

Joint Estimation of High-Dimensional Spiked Covariance Matrices via a Partially Shared Subspace

Changwon Yoon¹, Minwoo Kim², **Sungkyu Jung**², and Jeongyoun Ahn¹

¹KAIST ²Seoul National University

Abstract. High-dimensional covariance estimation is intrinsically difficult when the number of variables is comparable to, or larger than, the sample size. At the same time, many applications provide an auxiliary data set measured on the same variables and related to the target data through only part of its covariance structure. Pooling the two samples can then reduce variance but may induce substantial bias when dataset-specific directions are present. We propose a *partially shared subspace* (PSS) model and estimation framework that transfers only the covariance information supported by a common signal subspace while retaining distinct structure in each data source.

For independent target and background samples with spiked covariance matrices, after noise-level normalization we write

$$\Sigma_X = Q_S \Lambda_S Q_S^\top + Q_D \Lambda_D Q_D^\top + I_p, \quad \Sigma_Y = U_S \Gamma_S U_S^\top + U_D \Gamma_D U_D^\top + I_p,$$

where $\text{span}(Q_S) = \text{span}(U_S)$ is an unknown shared subspace of rank r_S , while Q_D and U_D contain dataset-specific spikes. The shared spikes may occur at arbitrary spectral positions and need not correspond to the leading eigenvectors. This formulation strictly generalizes standard common-principal-component and shared-subspace models while allowing the two data sets to have different total spiked ranks and unrelated distinct components.

We work in the proportional-growth regime $p/n_X \rightarrow c_X$ and $p/n_Y \rightarrow c_Y$. Random matrix theory implies that sample spiked eigenvectors are not consistent, but their alignments converge to deterministic limits. In particular, for population eigenvectors q_i, u_j and their sample counterparts $\hat{q}_i, \hat{u}_j$,

$$|\langle \hat{q}_i, \hat{u}_j \rangle| \xrightarrow{a.s.} \phi_{c_X}(\lambda_i) \phi_{c_Y}(\gamma_j) |\langle q_i, u_j \rangle|, \quad \phi_c(\ell) = \sqrt{\frac{1 - c/(\ell - 1)^2}{1 + c/(\ell - 1)}}.$$

This yields a principal-angle characterization of the shared subspace through the singular values of $\hat{Q}^\top \hat{U}$. We develop an iterative estimator that alternates between estimating r_S via a random-matrix-theory threshold and identifying the shared eigenvector indices through alignment with a pooled projection. The corresponding oracle rank and index estimators are consistent under a separation condition between the shared and distinct components.

After estimating the shared subspace, we combine the two samples through a weighted projection

$$\hat{P}_S^{(\alpha)} = \Pi_{\hat{r}_S} \left\{ \alpha \hat{P}_X + (1 - \alpha) \hat{P}_Y \right\}, \quad \hat{P}_X = \hat{Q} \hat{Q}^\top, \quad \hat{P}_Y = \hat{U} \hat{U}^\top.$$

An asymptotic Frobenius-loss calculation gives the approximate optimal weight

$$\alpha^* = \frac{1}{2} + \frac{L_Y - L_X}{2(r_S - T)},$$

so the amount of transfer depends on both sample sizes and spike strengths rather than on sample size alone. The target covariance is then reconstructed by projecting its spiked component onto the estimated shared subspace and its orthogonal complement. This construction also produces a principled estimator of the target-specific subspace, providing a high-dimensional solution to contrastive dimension reduction.

Simulation studies show accurate recovery of the shared rank and uniformly favorable target-covariance estimation across leading or interleaved shared spikes, non-orthogonal distinct subspaces, and mild model misspecification. In an S&P 500 application, using pre-pandemic returns as background improves minimum-variance portfolio construction during the early COVID-19 period; with $p = 499$, $n_X = 61$, and $n_Y = 378$, the method estimates nine shared directions and attains the lowest out-of-sample volatility among the considered estimators. In a second application to low-grade glioma RNA-seq data, using normal brain tissue as background yields a distinct subspace that cleanly separates astrocytoma and oligodendroglioma samples, whereas the shared subspace and ordinary PCA do not. These examples demonstrate that partial structural transfer can improve high-dimensional covariance estimation while preserving scientifically meaningful dataset-specific variation.

Keywords: high-dimensional covariance estimation; spiked covariance model; shared subspace; transfer learning; random matrix theory; contrastive dimension reduction.