Supplementary materials for paper: Revealing Common Statistical Behaviors in Heterogeneous Populations

Andrey Zhitnikov Rotem Mulayoff Tomer Michaeli

This supplementary document contains:

1. Proofs for Lemmas 1 and 2;
2. An alternative common-covariance estimation algorithm, which can be more efficient on parallel platforms, and an empirical comparison to Alg. 1;
3. Resting-state fMRI experiments on a group of ADHD subjects, and a comparison to graph-Lasso covariance estimation;
4. Analysis of the effect of kernel bandwidth on the pdf estimation in the ABP-PPG experiment.

1 Proof of Lemma 1

For convenience we write again the definitions

$$\psi(\boldsymbol{q}) \triangleq \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{q}, \tag{1}$$

$$g_j(\boldsymbol{q}) \triangleq \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \boldsymbol{q}, \tag{2}$$

$$h_m(\boldsymbol{q}) \triangleq \min_{j\in\{1,\ldots,m\}} g_j(\boldsymbol{q}), \tag{3}$$

and the assumptions of Theorem 1, which are

$$\mathbb{P}\left(\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{v}}) \leq \alpha\right) = 1 \tag{4}$$

for some $\alpha > 0$ and

$$\mathbb{P}\left(\boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{q} \leq \epsilon\right) > 0 \tag{5}$$

for every $\epsilon > 0$ and every unit-norm $\boldsymbol{q}$. Since $\{g_j(\boldsymbol{q})\}$ and $\{h_m(\boldsymbol{q})\}$ are random functions, we will sometimes explicitly write $g_j(\omega, \boldsymbol{q})$ and $h_m(\omega, \boldsymbol{q})$, where ω denotes an element from the sample space Ω. We begin by demonstrating the first part of the lemma, namely that

$$h_m(\omega, \boldsymbol{q}) \overset{\text{a.s.}}{\to} 0 \tag{6}$$

for any $\boldsymbol{q}$. To this end, fix $\boldsymbol{q}$ and $\omega \in \Omega$. By definition, the sequence $\{h_m(\omega, \boldsymbol{q})\}_{m=1}^{\infty}$ is monotonically non-increasing and bounded from below, namely $h_m(\omega, \boldsymbol{q}) \geq h_{m+1}(\omega, \boldsymbol{q}) \geq 0$ for every m. Therefore, this sequence converges for every ω. Almost sure convergence implies convergence in probability to the same limit. Therefore, to prove (6), we will show that $h_m(\omega, \boldsymbol{q}) \overset{\text{p}}{\to} 0$, namely that

$$\lim_{m\to\infty} \mathbb{P}\left(|h_m(\boldsymbol{q})| > \epsilon\right) \to 0 \tag{7}$$

for every $\epsilon > 0$. Note that $g_j(\boldsymbol{q}) \geq 0$ for every j, so that $h_m(\boldsymbol{q}) \geq 0$ for every m, and thus the absolute value in (7) can be omitted. Since the random variables $\{g_j(\boldsymbol{q})\}$ are independent and identically distributed, we have that

$$\mathbb{P}(h_m(\boldsymbol{q}) > \epsilon) = \mathbb{P}\left(\cap_{j=1}^{m}\{g_j(\boldsymbol{q}) > \epsilon\}\right) = \mathbb{P}\left(\{g_1(\boldsymbol{q}) > \epsilon\}\right)^m = \left(1 - \mathbb{P}(g_1(\boldsymbol{q}) \leq \epsilon)\right)^m \underset{m\to\infty}{\longrightarrow} 0, \tag{8}$$

where we used (5). This completes the proof of (6). Next, we prove the second part of the Lemma, namely that

$$h_m(\omega, \boldsymbol{q}_m) \overset{\text{a.s.}}{\to} 0 \tag{9}$$

for any converging sequence $\{q_m\}$. To this end, we will first prove that all the functions $\{h_m(\boldsymbol{q})\}$ are Lipschitz w.p. 1. Note from (4) and the definition of $g_j(\boldsymbol{q})$ in (2) that each $g_j(\boldsymbol{q})$ is Lipschitz w.p. 1. Therefore, all the functions $\{g_j(\boldsymbol{q})\}_{j=1}^m$ are simultaneously Lipschitz w.p. 1, with the same Lipschitz constant 2α (a countable union of a.s. events). Fix any $\omega \in \Omega$ within this a.s. event. The function $h_1(\omega, \boldsymbol{y})$ is Lipschitz since it equals $g_1(\omega, \boldsymbol{y})$. Let us show this also holds for $h_2(\omega, \boldsymbol{y})$. For any two vectors $\boldsymbol{x}, \boldsymbol{y}$, we have that

$$g_1(\omega, \boldsymbol{y}) \geq g_1(\omega, \boldsymbol{x}) - |g_1(\omega, \boldsymbol{x}) - g_1(\omega, \boldsymbol{y})| \geq h_2(\omega, \boldsymbol{x}) - 2\alpha\|\boldsymbol{x} - \boldsymbol{y}\|, \tag{10}$$

where we used (3). Similarly, we have that $g_2(\omega, \boldsymbol{y}) \geq h_2(\omega, \boldsymbol{x}) - 2\alpha\|\boldsymbol{x} - \boldsymbol{y}\|$. Therefore,

$$h_2(\omega, \boldsymbol{y}) = \min\{g_1(\omega, \boldsymbol{y}), g_2(\omega, \boldsymbol{y})\} \geq h_2(\omega, \boldsymbol{x}) - 2\alpha\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{11}$$

This implies that $h_2(\omega, \boldsymbol{x}) - h_2(\omega, \boldsymbol{y}) \leq 2\alpha\|\boldsymbol{x} - \boldsymbol{y}\|$, and by switching the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$, we also obtain

$$|h_2(\omega, \boldsymbol{x}) - h_2(\omega, \boldsymbol{y})| \leq 2\alpha\|\boldsymbol{x} - \boldsymbol{y}\|, \tag{12}$$

demonstrating that $h_2(\omega, \boldsymbol{q})$ is Lipschitz with the same constant 2α. In a similar way, we obtain that $h_m(\omega, \boldsymbol{q})$ is Lipschitz for every m. We are now ready to prove (9). We saw that all $\{h_m(\omega, \boldsymbol{q})\}$ are Lipschitz a.s. and that $|h_m(\omega, \boldsymbol{q})| \to 0$ a.s. for any $\boldsymbol{q}$. The intersection of these two events is also a.s. Fix any ω in this intersection. Then

$$|h_m(\omega, \boldsymbol{q}_m)| = |h_m(\omega, \boldsymbol{q}_m) - h_m(\omega, \boldsymbol{q}^*) + h_m(\omega, \boldsymbol{q}^*)| \leq 2\alpha\|\boldsymbol{q}_m - \boldsymbol{q}^*\| + |h_m(\omega, \boldsymbol{q})| \underset{m\to\infty}{\longrightarrow} 0. \tag{13}$$

This completes the proof.

2 Proof of Lemma 2

The functions $\{\phi(\boldsymbol{q}) + f_n(\boldsymbol{q})\}$ are continuous and $\mathcal{C}$ is a compact set, therefore $\arg\min_{\boldsymbol{q}\in\mathcal{C}}\{\phi(\boldsymbol{q}) + f_n(\boldsymbol{q})\} \neq \emptyset$. We have to prove that for every $\epsilon > 0$ there exists an N_ϵ such that

$$\|\boldsymbol{q}_n - \boldsymbol{q}^*\| < \epsilon \quad \forall n \geq N_\epsilon. \tag{14}$$

Fix $\epsilon > 0$ and define the set $\mathcal{Q} = \{\boldsymbol{q} \in \mathcal{C} : \|\boldsymbol{q} - \boldsymbol{q}^*\| \geq \epsilon\}$. If the set $\mathcal{Q}$ is empty, then (14) is obviously satisfied with $N_\epsilon = 1$ and the proof is complete. If $\mathcal{Q} \neq \emptyset$, then $\phi(\boldsymbol{q})$ attains a minimum over the set $\mathcal{Q}$. This follows from the fact that $\mathcal{Q}$ is compact and $\phi(\boldsymbol{q})$ is continuous over the set $\mathcal{Q} \subset \mathcal{C}$. Let

$$\tilde{\boldsymbol{q}} \in \underset{\boldsymbol{q}\in\mathcal{Q}}{\arg\min}\, \phi(\boldsymbol{q}) \tag{15}$$

and denote $\delta(\epsilon) = \phi(\tilde{\boldsymbol{q}}) - \phi(\boldsymbol{q}^*)$. From

$$\lim_{n\to\infty} f_n(\boldsymbol{q}^*) = 0, \tag{16}$$

we know that there exists an N_ϵ such that $0 \leq f_n(\boldsymbol{q}^*) < \delta(\epsilon)$ for every $n \geq N_\epsilon$. Therefore, for every $n \geq N_\epsilon$, we have that

$$\phi(\boldsymbol{q}^*) + f_n(\boldsymbol{q}^*) < \phi(\boldsymbol{q}^*) + \delta(\epsilon) = \phi(\tilde{\boldsymbol{q}}) \underbrace{\leq}_{f_n\geq 0} \min_{\boldsymbol{q}\in\mathcal{Q}}[\phi(\boldsymbol{q}) + f_n(\boldsymbol{q})]. \tag{17}$$

This demonstrates that for every $n > N_\epsilon$ there exists no point at distance larger than ϵ from $\boldsymbol{q}^*$, at which the function $\phi(\boldsymbol{q}) + f_n(\boldsymbol{q})$ attains a lower value than at $\boldsymbol{q}^*$. In other words, the minimum of $\phi(\boldsymbol{q}) + f_n(\boldsymbol{q})$ must be attained at a point $\boldsymbol{q}_n$, which satisfies (14). This completes the proof.

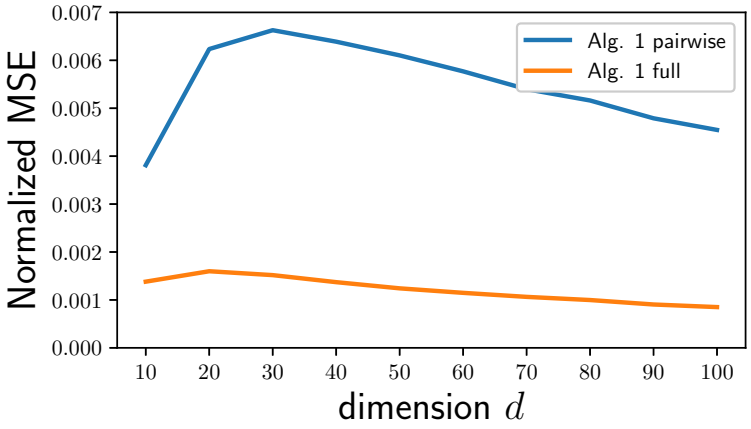

Figure 1: **Common covariance estimation with the full and pairwise versions of Alg. 1 at high SNR.** We used 100 noisy covariance matrices to estimate the underlying common covariance, using Alg. 1 in the full and pairwise form. In this simulation, SNR = 1. Approximation of the MSE is based on 100 trials.

3 Accelerated covariance estimation

To speed up the estimation of the common covariance on parallel platforms, one can break the task of estimating the full $d \times d$ covariance matrix $\mathbf{\Sigma_u}$ into $d \times (d-1)/2$ sub-tasks of estimating all 2×2 sub-covariance-matrices of $\mathbf{\Sigma_u}$. Those sub-tasks can be solved in parallel, while averaging the $d-1$ different estimates obtained for the diagonal entries. This procedure is not guaranteed to yield a positive semidefinite estimate $\hat{\mathbf{\Sigma}}_{\boldsymbol{u}}$, so that negative eigenvalues have to be truncated. However, this method enjoys the same asymptotic guarantees as the direct estimation of the full covariance.

To study the effect of this approach on the estimation accuracy for a finite number of subjects, we next compare between Alg. 1 and this pairwise version in a simulation. We take the common covariance to be the identity matrix $\mathbf{\Sigma_u} = \boldsymbol{I}$ and generate the noise covariance according to equation (27) of the main text, where now $\mathbf{\Lambda}_j = \mathrm{diag}\{\beta_1^j \dots \beta_d^j\}$ with

$$\beta_k^j \sim U[0, b] \quad \forall j = 1, \dots, m \; \forall k = 1, \dots, d. \tag{18}$$

For the rotation matrices, we first draw a matrix $\tilde{\boldsymbol{M}}_j$ with iid entries uniformly distributed on the interval $[-50, 50]$. Then, the matrix $\boldsymbol{M}_j$ is taken to be the $\boldsymbol{Q}$ matrix from the $\boldsymbol{QR}$ decomposition of $\tilde{\boldsymbol{M}}_j$. Figures 1 and 2 depict the MSE attained by both algorithms (normalized by the number of entries in the estimated matrix, d^2), as a function of the dimension d, for two different SNR levels. As can be seen, in all the tested settings, Alg. 1 in its full form gives more accurate results. We conclude that the potential improvement in running time offered by this approach, typically comes at the cost of reduced estimation accuracy.

4 Comparison of common covariance matrices of Control subjects and subjects with ADHD

We repeated the analysis of Sec. 5.3 in the main text on a group of 141 subjects diagnosed with ADHD. In addition to our estimator and the Riemann mean and Euclidean mean estimators, we also show results obtained with the graph Lasso estimator (Friedman et al., 2008) applied to all data points from all the subjects. This estimator assumes a sparse precision matrix and selects its L_1 regularization weight through cross-validation. As can be seen in Figs. 3 and 4 below, our estimator consistently detects higher correlations within known networks (around the main diagonal).

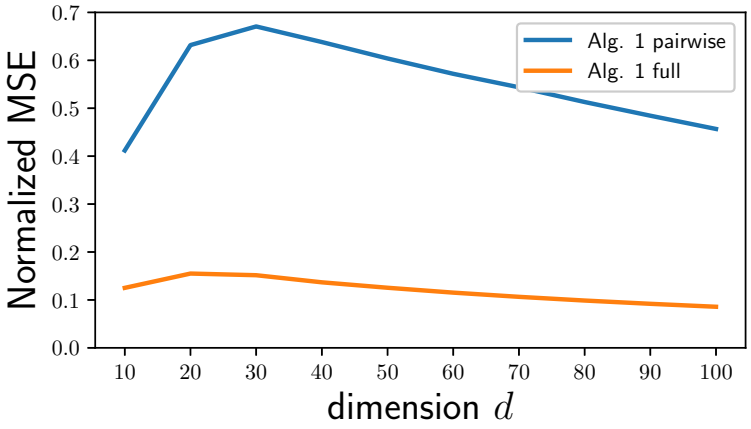

Figure 2: **Common covariance estimation with the full and pairwise versions of Alg. 1 at low SNR.** We used 100 noisy covariance matrices to estimate the underlying common covariance, using Alg. 1 in the full and pairwise form. In this simulation, SNR = 0.1. Approximation of the MSE is based on 100 trials.

5 The effect of the kernel bandwidth in naive KDE and in our common density estimator

In Sec. 5.4 of the main text, we demonstrated the ability of our common pdf estimator to reveal delicate structures that are not seen with naive KDE applied on all data points from all patients. To show that this phenomenon is not a matter of appropriate selection of bandwidth in the KDE, we now repeat this experiment with several bandwidths, ranging from ones leading to over-smoothness to ones leading to under-smoothness. As can be seen in Fig. 5 below, there exists no bandwidth with which the vertical trace is seen in the KDE estimate. At the same time, this trace is seen with all bandwidths in our pdf estimate.

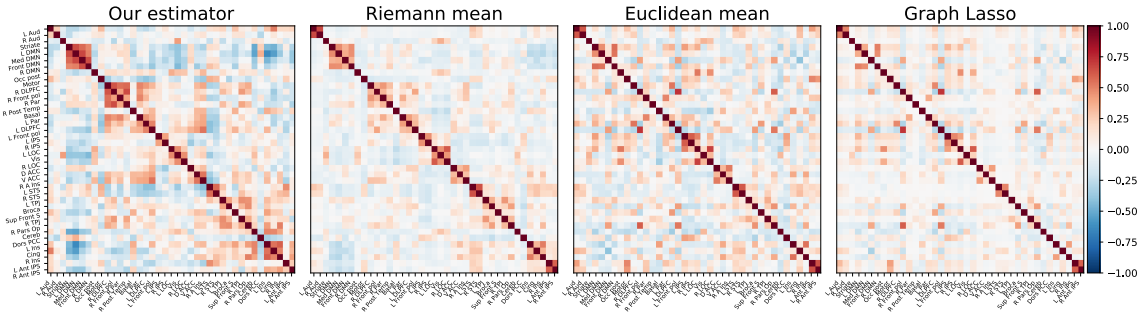

Figure 3: Comparison of group level correlation matrices of control subjects (same experiment as in the main text).

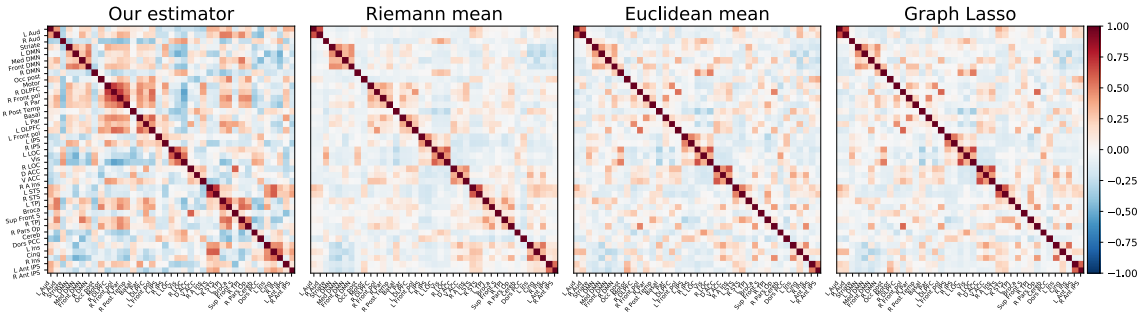

Figure 4: Comparison of group level correlation matrices of ADHD subjects.

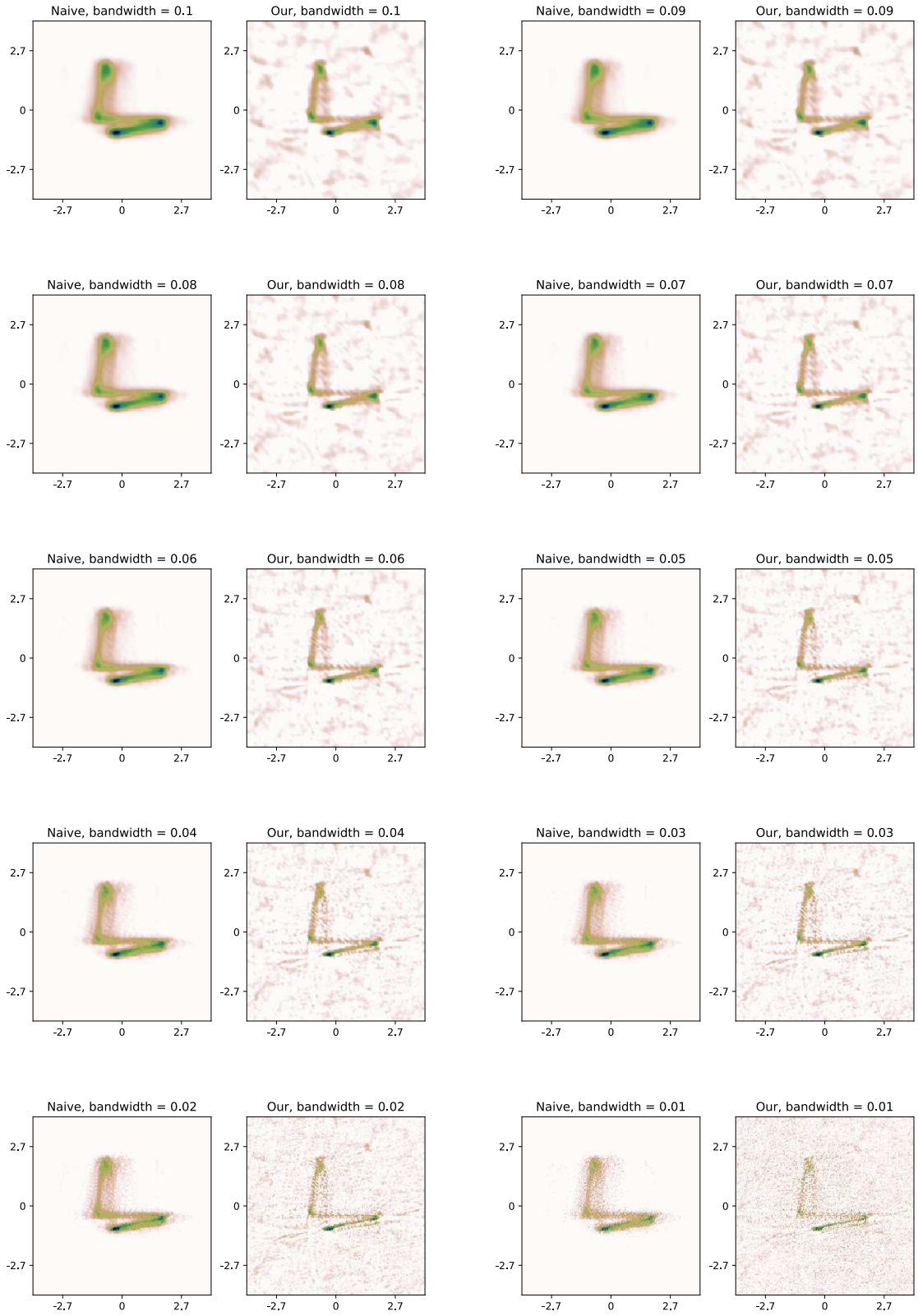

Figure 5: Common probability density estimation of PPG and ABP with a bandwidths ranging from 0.1 (top left) to 0.01 (bottom right).