Population-shrinkage of covariance to estimate better brain functional connectivity

Mehdi Rahim, Bertrand Thirion, and Gaël Varoquaux

Parietal Team, Inria - CEA, Paris-Saclay University, France
`mehdi.rahim@inria.fr`

Abstract. Brain functional connectivity, obtained from functional Magnetic Resonance Imaging at rest (r-fMRI), reflects inter-subject variations in behavior and characterizes neuropathologies. It is captured by the covariance matrix between time series of remote brain regions. With noisy and short time series as in r-fMRI, covariance estimation calls for penalization, and shrinkage approaches are popular. Here we introduce a new covariance estimator based on a non-isotropic shrinkage that integrates prior knowledge of the covariance distribution over a large population. The estimator performs shrinkage tailored to the Riemannian geometry of symmetric positive definite matrices, coupled with a probabilistic modeling of the subject and population covariance distributions. Experiments on a large-scale dataset show that such estimators resolve better intra- and inter-subject functional connectivities compared existing covariance estimates. We also demonstrate that the estimator improves the relationship across subjects between their functional-connectivity measures and their behavioral assessments.

1   Introduction

Functional connectivity captures markers of brain activity that can be linked to neurological or psychiatric phenotypes of subjects. It is commonly used in neuro-imaging population analyses to study between-group differences [1] or to extract biomarkers of a specific pathology [2]. Typically, functional connectivity is measured with an empirical covariance or Pearson correlation (i.e. normalized covariance) between time-series of different brain regions. However, r-fMRI suffers from low signal to noise ratio and small sample sizes. In such regime, the empirical covariance matrix is not a good estimate of covariance, in particular when the number of regions of interest (ROIs) is large. Penalized estimators are used to overcome such limitations by injecting prior [3, 4]. Beyond sparsity, which leads to costly optimization, high-dimensional *covariance shrinkage* has appealing theoretical properties [5, 6]. Such approaches use a convex combination between the empirical covariance and a target matrix –usually the identity– resulting in well-conditioned estimators with little computational cost. They are vastly used for connectivity estimation on r-fMRI [7], in genomics [8], or in signal processing [6]. However, existing covariance shrinkage methods use as prior a single shrinkage target, which seems modest compared to the information provided by the large cohorts of modern population neuro-imaging.

To inform better the estimation of a subject's functional connectivity, we propose a covariance shrinkage that integrates a probabilistic distribution of the covariances calculated from a prior population. The resulting estimator shrinks toward the population mean, but additionally accounting for the population dispersion, hence with a non-isotropic shrinkage, [9] proposed a similar approach with a prior based regularization of the empirical covariance. Such approach relies on the population mean only and discards the population dispersion. A challenge is that covariance matrices must be positive definite and are distributed on a Riemannian manifold [10, 11]. To derive efficient shrinkage rules that respect this intrinsic geometry we leverage a tangent-space representation of the manifold in our shrinkage model. Local Euclidean approximation of the Riemannian geometry enables mean-square-error estimation that can be controlled well. Such approach has been successfully applied for functional-connectivity estimation [2] and classification [12]. The resulting model is validated on r-fMRI scans of 498 healthy subjects from the Human Connectome Project (HCP) dataset [13]. Experimental results show that the proposed estimator gives functional-connectivity matrices that generalize better and capture better subject phenotypes.

2 Covariance shrinkage towards a prior distribution

We propose a covariance estimator that takes into account a prior representing the population distribution. Fig.1-(a) depicts an overview of the method. Input data are time-series extracted from r-fMRI scans on ROIs defined from a brain atlas. The proposed method relies on three key elements: *i)* estimating a prior distribution for covariances over a reference dataset; *ii)* building a tangent embedding of the manifold of covariance matrices that enables the use of the minimum mean squared error (MMSE) framework; *iii)* shrinking subject covariance according to the prior in the tangent space. We introduce in the following subsections the mathematical basis of the method.
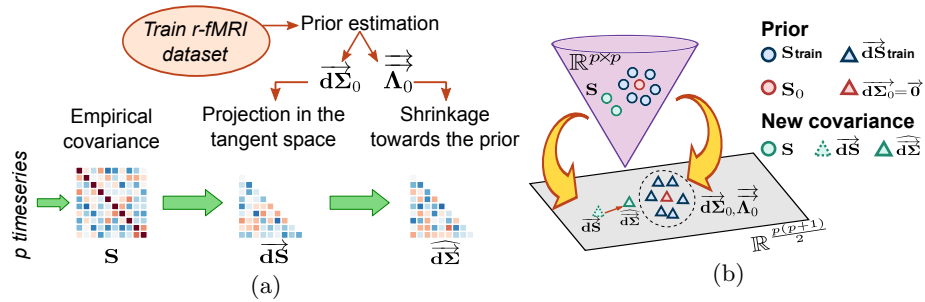
Fig. 1. (a) Shrunk embedding estimation workflow: the empirical covariance is estimated from r-fMRI time-series; it is projected onto a tangent space built from a prior population; the embedding is then shrunk towards the prior $(\overrightarrow{\mathbf{d\Sigma}}_0, \overrightarrow{\mathbf{\Lambda}}_0)$. (b) Principle of tangent embedding shrinkage towards population distribution.

Notations n and p denote the number of time-points and ROIs, respectively. We use boldface uppercase letters for matrices. We write $\rightarrow$ for vectors and $\Rightarrow$ for matrices in tangent space.

2.1 Tangent space embedding of the geometry of covariances

We model a subject's r-fMRI time series as drawn from a Gaussian distribution: $\mathbf{X} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$. For centered data the mean μ is 0. The covariance $\boldsymbol{\Sigma}$ captures functional connectivity. It lives on the positive definite cone. As pictured in Fig.1-(b), this cone can be seen as a Riemannian manifold endowed with an affine-invariant metric well suited to invariances of the Gaussian model [2, 14]. The manifold can be projected onto a vector space where Euclidean distances locally approximate Riemannian distances in the manifold. We use as a reference point the population average covariance across subjects, $\boldsymbol{\Sigma}_0$, and project $\boldsymbol{\Sigma}$ onto the corresponding tangent space $\mathbb{R}^{p \times p}$. The tangent-space vector $\mathbf{d\Sigma} \in \mathbb{R}^{p \times p}$ is then:

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

where $\frac{1}{2}$ denotes the matrix square-root and logm is the matrix logarithm. We adopt a convenient parametrization $\overrightarrow{\mathbf{d\Sigma}} \in \mathbb{R}^d$ with $d = p(p+1)/2$ where $\overrightarrow{\mathbf{d\Sigma}} = \{\sqrt{2} \, d\sigma_{i,j}, j < i, d\sigma_{i,i}, i = 1...p\}$. As the population average covariance $\boldsymbol{\Sigma}_0$, we simply use the Euclidean mean, since it yields more stable estimations compared to the Fréchet mean, as mentioned in [12]. An interesting property is that the ℓ_2 distance in the tangent space approximates the Kullback-Leibler (KL) divergence between two connectivity distribution. Hence it is possible to use second-order statistics to minimize the KL-divergence loss. This property will be used later for the shrinkage estimation.

2.2 Building the prior from the population distribution

We build a prior from a population-level model of the distribution of subjects functional-connectivity matrices. For this, we consider a set of covariances $\mathbf{S}_i$ from an r-fMRI dataset. First we compute a tangent-space embedding of this dataset by setting the reference $\boldsymbol{\Sigma}_0$ as the population empirical mean. We then model the vectors in the resulting tangent embedding as drawn from a normal distribution $\overrightarrow{\mathbf{d\Sigma}} \sim \mathcal{N}(\overrightarrow{\mathbf{d\Sigma}}_0 = \overrightarrow{0}, \overrightarrow{\boldsymbol{\Lambda}}_0)$, characterized by the mean and the covariance of the population. This distribution will be used as a prior for optimal shrinkage. Its covariance $\overrightarrow{\boldsymbol{\Lambda}}_0$ measures the element-wise dispersion of connectivity matrices in the tangent space. Assuming a normal distribution on the manifold –highest entropy assumption– this dispersion is given by the mean outer product of the tangent embedding over the train set, as mentioned in [14]:

$$\overrightarrow{\boldsymbol{\Lambda}}_0 = \frac{1}{N_{\mathrm{train}} - 1} \sum_{i=1}^{N_{\mathrm{train}}} \overrightarrow{\mathbf{dS}}_i \otimes \overrightarrow{\mathbf{dS}}_i, \tag{2}$$

where $\overrightarrow{\boldsymbol{\Lambda}}_0 \in \mathbb{R}^{d \times d}$ with $d = p(p+1)/2$. In practice, $\overrightarrow{\boldsymbol{\Lambda}}_0$ is very high dimensional and is learned from a finite population of subjects. We use instead a low-rank approximation as a regularization with a PCA decomposition: $\overrightarrow{\boldsymbol{\Lambda}}_* = \alpha \mathbf{I} + \mathbf{D}\mathbf{D}^T$, where α is set such that the explained variance ratio is above 70%.

2.3 Estimating the shrunk covariance as a posterior mean

We describe here how we use the prior distribution $(\overrightarrow{\mathbf{d\Sigma}}_0, \overrightarrow{\boldsymbol{\Lambda}}_0)$ for optimal shrinkage of $\overrightarrow{\mathbf{d\Sigma}}$ in tangent space. To derive the shrinkage from the prior distribution on $\overrightarrow{\mathbf{d\Sigma}}$, we rely on a Bayesian formulation: $p(\overrightarrow{\mathbf{d\Sigma}}|\overrightarrow{\mathbf{DS}}) \propto p(\overrightarrow{\mathbf{DS}}|\overrightarrow{\mathbf{d\Sigma}})p(\overrightarrow{\mathbf{d\Sigma}})$, where $\overrightarrow{\mathbf{DS}}$ is the embedded empirical covariance, and $p(\overrightarrow{\mathbf{d\Sigma}})$ the prior from the population. For $p(\overrightarrow{\mathbf{DS}}|\overrightarrow{\mathbf{d\Sigma}})$, the natural loss on covariances would be the KL divergence between Gaussian models. However, as mentioned earlier, we can use the quadratic loss that approximates it in the tangent space. The posterior mean –conditional on the data– then gives the minimum mean squared error (MMSE) estimator for $\overrightarrow{\mathbf{d\Sigma}}$ [15, Corollary 4.1.2.]:

$$\widehat{\overrightarrow{\mathbf{d\Sigma}}} = \mathbb{E}\big[\overrightarrow{\mathbf{d\Sigma}}|\overrightarrow{\mathbf{DS}}\big] = \underset{\overrightarrow{\mathbf{d\Sigma}}}{\operatorname{argmin}} \operatorname{MSE}(\overrightarrow{\mathbf{d\Sigma}}, \overrightarrow{\mathbf{DS}}). \tag{3}$$

To compute this expectancy we use:

- The prior established before: $p(\overrightarrow{\mathbf{d\Sigma}}) = \mathcal{N}(\overrightarrow{\mathbf{d\Sigma}}_0, \overrightarrow{\boldsymbol{\Lambda}}_0)$ where $(\overrightarrow{\mathbf{d\Sigma}}_0, \overrightarrow{\boldsymbol{\Lambda}}_0)$ are the estimated prior population distribution mean and covariance, respectively.
- The quadratic loss for the likelihood of the observed data $\overrightarrow{\mathbf{DS}}$ given $\overrightarrow{\mathbf{d\Sigma}}$: a Gaussian distribution centered on $\overrightarrow{\mathbf{dS}}$, the tangent-space projection of the empirical covariance, with covariance $\overrightarrow{\boldsymbol{\Lambda}}$: $p(\overrightarrow{\mathbf{DS}}|\overrightarrow{\mathbf{d\Sigma}}) = \mathcal{N}(\overrightarrow{\mathbf{dS}}, \overrightarrow{\boldsymbol{\Lambda}})$. $\overrightarrow{\boldsymbol{\Lambda}}$ cannot be fully estimated from limited data, hence we take $\overrightarrow{\boldsymbol{\Lambda}} = \lambda \mathbf{I}$, where λ acts as a shrinkage control parameter.

Using Bayes rule for multivariate Gaussian distributions [16], the posterior is $p(\overrightarrow{\mathbf{d\Sigma}}|\overrightarrow{\mathbf{DS}}) = \mathcal{N}(\widehat{\overrightarrow{\mathbf{d\Sigma}}}, \overrightarrow{\mathbf{C}})$, where the posterior covariance is $\overrightarrow{\mathbf{C}}^{-1} = \overrightarrow{\boldsymbol{\Lambda}}^{-1} + \overrightarrow{\boldsymbol{\Lambda}}_0^{-1}$, and the posterior mean $\widehat{\overrightarrow{\mathbf{d\Sigma}}}$ is:

$$\widehat{\overrightarrow{\mathbf{d\Sigma}}} = \left(\overrightarrow{\boldsymbol{\Lambda}}^{-1} + \overrightarrow{\boldsymbol{\Lambda}}_0^{-1}\right)^{-1} \left(\overrightarrow{\boldsymbol{\Lambda}}^{-1}\overrightarrow{\mathbf{dS}} + \overrightarrow{\boldsymbol{\Lambda}}_0^{-1}\overrightarrow{\mathbf{d\Sigma}}_0\right). \tag{4}$$

Since the mean of the prior $\overrightarrow{\mathbf{d\Sigma_0}}$ is null and approximating the prior covariance $\overrightarrow{\boldsymbol{\Lambda}}_0$ results in $\overrightarrow{\boldsymbol{\Lambda}}_*$, we have:

$$\text{Shrunk embedding:} \quad \widehat{\overrightarrow{\mathbf{d\Sigma}}} = (\overrightarrow{\boldsymbol{\Lambda}}^{-1} + \overrightarrow{\boldsymbol{\Lambda}}_*^{-1})^{-1}\overrightarrow{\boldsymbol{\Lambda}}^{-1}\overrightarrow{\mathbf{dS}}. \tag{5}$$

We observe that Eq.(5) is a generalization of classic shrinkage estimators [5, 8] that relies on a convex combination of a prior with the empirical covariance matrix. Here, the shrinkage is in the tangent space and the amount of shrinkage is controlled by the likelihood covariance parameter λ. In our experiments, we set λ with a cross-validation on a subset of the train dataset. Using $\overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0 \propto \mathbf{Id}$ recovers standard covariance shrinkage rules.

## 3 Experimental validation: shrunk embedding on HCP

The proposed covariance estimator is evaluated through systematic comparisons with state-of-the-art approaches. We assess in our experiments: *i)* the fidelity of the estimates across two sessions of the same subject; *ii)* the estimator capacity to characterize functional connectivity similarities between twins, siblings, and random subjects; *iii)* the relationship of the estimates to behavioral scores variations across the HCP population using canonical correlation analysis (CCA).

The HCP dataset. We use r-fMRI scans from the 500-subjects release of the HCP dataset: 498 healthy subjects including twins, siblings, and non-siblings. Each subject has two 15 min-long r-fMRI sessions comprising 1 200 time-points. In addition to the imaging modalities, each subject has a rich phenotype description and behavioral assessments. The HCP r-fMRI data are already spatially preprocessed and normalized to the MNI space. For a given brain atlas, we extract r-fMRI timeseries, detrended them, regressed out motion parameters, and band filter (0.01-0.1Hz) using the `Nilearn` library.

The prior embedding distribution $(\overrightarrow{\mathbf{d\Sigma}}_0, \overrightarrow{\overrightarrow{\mathbf{\Lambda}}}_0)$ is built on half of the HCP subjects ($n = 249$) randomly selected, and by keeping twins and siblings in the same group, as advocated in [17]. The remaining set –249 subjects as well– is used to test the covariance estimators: their fidelity to subject data, how well they separate subjects, and how well they relate to behavior.

i) Intra-subject shrunk embedding fidelity. To illustrate the benefits of estimating the connectivity with the shrunk embedding method, we assess the estimator fidelity by the log-likelihood of the data from a r-fMRI session –rest2– in a model estimated on a different session –rest1– from the same subject. For a model $\mathbf{\Sigma}$ estimated on rest1, the log-likelihood of data from rest2, characterized by its empirical covariance matrix $\mathbf{S}$, is: $1/2(-\text{tr}(\mathbf{S\Sigma}^{-1}) + \det(\mathbf{\Sigma}^{-1}) - p\log(2\pi))$. To compute this log-likelihood, we *back-project* the estimates from the tangent space to the actual covariance matrices. We compare five covariance models for each subject, either by shrinking the covariance towards the identity or the prior. The amount of the shrinkage is set through a cross-validation between rest1 and rest2. As there is no consensus on which set of ROIs should be used, we run experiments on three different brain atlases: MSDL atlas ($p = 39$), BASC atlas ($p = 64$), and Harvard-Oxford atlas ($p = 96$).

Figure 2 shows the log-likelihoods of each estimator, relative to the mean. The results demonstrate that shrinking the tangent embedding of the covariance

Intra-subject connectivity differences on HCP dataset
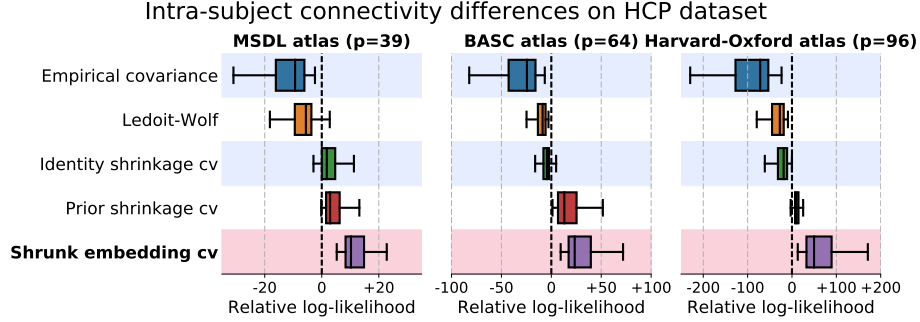
Fig. 2. Fidelity to subject data – Relative log-likelihoods of estimates across intra-subject sessions (rest1, rest2). Shrinking the covariance embedding towards the prior outperforms other estimators. Results are consistent across different brain atlases.

towards the prior produces the highest likelihood values. It suggests that integrating the population distribution efficiently regularizes connectivity components that are sensitive to inter-session variability, whereas shrinkage to identity is less optimal. There is a systematic gain when estimating the connectivity with shrunk embedding compared to only using the mean covariance as target of the shrinkage model, as in [9]. We also observe that shrinkage estimates generalize better than the empirical covariance estimator, and the optimal shrinkage is better estimated with cross-validation than with the Ledoit-Wolf method. Finally, the results are consistent across all brain atlases.

ii) Shrunk embedding highlights twins similarities. We assess the shrunk embedding estimator capacity to map inter-subject differences. For this, we compare distances of 20 twin-to-twin, with 20 twin-to-sibling, and with 20 twin-to-non-sibling. We use Mahalanobis distance: $\sqrt{(\widehat{\vec{\mathbf{d\Sigma}}_i} - \widehat{\vec{\mathbf{d\Sigma}}_j})\vec{\mathbf{\Lambda}}_0^{-1}(\widehat{\vec{\mathbf{d\Sigma}}_i} - \widehat{\vec{\mathbf{d\Sigma}}_j})}$, where $\widehat{\vec{\mathbf{d\Sigma}}_i}$ and $\widehat{\vec{\mathbf{d\Sigma}}_j}$ are the shrunk embedding estimates of two subjects i and j respectively, using BASC atlas.

Fig. 3. Separating unrelated subjects – Comparing connectivity distances between twins, siblings, and others in HCP dataset. Population covariance with the shrunk embedding estimator better distinguishes functional connectivity-based similarities between twins compared with non-twins ($*$: $p < 0.1$, two-sample t-test).

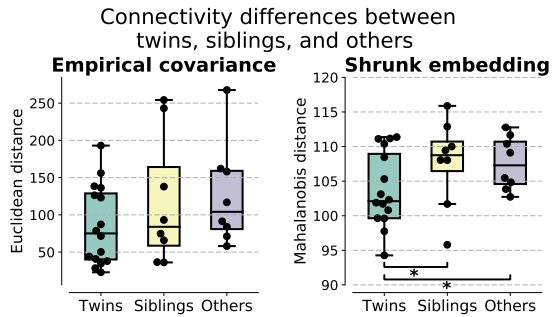Connectivity differences between twins, siblings, and others

Figure 3 shows the distribution of the distances for the three groups with the shrunk embedding and the empirical covariance. The shrunk embedding-based distances of twins are smaller than siblings and non-siblings, whereas the empirical covariance-based distance has more spread distances and less between-group differences. Even though the sample size is relatively small and calls for further validation, these results highlight the potential of using the population dispersion as a prior to characterize phenotypical and behavioral differences.
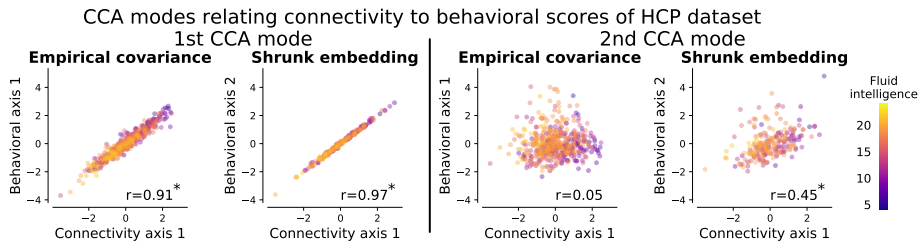
iii) Shrunk embedding association to behavioral scores. We investigate the relationship between HCP population behavioral assessments and their respective functional connectivity profiles. Following [1], we apply CCA with 100 components on 158 selected behavioral scores to relate them to 2 080 shrunk embedding features estimated on BASC atlas. The significance of the CCA modes is assessed through a permutations test with 10 000 iterations.

Results show two statistically significant CCA modes as depicted in Fig. 4 ($p < 10^{-4}$). While only one CCA mode is revealed by using the empirical covariance –as in[1]– the shrunk embedding yields two CCA modes with significant covariations between the functional connectivity and the behavioral assessments. The representation of the fluid intelligence value of each subject helps to visualize its correlation with the connectivity measures.

4 Conclusion

We introduced a covariance model that integrates population distribution knowledge for optimal shrinkage of the covariance. It combines the tangent space embedding representation of covariance matrices with a Bayesian estimate for the shrinkage. Compared to existing covariance shrinkage estimator, our contribution leverages additional prior information –the dispersion of a reference population of covariances– for non isotropic shrinkage. It gives rise to simple closed-form equations, and is thus suitably fast for large cohorts.

For brain functional connectivity, the proposed shrunk embedding model produces better estimation of connectivity matrices on the HCP dataset. It reduces

Fig. 4. Relating to behavior – Subjects distribution over CCA modes relating behavioral scores and connectivity estimates. Shrunk embedding gives two significantly correlated modes while empirical covariance only one ($*$:$p < 10^{-4}$, permutation test).

intra-subject variability and highlights more accurately co-variations between connectivity profiles and subjects behavioral assessments.

Further analysis of statistical properties could determine a *minimax* choice of the shrinkage amount that minimize the worst-case error for our estimator. Future work in brain imaging calls for more study of the generality of the population prior, for instance across distinct datasets. Our group-level analysis results show that the shrunk embedding captures better connectivity-phenotype covariation. It should next be used to build connectivity-based predictive models, predicting neurological or psychiatric disorders and health outcomes from clinical r-fMRI data.

References

1. Smith, S.M., Nichols, T.E., et al.: A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat Neurosci (2015)
2. Varoquaux, G., et al.: Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In: MICCAI. (2010)
3. Smith, S.M., Miller, K.L., et al.: Network modelling methods for FMRI. Neuroimage **54** (2011) 875
4. Varoquaux, G., Gramfort, A., et al.: Brain covariance selection: better individual functional connectivity models using population prior. In: NIPS. (2010) 2334
5. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis **88**(2) (2004) 365–411
6. Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O.: Shrinkage algorithms for MMSE covariance estimation. IEEE Transactions on Signal Processing **58** (2010) 5016
7. Brier, M.R., Mitra, A., et al.: Partial covariance based functional connectivity computation using Ledoit–Wolf covariance regularization. NeuroImage (2015)
8. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology **4**(1) (2005)
9. Crimi, A., et al.: Maximum a posteriori estimation of linear shape variation with application to vertebra and cartilage modeling. IEEE Trans on Med Imag (2011)
10. Lenglet, C., Rousson, M., Deriche, R., Faugeras, O.: Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. Journal of Mathematical Imaging and Vision **25**(3) (2006) 423–444
11. Fletcher, P.T., Joshi, S.: Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing **87**(2) (2007) 250–262
12. Ng, B., Dressler, M., et al.: Transport on riemannian manifold for functional connectivity-based classification. In: MICCAI. (2014) 405–412
13. Van Essen, D.C., Smith, S.M., et al.: The WU-minn human connectome project: An overview. NeuroImage **80** (2013) 62–79
14. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal of Computer Vision **66**(1) (2006) 41–66
15. Lehmann, E.L., Casella, G.: Theory of point estimation. (2006)
16. Bishop, C.M.: Pattern recognition and machine learning. (2006)
17. Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., Smith, S.M.: Multilevel block permutation. NeuroImage **123** (2015) 253–268