, we compare connectivity estimators reproducibility within subjects as in section 4.4 by choosing different brain atlases. We include two publicly available brain atlases :

- Multi-scale dictionary learning atlas (MSDL) (Varoquaux et al., 2011) with 39 regions.
- Harvard-Oxford atlas from `FSL` software (Jenkinson et al., 2012) with 96 regions.

Figure C.8 shows estimator reproducibility results on HCP using MSDL and Harvard-Oxford atlas. These two atlases yield comparable results to those observed in section 4.4 where the prior has similar contribution to the covariance stability when increasing or decreasing the number of regions.

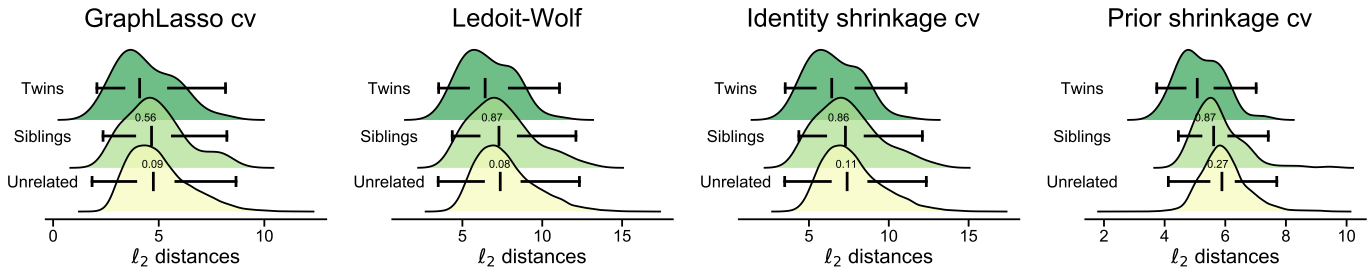

Figure B.7: **Within phenotype similarities.** Plots represent pairwise distance distributions. Digits represent differences between each two distributions. Distances between functional connectivity profiles are better highlighted with PoSCE. Overall, shrinkage yield better similarities compared to ℓ_1 penalized covariance.

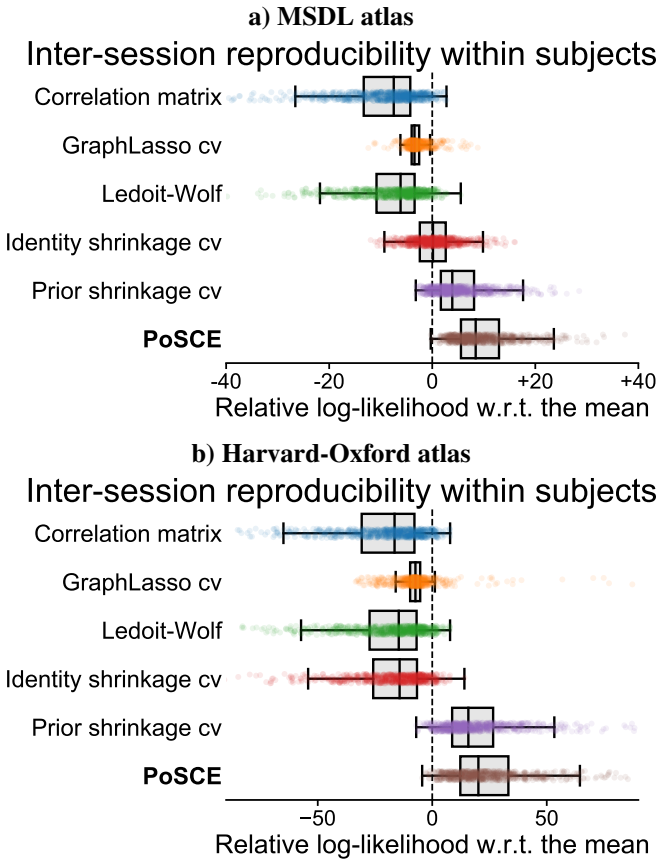

Figure C.8: **Fidelity to subject data.** Using different brain atlases yield similar results : the population prior gives more stable within-subject estimates.

Appendix D. Prior visualization

We provide visualizations of the calculated population distributions. Figure D.9 shows population distribution characteristics of HCP and CamCAN. It depicts the mean $\vec{d}\,\Sigma_0$ and the first two principal component analysis (PCA) modes of the covariance $\vec{\Lambda}$. We note that the mean matrices are similar for HCP and CamCAN datasets where pairwise connectivities are dense and stronger within each functional network (auditory, default mode network, attention, ...).

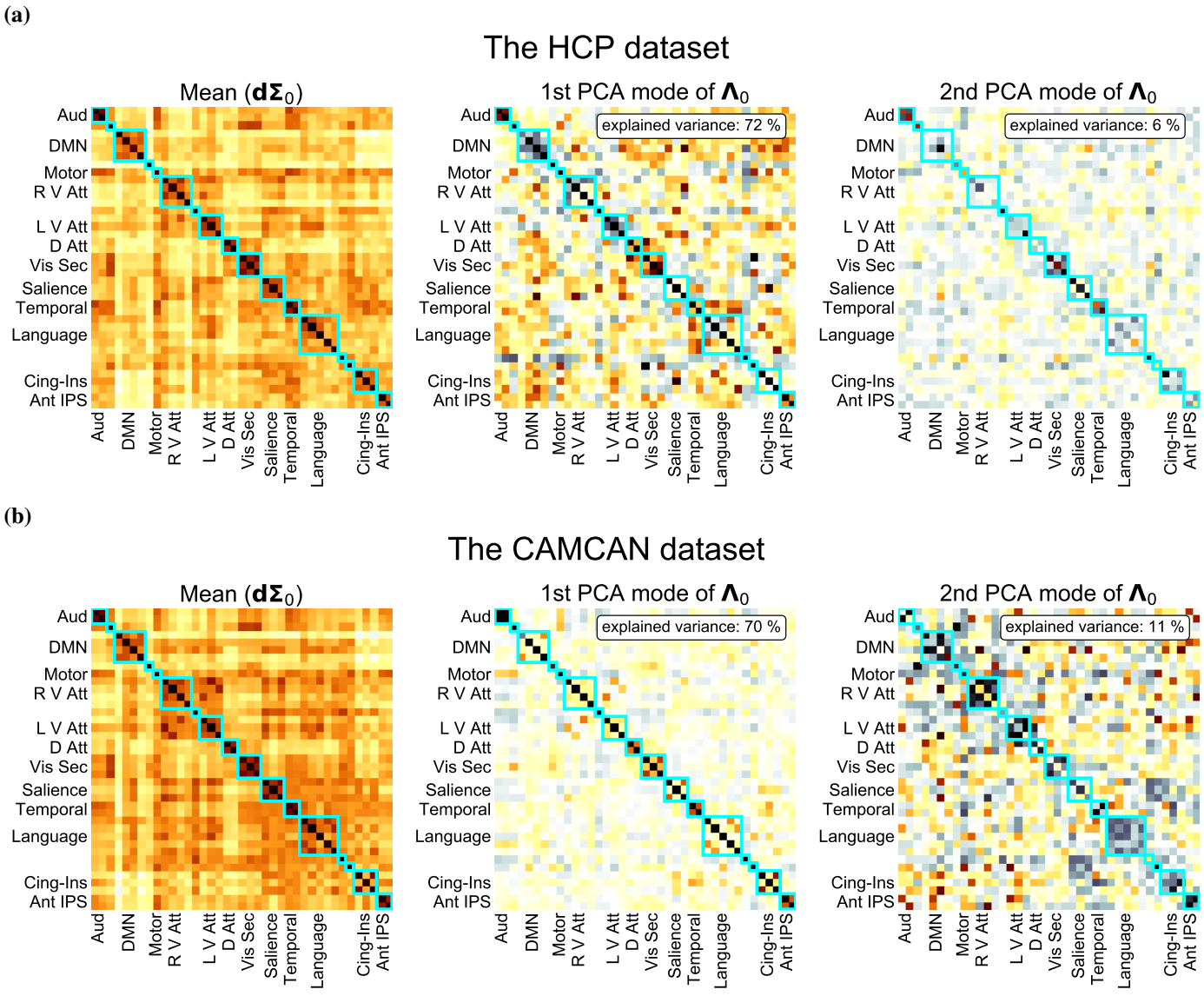

Figure D.9: **Population prior distributions.** Mean and covariance modes of the HCP and the CamCAN datasets.