Principal Component Analysis: Basic Results and Proofs

(Instructor Name)

November 6, 2025

1 Setup and Notation

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be observations of a random vector $X \in \mathbb{R}^d$. We assume throughout this handout that the data have been centered, i.e.

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = 0.$$

The *empirical covariance matrix* is

$$\Sigma_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \in \mathbb{R}^{d\times d}.$$

We denote by $\|\cdot\|_2$ the Euclidean norm and by $\langle \cdot, \cdot \rangle$ the associated inner product.

Let $\Sigma_n = V\Lambda V^\top$ be an eigendecomposition of Σ_n, with

$$V = [v_1, \dots, v_d] \in \mathbb{R}^{d\times d} \quad \text{orthogonal}, \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d),$$

and eigenvalues ordered as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

2 PCA as Variance Maximization

The classical one-dimensional PCA direction is defined as the unit vector $u \in \mathbb{R}^d$ that maximizes the sample variance of the projections $\langle u, X_i \rangle$.

Definition 2.1 (Empirical variance along a unit direction)**.** For $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$, the empirical variance of the projected data $u^\top X_i$ is

$$\widehat{\operatorname{Var}}(u^\top X) := \frac{1}{n}\sum_{i=1}^{n} (u^\top X_i)^2.$$

Lemma 2.2. *For any $u \in \mathbb{R}^d$,*

$$\widehat{\operatorname{Var}}(u^\top X) = u^\top \Sigma_n u.$$

Proof. Since the data are centered, $\frac{1}{n}\sum_{i=1}^{n} u^\top X_i = 0$. Hence

$$\widehat{\operatorname{Var}}(u^\top X) = \frac{1}{n}\sum_{i=1}^{n} (u^\top X_i)^2 = \frac{1}{n}\sum_{i=1}^{n} u^\top X_i X_i^\top u = u^\top \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \right) u = u^\top \Sigma_n u.$$

□

Thus the first principal component solves

$$\max_{u\in\mathbb{R}^d} u^\top \Sigma_n u \quad \text{subject to} \quad \|u\|_2 = 1.$$

Theorem 2.3 (Rayleigh quotient and first principal component)**.** *Let $\Sigma_n = V\Lambda V^\top$ as above. Then:*

1. *For any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,*
$$\lambda_1 \geq u^\top \Sigma_n u \geq \lambda_d.$$

2. *The maximum λ_1 is attained when $u = v_1$ (or any vector in the span of eigenvectors associated with λ_1).*

Proof. Write u in the eigenbasis: $u = Vz$ with $z \in \mathbb{R}^d$, $\|z\|_2 = \|u\|_2 = 1$. Then

$$u^\top \Sigma_n u = (Vz)^\top V\Lambda V^\top (Vz) = z^\top \Lambda z = \sum_{j=1}^{d} \lambda_j z_j^2.$$

Since $\lambda_1 \geq \cdots \geq \lambda_d$ and $\sum_j z_j^2 = 1$, we have

$$\lambda_d \leq \sum_{j=1}^{d} \lambda_j z_j^2 \leq \lambda_1.$$

The upper bound is attained by taking $z = e_1$ (first coordinate vector), i.e. $u = v_1$, and the lower bound by $z = e_d$, i.e. $u = v_d$. □

For higher principal components, we add orthogonality constraints. The k-th principal component direction u_k is defined as a maximizer of $u^\top \Sigma_n u$ over unit vectors u orthogonal to $u_1, \ldots, u_{k-1}$. This is again obtained as $u_k = v_k$.

3 Empirical Reconstruction Error and Trace Form

Let $U \in \mathbb{R}^{d\times k}$ be a matrix with orthonormal columns, $U^\top U = I_k$. The orthogonal projector onto the subspace $\mathcal{S} = \mathrm{span}(U)$ is $P_U := UU^\top$. The reconstruction of a data point X_i through this subspace is $P_U X_i$, and the reconstruction error is $X_i - P_U X_i$.

Definition 3.1 (Empirical reconstruction risk)**.** The empirical reconstruction risk of a subspace represented by U is

$$\widetilde{R}_n(U) := \frac{1}{2n} \sum_{i=1}^{n} \|X_i - UU^\top X_i\|_2^2.$$

The factor $1/2$ is conventional and simplifies later expressions.

Proposition 3.2 (Reconstruction risk as constant minus trace)**.** *For any $U \in \mathbb{R}^{d\times k}$ with $U^\top U = I_k$,*

$$\widetilde{R}_n(U) = C_n - \frac{1}{2}\mathrm{Tr}(U^\top \Sigma_n U),$$

where

$$C_n := \frac{1}{2n} \sum_{i=1}^{n} \|X_i\|_2^2$$

is independent of U.

Proof. Fix U and denote $P = UU^\top$. For each i,

$$\|X_i - PX_i\|_2^2 = \|X_i\|_2^2 + \|PX_i\|_2^2 - 2\langle X_i, PX_i\rangle.$$

Since P is an orthogonal projector ($P^2 = P$, $P^\top = P$),

$$\|PX_i\|_2^2 = \langle PX_i, PX_i\rangle = \langle X_i, P^\top PX_i\rangle = \langle X_i, PX_i\rangle.$$

Hence

$$\|X_i - PX_i\|_2^2 = \|X_i\|_2^2 - \langle X_i, PX_i \rangle.$$

Using trace notation,

$$\langle X_i, PX_i \rangle = X_i^\top P X_i = \mathrm{Tr}(X_i^\top P X_i) = \mathrm{Tr}(P X_i X_i^\top) = \mathrm{Tr}(U^\top X_i X_i^\top U).$$

Therefore

$$\|X_i - UU^\top X_i\|_2^2 = \|X_i\|_2^2 - \mathrm{Tr}(U^\top X_i X_i^\top U).$$

Summing and dividing by $2n$,

$$\begin{aligned}
\widetilde{R}_n(U) &= \frac{1}{2n} \sum_{i=1}^n \|X_i\|_2^2 - \frac{1}{2n} \sum_{i=1}^n \mathrm{Tr}(U^\top X_i X_i^\top U) \\
&= C_n - \frac{1}{2} \mathrm{Tr}\left(U^\top \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right) U \right) \\
&= C_n - \frac{1}{2} \mathrm{Tr}(U^\top \Sigma_n U).
\end{aligned}$$

□

As a direct corollary:

Corollary 3.3 (Reconstruction error minimization ⇔ trace maximization)**.**

$$\arg \min_{U^\top U = I_k} \widetilde{R}_n(U) = \arg \max_{U^\top U = I_k} \mathrm{Tr}(U^\top \Sigma_n U).$$

4 The Block Rayleigh Quotient and Top-k Eigenvectors

We now study the maximization problem

$$\max_{U \in \mathbb{R}^{d \times k}} \mathrm{Tr}(U^\top \Sigma U) \quad \text{subject to } U^\top U = I_k,$$

where Σ is a symmetric positive semi-definite matrix. For the empirical case, simply take $\Sigma = \Sigma_n$.

Theorem 4.1 (Block Rayleigh quotient)**.** *Let $\Sigma = V \Lambda V^\top$ be a symmetric positive semi-definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Then:*

1. *For any $U \in \mathbb{R}^{d \times k}$ with $U^\top U = I_k$,*

$$\mathrm{Tr}(U^\top \Sigma U) \leq \sum_{j=1}^k \lambda_j.$$

2. *Equality is attained if and only if the column space of U is an invariant subspace spanned by eigenvectors associated with $\lambda_1, \ldots, \lambda_k$, i.e.*

$$\mathrm{span}(U) = \mathrm{span}\{v_1, \ldots, v_k\}.$$

Proof. Write U in the eigenbasis of Σ: $U = VB$, where $B \in \mathbb{R}^{d \times k}$. Since V is orthogonal and $U^\top U = I_k$,

$$I_k = U^\top U = B^\top V^\top V B = B^\top B.$$

Therefore the columns of B are orthonormal.

Compute

$$\mathrm{Tr}(U^\top \Sigma U) = \mathrm{Tr}\Big((VB)^\top V\Lambda V^\top (VB)\Big) = \mathrm{Tr}(B^\top \Lambda B).$$

Let $b_i^\top$ denote the i-th row of B, so $b_i \in \mathbb{R}^k$ and $B = (b_{ij})_{i,j}$. Since Λ is diagonal,

$$B^\top \Lambda B = \sum_{i=1}^{d} \lambda_i \, b_i b_i^\top,$$

and thus

$$\mathrm{Tr}(B^\top \Lambda B) = \sum_{i=1}^{d} \lambda_i \, \mathrm{Tr}(b_i b_i^\top) = \sum_{i=1}^{d} \lambda_i \, \|b_i\|_2^2.$$

Next, use the constraint $B^\top B = I_k$. Taking traces,

$$\mathrm{Tr}(B^\top B) = \mathrm{Tr}(I_k) = k.$$

On the other hand,

$$\mathrm{Tr}(B^\top B) = \sum_{i=1}^{d} \|b_i\|_2^2.$$

Thus

$$\sum_{i=1}^{d} \|b_i\|_2^2 = k. \tag{1}$$

We also have $0 \le \|b_i\|_2^2 \le 1$ for each i. Indeed, $G := BB^\top$ is a projection matrix onto the column space of B, hence G is positive semi-definite and $G^2 = G$. In particular its diagonal entries satisfy $0 \le G_{ii} \le 1$. But

$$G_{ii} = e_i^\top BB^\top e_i = \|b_i\|_2^2.$$

We therefore need to maximize

$$S := \sum_{i=1}^{d} \lambda_i \|b_i\|_2^2$$

subject to

$$0 \le \|b_i\|_2^2 \le 1 \quad \text{and} \quad \sum_{i=1}^{d} \|b_i\|_2^2 = k.$$

Since $\lambda_1 \ge \cdots \ge \lambda_d$, the value of S is maximized by allocating as much of the "mass" k as possible on the largest eigenvalues, i.e. on indices $i = 1, \ldots, k$. Because each $\|b_i\|_2^2 \le 1$, at most 1 unit of mass can be placed on each index. Hence the optimal allocation is

$$\|b_i\|_2^2 = \begin{cases} 1, & i = 1, \ldots, k, \\ 0, & i = k+1, \ldots, d. \end{cases}$$

For this allocation,

$$S_{\max} = \sum_{i=1}^{k} \lambda_i.$$

More formally, for any feasible $(\|b_i\|_2^2)$,

$$S = \sum_{i=1}^{d} \lambda_i \|b_i\|_2^2 \le \sum_{i=1}^{k} \lambda_i \|b_i\|_2^2 + \lambda_{k+1} \sum_{i=k+1}^{d} \|b_i\|_2^2 \le \lambda_1 \sum_{i=1}^{k} \|b_i\|_2^2 + \lambda_{k+1}(k - \sum_{i=1}^{k} \|b_i\|_2^2),$$

and using $\sum_{i=1}^{d} \|b_i\|_2^2 = k$ gives

$$S \leq \sum_{i=1}^{k} \lambda_i,$$

with equality only if $\|b_i\|_2^2 = 1$ for $i \leq k$ and $\|b_i\|_2^2 = 0$ for $i > k$.

In this case, B has zero rows from $k+1$ to d, and the first k rows form an orthonormal system. Then

$$U = VB$$

has its columns lying entirely in $\text{span}(v_1, \ldots, v_k)$. Conversely, any U whose column space is $\text{span}(v_1, \ldots, v_k)$ attains the maximum. □

Combining Proposition 3.2 and Theorem 4.1, we recover the standard PCA characterization:

Corollary 4.2 (Empirical PCA subspace)**.** *Any $U_n \in \mathbb{R}^{d\times k}$ with $U_n^\top U_n = I_k$ and*

$$\text{span}(U_n) = \text{span}\{v_1, \ldots, v_k\}$$

minimizes the empirical reconstruction risk $\widetilde{R}_n(U)$ among all U with $U^\top U = I_k$.

5 Population PCA and Risk Representation

Let $X \in \mathbb{R}^d$ be a random vector with mean $m = \mathbb{E}[X]$ and covariance

$$\Sigma := \mathbb{E}\big[(X-m)(X-m)^\top\big].$$

Definition 5.1 (Population reconstruction risk)**.** For $U \in \mathbb{R}^{d\times k}$ with $U^\top U = I_k$, define

$$R(U) := \frac{1}{2}\mathbb{E}\Big[\|X - m - UU^\top(X-m)\|_2^2\Big].$$

Proposition 5.2 (Population risk as constant minus trace)**.** *For any $U \in \mathbb{R}^{d\times k}$ with $U^\top U = I_k$,*

$$R(U) = C - \frac{1}{2}\text{Tr}(U^\top \Sigma U),$$

where $C = \frac{1}{2}\mathbb{E}\big[\|X-m\|_2^2\big]$ is independent of U.

Proof. Let $Y := X - m$. Then $\mathbb{E}[Y] = 0$ and $\Sigma = \mathbb{E}[YY^\top]$. For fixed U, denote $P = UU^\top$ as before. Then

$$R(U) = \frac{1}{2}\mathbb{E}\left[\|Y - PY\|_2^2\right].$$

We can repeat the same argument as in Proposition 3.2 with expectations in place of empirical means:

$$\|Y - PY\|_2^2 = \|Y\|_2^2 - Y^\top PY.$$

Taking expectations,

$$R(U) = \frac{1}{2}\mathbb{E}\|Y\|_2^2 - \frac{1}{2}\mathbb{E}[Y^\top PY].$$

The first term is C. For the second term,

$$\mathbb{E}[Y^\top PY] = \mathbb{E}[\text{Tr}(PYY^\top)] = \text{Tr}\big(P\mathbb{E}[YY^\top]\big) = \text{Tr}(U^\top \Sigma U),$$

using linearity of trace and expectation. Hence

$$R(U) = C - \frac{1}{2}\text{Tr}(U^\top \Sigma U).$$

□

Corollary 5.3 (Population PCA subspace)**.** *Let $\Sigma = U\Lambda U^\top$ be an eigendecomposition with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. Any matrix $U^* \in \mathbb{R}^{d\times k}$ with orthonormal columns satisfying*

$$\mathrm{span}(U^*) = \mathrm{span}\{u_1, \dots, u_k\}$$

minimizes $R(U)$ over all U with $U^\top U = I_k$.

Proof. By Proposition 5.2, minimizing $R(U)$ is equivalent to maximizing $\mathrm{Tr}(U^\top \Sigma U)$, and we can apply Theorem 4.1 with Σ in place of Σ_n. □

6 Dependence on the Subspace Only

Proposition 6.1 (Invariance under change of basis)**.** *Let $U, V \in \mathbb{R}^{d\times k}$ satisfy $U^\top U = V^\top V = I_k$. Then the following are equivalent:*

1. $\mathrm{span}(U) = \mathrm{span}(V)$*;*
2. *there exists an orthogonal matrix $Q \in \mathbb{R}^{k\times k}$ such that $V = UQ$;*
3. $UU^\top = VV^\top$*.*

Moreover, if any of the above holds, then

$$R(U) = R(V) \quad \text{and} \quad \widetilde{R}_n(U) = \widetilde{R}_n(V).$$

Proof. (1) ⇒ (2): If $\mathrm{span}(U) = \mathrm{span}(V)$ and both U and V have orthonormal columns, then the columns of V are orthonormal vectors in the span of the columns of U. Thus there exists an orthogonal matrix Q such that $V = UQ$.

(2) ⇒ (3): If $V = UQ$ with $Q^\top Q = I_k$, then

$$VV^\top = UQQ^\top U^\top = UU^\top.$$

(3) ⇒ (1): If $UU^\top = VV^\top =: P$, then P is an orthogonal projector, and its image is precisely the subspace onto which it projects. Thus

$$\mathrm{span}(U) = \mathrm{Im}(P) = \mathrm{span}(V).$$

For the risk equality, note that $R(U)$ and $\widetilde{R}_n(U)$ only depend on $P = UU^\top$. If $UU^\top = VV^\top$, then clearly $R(U) = R(V)$ and $\widetilde{R}_n(U) = \widetilde{R}_n(V)$. □

Remark 6.2. This proposition justifies viewing PCA as choosing a *subspace* rather than a specific orthonormal basis. The natural parameter space is the Grassmann manifold of k-dimensional subspaces in $\mathbb{R}^d$, rather than the Stiefel manifold of $d \times k$ orthonormal matrices.

7 Principal Angles and Subspace Distance

We now introduce principal angles between subspaces and relate them to a natural distance between projectors. This connects the geometric error of PCA to matrix norms.

Definition 7.1 (Principal angles)**.** Let $U, V \in \mathbb{R}^{d\times k}$ have orthonormal columns and let

$$U^\top V = S \in \mathbb{R}^{k\times k}.$$

Let $\sigma_1 \geq \cdots \geq \sigma_k \geq 0$ be the singular values of S. The *principal angles* between the subspaces $\mathcal{U} = \mathrm{span}(U)$ and $\mathcal{V} = \mathrm{span}(V)$ are defined by

$$\theta_j := \arccos(\sigma_j), \quad j = 1, \dots, k.$$

Remark 7.2. We have $0 \le \sigma_j \le 1$, hence $\theta_j \in [0, \pi/2]$. The principal angles are symmetric in $(\mathcal{U}, \mathcal{V})$ and do not depend on the particular orthonormal bases U and V.

A common way to measure the distance between subspaces is to look at the difference of their orthogonal projectors.

Definition 7.3 (Projection distance)**.** Let $P_U = UU^\top$ and $P_V = VV^\top$ be the orthogonal projectors onto $\mathcal{U}$ and $\mathcal{V}$, respectively. The *projection-Frobenius distance* between $\mathcal{U}$ and $\mathcal{V}$ is

$$d_{\mathrm{proj}}(\mathcal{U}, \mathcal{V}) := \|P_U - P_V\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm.

The next proposition makes the link to principal angles explicit.

Proposition 7.4 (Projectors and principal angles)**.** *With notation as above, we have*

$$\|P_U - P_V\|_F^2 = 2\sum_{j=1}^{k} \sin^2\theta_j = 2k - 2\|U^\top V\|_F^2.$$

Proof. First, note that

$$\|P_U - P_V\|_F^2 = \mathrm{Tr}\big((P_U - P_V)^\top (P_U - P_V)\big) = \mathrm{Tr}(P_U^2) + \mathrm{Tr}(P_V^2) - 2\mathrm{Tr}(P_U P_V),$$

using $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and the symmetry of P_U, P_V.

Since $P_U^2 = P_U$ and $P_V^2 = P_V$, and both are rank-k projectors,

$$\mathrm{Tr}(P_U^2) = \mathrm{Tr}(P_U) = k, \qquad \mathrm{Tr}(P_V^2) = \mathrm{Tr}(P_V) = k.$$

Moreover,

$$\mathrm{Tr}(P_U P_V) = \mathrm{Tr}(UU^\top VV^\top) = \mathrm{Tr}(U^\top VV^\top U) = \mathrm{Tr}\big((U^\top V)(U^\top V)^\top\big) = \|U^\top V\|_F^2.$$

Therefore

$$\|P_U - P_V\|_F^2 = k + k - 2\|U^\top V\|_F^2 = 2k - 2\|U^\top V\|_F^2.$$

Now write the singular value decomposition of $U^\top V$:

$$U^\top V = Q\Sigma R^\top,$$

where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_k)$ contains the singular values. The Frobenius norm is unitarily invariant, hence

$$\|U^\top V\|_F^2 = \|\Sigma\|_F^2 = \sum_{j=1}^{k} \sigma_j^2.$$

Substituting,

$$\|P_U - P_V\|_F^2 = 2k - 2\sum_{j=1}^{k} \sigma_j^2 = 2\sum_{j=1}^{k} (1 - \sigma_j^2) = 2\sum_{j=1}^{k} \sin^2\theta_j,$$

since $\sigma_j = \cos\theta_j$. □

Remark 7.5. The quantity

$$\sqrt{\sum_{j=1}^{k} \sin^2\theta_j}$$

is sometimes called the *chordal distance* between subspaces. Proposition 7.4 shows that, up to a factor $\sqrt{2}$, this is exactly the Frobenius norm of the difference of projectors. This makes it convenient to use matrix norms when analyzing the geometric error of PCA.

8 A Davis–Kahan Type Perturbation Bound

We now state a basic version of the Davis–Kahan $\sin\Theta$ theorem, which controls the distance between invariant subspaces of two symmetric matrices in terms of a spectral gap and the size of the perturbation. In the PCA setting, this links the population and empirical principal subspaces.

8.1 Statement of the Theorem

Let Σ be the population covariance and Σ_n the empirical covariance. Assume Σ has eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_d.$$

Let $U^* \in \mathbb{R}^{d\times k}$ collect the top-k eigenvectors of Σ, and let $U_n \in \mathbb{R}^{d\times k}$ collect the top-k eigenvectors of Σ_n.

Define the matrix of principal angles between the subspaces $\mathcal{U}^* = \operatorname{span}(U^*)$ and $\mathcal{U}_n = \operatorname{span}(U_n)$ as follows: let $\Theta \in \mathbb{R}^{k\times k}$ be diagonal with entries $\theta_1, \dots, \theta_k$ the principal angles. We use the notation

$$\sin\Theta := \operatorname{diag}(\sin\theta_1, \dots, \sin\theta_k),$$

and define norms of $\sin\Theta$ by

$$\|\sin\Theta\|_2 = \max_j \sin\theta_j, \qquad \|\sin\Theta\|_F^2 = \sum_{j=1}^k \sin^2\theta_j.$$

Theorem 8.1 (Davis–Kahan $\sin\Theta$ theorem, simplified)**.** *Let Σ and Σ_n be symmetric matrices as above and let $\delta := \lambda_k - \lambda_{k+1} > 0$ be the eigengap. Then*

$$\|\sin\Theta\|_2 \leq \frac{\|\Sigma_n - \Sigma\|_2}{\delta},$$

and, consequently,

$$\|\sin\Theta\|_F \leq \sqrt{k}\,\|\sin\Theta\|_2 \leq \sqrt{k}\,\frac{\|\Sigma_n - \Sigma\|_2}{\delta}.$$

Here $\|\cdot\|_2$ is the spectral norm (largest singular value).

Remark 8.2. By Proposition 7.4, the Frobenius distance between projectors satisfies

$$\|U^*U^{*\top} - U_nU_n^\top\|_F^2 = 2\sum_{j=1}^k \sin^2\theta_j = 2\|\sin\Theta\|_F^2,$$

hence Theorem 8.1 also yields

$$\|U^*U^{*\top} - U_nU_n^\top\|_F \leq \sqrt{2k}\,\frac{\|\Sigma_n - \Sigma\|_2}{\delta}.$$

8.2 Proof Sketch

A full proof requires a few linear algebra identities; we outline the main ideas.

Proof sketch. Let $P^* = U^*U^{*\top}$ and $P_n = U_nU_n^\top$ be the projectors onto the top-k eigenspaces of Σ and Σ_n, respectively. Assume for simplicity that Σ and Σ_n are diagonalizable in orthonormal bases, which they are as symmetric matrices.

Step 1: Block decomposition. Choose an orthonormal basis in which

$$\Sigma = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix},$$

where $\Lambda_1 \in \mathbb{R}^{k\times k}$ contains $\lambda_1, \dots, \lambda_k$ and $\Lambda_2 \in \mathbb{R}^{(d-k)\times(d-k)}$ contains $\lambda_{k+1}, \dots, \lambda_d$. In this basis,

$$P^* = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}.$$

Write, in the same basis,

$$P_n = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}.$$

The off-diagonal block B describes how much the empirical top-k subspace "leans" into the orthogonal complement; it is closely related to $\sin\Theta$.

Step 2: SinΘ and projectors. One can show that

$$\|\sin\Theta\|_2 = \|(I - P^*)P_n\|_2 = \|B^\top\|_2 = \|B\|_2.$$

Intuitively, $(I - P^*)P_n$ projects first onto the empirical subspace and then onto the orthogonal complement of the true subspace, measuring the mismatch between them.

Step 3: Using the spectral gap. Consider the operator

$$\Sigma P_n - P_n\Sigma.$$

In the block basis above and using the form of Σ, one can compute this commutator explicitly and relate it to the block B. On the other hand,

$$\Sigma P_n - P_n\Sigma = (\Sigma - \Sigma_n)P_n + \Sigma_n P_n - P_n\Sigma_n + P_n(\Sigma_n - \Sigma).$$

But P_n is the spectral projector onto eigenvectors of Σ_n associated with the top-k eigenvalues, so $\Sigma_n P_n = P_n\Sigma_n$. Thus

$$\Sigma P_n - P_n\Sigma = (\Sigma - \Sigma_n)P_n + P_n(\Sigma_n - \Sigma).$$

In particular,

$$\|\Sigma P_n - P_n\Sigma\|_2 \le 2\|\Sigma_n - \Sigma\|_2.$$

On the other hand, using the block forms for Σ and P_n and the eigengap condition $\lambda_k > \lambda_{k+1}$, one can show that

$$\|\Sigma P_n - P_n\Sigma\|_2 \ge \delta\|B\|_2,$$

where $\delta = \lambda_k - \lambda_{k+1}$. Intuitively, moving mass between the top-k and bottom-$(d-k)$ subspaces costs at least δ in the commutator.

Combining the two inequalities gives

$$\delta\|B\|_2 \le \|\Sigma P_n - P_n\Sigma\|_2 \le 2\|\Sigma_n - \Sigma\|_2.$$

With a slightly more refined argument (or absorbing the factor 2 into δ via conventions on the spectral clusters), one obtains the bound

$$\|B\|_2 \le \frac{\|\Sigma_n - \Sigma\|_2}{\delta}.$$

Since $\|B\|_2 = \|\sin\Theta\|_2$, this is the desired result. □

Remark 8.3. The Davis–Kahan theorem is very general: it applies to any symmetric (or Hermitian) matrices and spectral subspaces separated by a gap, not only to covariance matrices. In PCA, it is a key tool to turn a matrix concentration bound on $\|\Sigma_n - \Sigma\|_2$ into a geometric error bound for the empirical principal subspace.

9 Matrix Concentration for the Sample Covariance

To turn Theorem 8.1 into a finite-sample PCA error bound, we need a high-probability bound on $\|\Sigma_n - \Sigma\|_2$. A standard assumption is sub-Gaussian tails.

Definition 9.1 (Sub-Gaussian random vector)**.** A random vector $X \in \mathbb{R}^d$ is *sub-Gaussian* with parameter $K \geq 0$ if for all $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$ and all $t \geq 0$,

$$\Pr\left(|u^\top X| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2K^2}\right).$$

Equivalently, for all such u,

$$\mathbb{E}\exp\left(\lambda\, u^\top X\right) \leq \exp\left(\frac{\lambda^2 K^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

We assume X has mean zero and covariance Σ.

Theorem 9.2 (Sample covariance concentration, sub-Gaussian case)**.** *Let $X_1, \dots, X_n$ be i.i.d. copies of a mean-zero sub-Gaussian random vector $X \in \mathbb{R}^d$ with parameter K and covariance matrix Σ. Let*

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top.$$

Then there exist universal constants $C, c > 0$ such that for all $t \geq 0$,

$$\Pr\left(\|\Sigma_n - \Sigma\|_2 \leq CK^2\|\Sigma\|_2\left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n}\right)\right) \geq 1 - 2\exp(-ct).$$

Remark 9.3.

- If $n \gtrsim d$, the dominant term is

$$\|\Sigma_n - \Sigma\|_2 = O_{\mathbb{P}}\left(K^2\|\Sigma\|_2\sqrt{\frac{d}{n}}\right).$$

- For isotropic X (i.e. $\Sigma = I_d$), this reduces to

$$\|\Sigma_n - I_d\|_2 \lesssim K^2\left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n}\right)$$

  with high probability.

Proof idea. One approach is to use an ε-net argument on the unit sphere S^{d-1} combined with Bernstein-type inequalities for quadratic forms $u^\top(\Sigma_n - \Sigma)u$; see, e.g., Vershynin's notes on non-asymptotic random matrix theory. Another approach is to invoke general matrix Bernstein inequalities for sums of independent random matrices $X_i X_i^\top - \Sigma$, together with sub-Gaussian tail control. We omit the detailed proof here. □

Combining Theorems 8.1 and 9.2 yields a clean finite-sample bound for PCA.

Corollary 9.4 (Finite-sample PCA subspace error, high probability)**.** *Under the assumptions of Theorems 8.1 and 9.2, let $\delta := \lambda_k - \lambda_{k+1} > 0$ be the eigengap of the population covariance Σ. Then there exist constants $C', c > 0$ such that for all $t \geq 0$,*

$$\Pr\left(\|U^* U^{*\top} - U_n U_n^\top\|_F \leq C' K^2 \sqrt{k}\,\frac{\|\Sigma\|_2}{\delta}\left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n}\right)\right) \geq 1 - 2\exp(-ct),$$

where U^* and U_n are the population and empirical k-PCA bases as before.

Proof. By Theorem 8.1,

$$\|U^* U^{*\top} - U_n U_n^\top\|_F \le \sqrt{2k}\,\frac{\|\Sigma_n - \Sigma\|_2}{\delta}.$$

Apply Theorem 9.2 to bound $\|\Sigma_n - \Sigma\|_2$ and absorb $\sqrt{2}$ into the constant C'. □

Remark 9.5. Corollary 9.4 gives the usual $O(\sqrt{kd/n})$ rate (up to log factors and constants) for the Frobenius error between the empirical and population principal subspaces in the sub-Gaussian setting, with explicit dependence on the eigengap δ and the scale $\|\Sigma\|_2$.