

Lecture 17: High-dimensional statistical phenomena

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Priority guide

- High** Core concepts and calculations to master; directly examined.
- Mid** Supporting concepts useful in other lectures; in scope for examinations.
- Low** Enrichment supporting understanding; will not be examined.
- Further reading** Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

Exam scope for this lecture. Section 3 (community detection) and its linked Exercises 4 and 5 are optional and will not be tested in exams. Their priority labels guide reading within this optional material. Other high- and mid-priority material remains in examination scope.

High priority *Overview*

Dimension changes the quality of familiar estimators. We use covariance error and **eigengaps** to study PCA, separate prediction risk into bias and variance on both sides of interpolation, and use eigenvectors to recover network communities. The main route develops PCA behavior as dimension grows and minimum-norm regression. The detailed interpolation-risk curves and all of Section 3 on community detection are optional self-study.

Recap from earlier lectures

From Lectures 1, 7, 8, 13 and 16: spectra and regression

Reminder only: this block introduces no new exam requirements.

- **Random matrices and covariance.** Operator-norm bounds control sample-covariance error; the dimension-to-sample-size ratio matters. Used in Section 1.
- **Weyl and PCA.** Eigenvalue perturbation controls spectral error; leading directions describe maximal projected variance. Used in Section 1.
- **Minimum-norm interpolation.** Among exact fits, the minimum-norm fit lies in the row space of the design. Used in Section 2.
- **OLS projection.** Least squares splits the observation vector into its fitted component and orthogonal residual. Used in Section 2.
- **Spectral relaxation.** Community assignments can be approximated by eigenvectors of a suitable adjacency matrix. Used in optional Section 3.

An asterisk in a topic heading marks optional material; the priority labels specify the reading scope.

Learning goals

- Distinguish fixed- and growing-dimension PCA regimes.
- Interpret the PCA phase transition and squared overlap.
- Compare regression risk below and above interpolation.
- Explore spectral community detection as optional material.

High priority *Introduction*

For convenience, we refer to the Euclidean norm as $\|\cdot\|$ and $\|\cdot\|_2$. As before, we use c or C to denote a positive absolute constant. We write c_K or C_K to denote a scalar that depends only on K . In particular, if K is a constant, then c_K and C_K are both constants. We denote by $\mathbb{S}^{n-1}$ the unit sphere: $\{x \in \mathbb{R}^n : \|x\| = 1\}$. The abbreviation psd means positive semidefinite.

1 Principal component analysis

High priority *Principal component analysis*

In this section, we examine a classical statistical method, principal component analysis (PCA), from a theoretical perspective. We will focus mostly on estimation consistency.

Suppose that $x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} X_0$ are our observed data where $\mathbb{E}[X_0] = 0$ and the second moments of X_0 exist. Denote $\Sigma = \text{Cov}(X_0)$. Let the eigen-decomposition of Σ be

$$\Sigma = \sum_{j=1}^p \lambda_j u_j u_j^\top$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ are in descending order by convention, and $u_1, u_2, \dots, u_p \in \mathbb{R}^p$ form an orthonormal basis. We form the empirical covariance¹ $\hat{\Sigma}$ and perform eigen-decomposition correspondingly

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top, \quad \hat{\Sigma} = \sum_{j=1}^p \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p \geq 0$ are in descending order by convention, and $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_p \in \mathbb{R}^p$ form an orthonormal basis. By definition, $u_1, u_2, \dots, u_p$ and $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_p$ are all unit vectors.

The vector u_1 is called the 1st population principal direction, and $x_1^\top u_1, x_2^\top u_1, \dots, x_n^\top u_1$ are called the 1st population principal component scores. Similarly, the vector $\hat{u}_1$ is called the 1st (empirical) principal direction, and $x_1^\top \hat{u}_1, x_2^\top \hat{u}_1, \dots, x_n^\top \hat{u}_1$ are called the 1st empirical principal component scores. PCA is useful for dimension reduction, visualization, clustering, etc.

Instructor's Note *The gap is the reference scale*

If the top two directions explain almost the same variance, a small perturbation can choose a very different winner. With a clear gap, the winning direction is stable. The sign is only a label for that direction: reversing an arrow does not change its axis.

¹For convenience, we assumed zero mean, so we do not need to center each observation by $x_i - \bar{x}$. If the mean is unknown, use the centered empirical covariance instead.

1.1 Fixed dimension

High priority *Fixed dimension*

In the classical regime, the dimension p is fixed, and PCA generally yields consistent estimators. We will prove a simple consistency result.

Theorem 1 (Consistency of PCA in fixed dimension). *Suppose that $\lambda_1 > \lambda_2$. Then as $n \rightarrow \infty$, we have*

$$\max_{k \leq p} |\hat{\lambda}_k - \lambda_k| \xrightarrow{p} 0, \quad \min_{s \in \{\pm 1\}} \|s\hat{u}_1 - u_1\|_2 \xrightarrow{p} 0.$$

Remark 1. *First, recall that eigenvectors are not unique for an eigenvalue: any $u = tu_1$ with $t \neq 0$ solves $\Sigma u = \lambda_1 u$, so both u_1 and $-u_1$ are unit top eigenvectors corresponding to λ_1 . This is why in the theorem statement, we have the freedom to choose a sign $s \in \{\pm 1\}$ when measuring the estimation error of the top principal direction². Sometimes this flexibility of choosing a sign is called “up to sign”.*

Second, if we don't assume $\lambda_1 > \lambda_2$, then the first eigenvalue may have multiplicity more than 1: $\lambda_1 = \lambda_2 = \dots = \lambda_k$. In this case, any vector $u = [u_1, \dots, u_k]a$ with $\|a\|_2 = 1$ is a top eigenvector; an orthogonal matrix $R \in O(k)$ rotates the entire basis $[u_1, \dots, u_k]R$ within that eigenspace. Every column of the rotated basis still satisfies $\Sigma u = \lambda_1 u$. In this case, we need to study the consistency of the top eigenspace $\text{span}\{u_1, \dots, u_k\}$ under an appropriate distance for subspaces.

Proof. By the LLN, we have

$$\hat{\Sigma} \xrightarrow{p} \Sigma$$

By Weyl's inequality,

$$\max_{k \leq p} |\hat{\lambda}_k - \lambda_k| \leq \|\hat{\Sigma} - \Sigma\|_{\text{op}},$$

so $\|\hat{\Sigma} - \Sigma\|_{\text{op}} \xrightarrow{p} 0$ implies $\max_{k \leq p} |\hat{\lambda}_k - \lambda_k| \xrightarrow{p} 0$. We can prove the consistency of the top eigenvector in a similar fashion, by using a well-known perturbation bound as stated below. $\square$

Theorem 2 (Davis–Kahan Sin- Θ Theorem: simple version). *Let $u_1, \tilde{u}_1$ be the first eigenvectors of two real symmetric matrices $\Sigma, \tilde{\Sigma} \in \mathbb{R}^{p \times p}$ respectively normalized such that $\|u_1\|_2 = \|\tilde{u}_1\|_2 = 1$. Here λ_1, λ_2 are the two largest eigenvalues of Σ . Assuming $\lambda_1 > \lambda_2$, we have the following deterministic result*

$$\min_{s \in \{\pm 1\}} \|s\tilde{u}_1 - u_1\|_2 \leq \frac{2\sqrt{2} \|\tilde{\Sigma} - \Sigma\|_{\text{op}}}{\lambda_1 - \lambda_2}.$$

Proof. If $\lambda_1 - \lambda_2 < 2\|\tilde{\Sigma} - \Sigma\|_{\text{op}}$, then we have trivial bound

$$\text{LHS} \leq \sqrt{2} \leq \sqrt{2} \cdot \frac{2\|\tilde{\Sigma} - \Sigma\|_{\text{op}}}{\lambda_1 - \lambda_2} = \text{RHS}.$$

In the proof below, we assume $\lambda_1 - \lambda_2 \geq 2\|\tilde{\Sigma} - \Sigma\|_{\text{op}}$. Denote the normalized eigenvectors of Σ by $u_1, u_2, \dots, u_p$ so that $U = [u_1, u_2, \dots, u_p]$ is an orthogonal matrix. As usual, we write $\langle u, v \rangle = u^\top v$ for vectors u, v .

We pick $s_0 \in \{\pm 1\}$ so that $s_0 = 1$ if $\langle \tilde{u}_1, u_1 \rangle \geq 0$ and $s_0 = -1$ otherwise. Observe

$$\begin{aligned} \min_{s \in \{\pm 1\}} \|s\tilde{u}_1 - u_1\|_2 &\leq \|s_0\tilde{u}_1 - u_1\|_2 = \sqrt{2 - 2\langle s_0\tilde{u}_1, u_1 \rangle} \\ &\leq \sqrt{2 - 2\langle s_0\tilde{u}_1, u_1 \rangle} \cdot \sqrt{1 + \langle s_0\tilde{u}_1, u_1 \rangle} \\ &= \sqrt{2} \sqrt{1 - \langle s_0\tilde{u}_1, u_1 \rangle^2} \\ &= \sqrt{2} \sqrt{1 - \langle \tilde{u}_1, u_1 \rangle^2}. \end{aligned}$$

²If we consider the equivalence classes in $\mathbb{S}^{p-1}$ with $u \sim -u$, then we have consistency for the distance $d([u], [v]) = \min_{s \in \{\pm 1\}} \|su - v\|_2$ between these classes.

Since $\tilde{u}_1$ is a unit vector and U is an orthogonal matrix,

$$1 = \|\tilde{u}_1\|_2^2 = \|U^\top \tilde{u}_1\|_2^2 = \langle u_1, \tilde{u}_1 \rangle^2 + \sum_{k=2}^p \langle u_k, \tilde{u}_1 \rangle^2.$$

It suffices to show

$$\sum_{k>1} \langle u_k, \tilde{u}_1 \rangle^2 \leq \frac{4 \|\tilde{\Sigma} - \Sigma\|_{\text{op}}^2}{(\lambda_1 - \lambda_2)^2}$$

By the definition of eigenvectors,

$$\begin{aligned} \tilde{\Sigma} \tilde{u}_1 &= \tilde{\lambda}_1 \tilde{u}_1 \\ \Sigma u_k &= \lambda_k u_k, \quad \text{for any } k. \end{aligned}$$

It follows that

$$\begin{aligned} \langle u_k, \tilde{\Sigma} \tilde{u}_1 \rangle &= \tilde{\lambda}_1 \langle u_k, \tilde{u}_1 \rangle \\ \langle \tilde{u}_1, \Sigma u_k \rangle &= \lambda_k \langle u_k, \tilde{u}_1 \rangle \quad \text{for any } k. \end{aligned}$$

Since $\tilde{\Sigma}$ and Σ are symmetric, taking the difference we find

$$\langle u_k, (\tilde{\Sigma} - \Sigma) \tilde{u}_1 \rangle = (\tilde{\lambda}_1 - \lambda_k) \langle u_k, \tilde{u}_1 \rangle. \quad (1)$$

By Weyl's inequality, $\tilde{\lambda}_1 - \lambda_1 \geq -\|\tilde{\Sigma} - \Sigma\|_{\text{op}}$, so for $k > 1$

$$\tilde{\lambda}_1 - \lambda_k \geq \tilde{\lambda}_1 - \lambda_2 \geq \lambda_1 - \lambda_2 - \|\tilde{\Sigma} - \Sigma\|_{\text{op}} \geq \frac{1}{2}(\lambda_1 - \lambda_2) \quad (2)$$

where we used $\lambda_1 - \lambda_2 \geq 2\|\tilde{\Sigma} - \Sigma\|_{\text{op}}$. Thus, using the inequality in (2) for (1), we have

$$\sum_{k>1} \langle u_k, \tilde{u}_1 \rangle^2 \leq \sum_{k>1} \frac{\langle u_k, (\tilde{\Sigma} - \Sigma) \tilde{u}_1 \rangle^2}{(\tilde{\lambda}_1 - \lambda_k)^2} \leq \frac{4 \|\tilde{\Sigma} - \Sigma\|_{\text{op}}^2 \|\tilde{u}_1\|_2^2}{(\lambda_1 - \lambda_2)^2} \leq \frac{4 \|\tilde{\Sigma} - \Sigma\|_{\text{op}}^2}{(\lambda_1 - \lambda_2)^2}$$

In the above derivation, we used basic inequalities

$$\sum_{k>1} \langle u_k, v \rangle^2 \leq \sum_{k=1}^p \langle u_k, v \rangle^2 = \|v\|_2^2$$

for any vector v and $\|Av\|_2 \leq \|A\|_{\text{op}} \|v\|_2$ (here v is unit). □

For the first PCA consistency assertion, finite second moments give integrability of every coordinate product by Cauchy–Schwarz. The LLN applies entrywise; in fixed dimension this implies operator-norm convergence. A fixed positive eigengap and the displayed perturbation bound then complete the eigenvector proof.

An eigengap stabilizes direction. A small operator perturbation gives a small top-eigenvector error, up to sign, when divided by a positive population eigengap. A shrinking gap can defeat consistency even if the covariance error tends to zero.

High priority *Practice: What does an eigengap control?*

Exercise prompt. For $\Sigma = \text{diag}(3, 1)$ and a symmetric perturbation E with $\|E\|_{\text{op}} \leq 0.1$, bound the top eigenvector error up to sign.

See Exercise 1 in Section 6 for the full statement and complete solution.

1.2 Growing dimension

High priority *Growing dimension*

Now, we consider the scenario where p is allowed to grow with n . Let us write $p = p_n$ and $\Sigma = \Sigma_n$ to emphasize the dependence on n . We need certain assumptions on the tail behavior of X_0 . For simplicity, assume that $X_0 \sim N(0, \Sigma_n)$, although it is possible to relax to subgaussian random vectors.

We should note that when p_n grows, the covariance matrix has p_n eigenvalues. We usually impose additional structure on Σ_n to facilitate theoretical analysis. One common structure is the *spiked model*:

$$\Sigma_n = (\lambda_1 - \lambda_2)u_1u_1^\top + \lambda_2 I_{p_n} \quad (3)$$

Under this assumption, the eigenvalues of Σ_n are λ_1 and λ_2 with multiplicity $p_n - 1$. To measure consistency, we use the usual Euclidean norm. Note

$$\frac{1}{2} \min_{s \in \{\pm 1\}} \|s\hat{u}_1 - u_1\|_2^2 = 1 - |\langle \hat{u}_1, u_1 \rangle|.$$

Inconsistency of PCA in high dimensions. We adopt the asymptotic view as before. We have

Theorem 3. Assume $\|\Sigma_n\|_{\text{op}} \leq C$ and $\lambda_1(\Sigma_n) - \lambda_2(\Sigma_n) \geq \kappa > 0$ holds for certain constants C and κ independent of n . Suppose that $p_n/n \rightarrow \gamma$ where $\gamma \in [0, \infty)$ is a constant.

- if $\gamma = 0$, then

$$\min_{s \in \{\pm 1\}} \|s\hat{u}_1 - u_1\|_2 \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty. \quad (\text{consistency of PCA})$$

- if $\gamma > 0$, then under (3) with fixed $0 < \lambda_2 < \lambda_1$, we have $|\langle \hat{u}_1, u_1 \rangle| \xrightarrow{P} \eta_\gamma$ where $\eta_\gamma < 1$. Moreover, there exists some threshold γ^* such that $\eta_\gamma = 0$ if $\gamma > \gamma^*$, namely [5]

$$|\langle \hat{u}_1, u_1 \rangle| \xrightarrow{P} 0, \quad (\text{inconsistency of PCA})$$

Beyond this threshold, the empirical and population principal directions are asymptotically orthogonal. Below it, a nonzero overlap can remain, even though consistency fails.

The first part of the theorem is a consequence of Theorem 2 in the previous lecture, which implies the following inequality holds with probability $1 - 2e^{-cp}$,

$$\|\hat{\Sigma}_n - \Sigma_n\|_{\text{op}} \leq C' \max \left\{ \sqrt{\frac{p}{n}}, \frac{p}{n} \right\} \|\Sigma\|_{\text{op}}$$

for a constant C' independent of n . Combining this bound with Davis–Kahan’s inequality, we get the consistency of $\hat{u}_1$ up to sign.

To make the probability tend to one even when p_n stays bounded or varies irregularly, use the tail-parameter version of the covariance bound with $t_n = \sqrt{\log n}$. Then both $\sqrt{p_n/n} + t_n/\sqrt{n}$ and the failure probability tend to zero. The uniform positive eigengap completes the argument.

1.3 Phase transition phenomenon of PCA

Mid priority *Phase transition phenomenon of PCA*

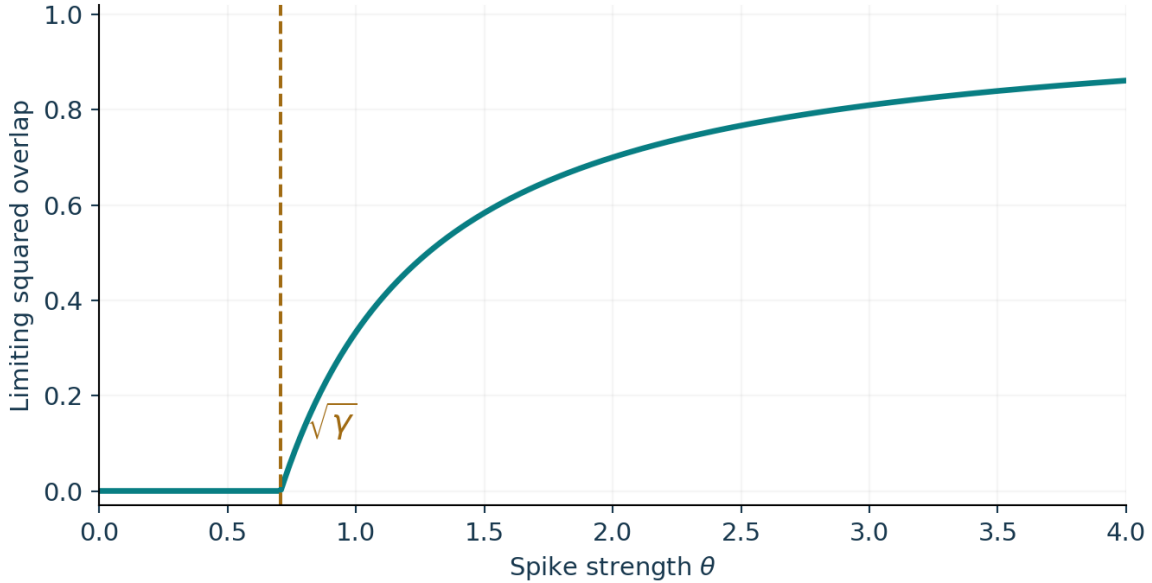
We present the precise eigenvalue and eigenvector limits for $0 < \gamma < 1$; see Paul (2007), Theorems 1, 2 and 4. Consider the spiked covariance model: let $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$, $x_i = \sqrt{\theta} \xi_i u + g_i$, where $\xi_i \sim \mathcal{N}(0, 1)$, $g_i \sim \mathcal{N}(0, I_p)$; $u = u_n \in \mathbb{S}^{p-1}$ is deterministic, $\theta > 0$ is fixed, and all (ξ_i, g_i) are mutually independent. Equivalently, we assume $x_i \sim \mathcal{N}(0, \theta u u^\top + I_p)$.

Using earlier notation, $\Sigma = \theta u u^\top + I_p$ and the eigengap is $\lambda_1 - \lambda_2 = \theta$. Intuitively, u is the shared signal direction, $\sqrt{\theta} \xi_i$ is its random amplitude in observation i , and g_i is noise.

Theorem 4. Denote $\hat{\Sigma} = \frac{X^\top X}{n}$, and assume $\frac{p_n}{n} \rightarrow \gamma \in (0, 1)$. Then as $n \rightarrow \infty$,

$$\lambda_1(\hat{\Sigma}) \xrightarrow{a.s.} \begin{cases} (\theta + 1)(1 + \frac{\gamma}{\theta}) & \text{if } \theta > \sqrt{\gamma} \\ (1 + \sqrt{\gamma})^2 & \text{if } \theta \leq \sqrt{\gamma}. \end{cases}$$

$$|\langle \hat{u}_1, u \rangle|^2 \xrightarrow{a.s.} \begin{cases} \frac{1 - \gamma/\theta^2}{1 + \gamma/\theta} & \text{if } \theta > \sqrt{\gamma} \\ 0 & \text{if } \theta \leq \sqrt{\gamma}. \end{cases}$$



The theorem's limiting squared overlap at $\gamma = 1/2$ becomes positive above $\theta = \sqrt{\gamma}$. Positive overlap is weaker than consistency: at any finite spike strength the limit remains below one. Based on Paul (2007), Theorem 4, linked above.

2 Linear regression in high dimensions

Mid priority *Linear regression in high dimensions*

Consider the linear regression with observations (x_i, y_i) where $x_i \in \mathbb{R}^p$ and $i = 1, 2, \dots, n$. As usual, we posit a linear relationship between x_i and y_i up to some error term:

$$y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Denote $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$, $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$. When X has full column rank (so $n \geq p$), the least-squares estimator is given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

For any new input vector x_0 , we make prediction based on $\hat{\beta}$:

$$\hat{f}(x_0) = x_0^\top \hat{\beta}.$$

Our goal is to understand the prediction curve—especially the “double descent phenomenon” in Figure 2—a bit better by deriving theoretical results under reasonable assumptions.

We will partially explain the picture, with a focus on understanding the spike near $p = n$, the **interpolation threshold**. The four regimes below use isotropic Gaussian design and fixed noise variance. We compare signal sequences of fixed norm when discussing bias. This correctly specified linear model explains the spike and the behavior beyond it; it does not reproduce the initial fall in approximation bias in the schematic.

1. **Low dimensions:** $p_n/n \rightarrow 0$. The ordinary least squares (OLS) estimator has zero bias, and the prediction variance is approximately $\sigma^2 p_n/n$.
2. **High dimensions before the spike:** $p_n/n \rightarrow c \in (0, 1)$. The OLS estimator still has zero bias, but its prediction variance approaches $\sigma^2 c/(1 - c)$, which grows sharply as c approaches 1.
3. **High dimensions after the spike:** $p_n/n \rightarrow c \in (1, \infty)$. The min-norm estimator (aka ridgeless regression solution) has both bias and variance: the variance decreases as c grows, while the squared bias increases for fixed signal norm. Their sum need not be monotone.
4. **Very high dimensions:** $p_n/n \rightarrow \infty$. The minimum-norm interpolator retains a vanishing fraction of a fixed signal. Its squared bias approaches $\|\beta\|^2$, while its prediction variance is approximately $\sigma^2 n/p_n$.

2.1 Regime 1: low dimensions

High priority *Regime 1: low dimensions*

Throughout Section 2, the coefficient $\beta = \beta_n$ is deterministic, the noise vector ε is independent of X , and $\mathbb{E} \varepsilon = 0$, $\text{Cov}(\varepsilon) = \sigma^2 I_n$ with fixed $\sigma^2 > 0$. This includes fixed design: a deterministic matrix is independent of any random vector. A random test vector x_0 is independent of both X and ε . Conditional prediction risk averages over the training noise and this fresh test vector, keeping X fixed; it predicts the regression mean $x_0^\top \beta$. Predicting a fresh noisy response adds its independent noise variance σ^2 .

Let us derive the associated estimator error $\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X]$ and its prediction error $\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X]$. Note that

$$\hat{\beta} = (X^\top X)^{-1} X^\top (X\beta + \varepsilon) = \beta + (X^\top X)^{-1} X^\top \varepsilon.$$

Therefore, we have

$$\begin{aligned} \mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] &= \mathbb{E}[\text{Tr}((X^\top X)^{-1} X^\top \varepsilon \varepsilon^\top X (X^\top X)^{-1}) | X] \\ &= \text{Tr}(\mathbb{E}[(X^\top X)^{-1} X^\top \varepsilon \varepsilon^\top X (X^\top X)^{-1} | X]) \\ &= \text{Tr}((X^\top X)^{-1} X^\top \sigma^2 I_n X (X^\top X)^{-1}) \\ &= \sigma^2 \text{Tr}((X^\top X)^{-1}). \end{aligned}$$

where in the first equality we used the identity $\|u\|^2 = \text{Tr}(uu^\top)$ for any (column) vector u . From a similar calculation, we obtain

$$\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X, x_0] = \sigma^2 x_0^\top (X^\top X)^{-1} x_0.$$

For a deterministic scale $b_n > 0$, the notation $Z_n = (1 + o_P(1))b_n$ means $Z_n/b_n \xrightarrow{P} 1$.

Proposition 1. Suppose that $x_0, x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} N(0, I_p)$. Also, suppose $p = p_n$ depends on n and $p_n/n = o(1)$ as $n \rightarrow \infty$. We have

$$\begin{aligned}\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] &= (1 + o_P(1)) \frac{\sigma^2 p_n}{n}, \\ \mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] &= (1 + o_P(1)) \frac{\sigma^2 p_n}{n}.\end{aligned}$$

Remark 2. This result shows that the prediction mean square error (MSE) of $\hat{f}(x_0)$ increases approximately linearly in dimension p_n . Since the least-squares estimator is conditionally unbiased, the MSE is purely due to the variance.

Proof. Write $\hat{\Sigma} = X^\top X/n$ and $a_n = \|\hat{\Sigma} - I_p\|_{\text{op}}$. The covariance theorem with a diverging tail parameter $t_n = o(\sqrt{n})$ gives $a_n = o_P(1)$ when $p_n/n \rightarrow 0$. On the event $a_n < 1$, every eigenvalue lies in $[1 - a_n, 1 + a_n]$, so

$$\max_j |\lambda_j(\hat{\Sigma})^{-1} - 1| \leq \frac{a_n}{1 - a_n} = o_P(1).$$

Averaging these reciprocal eigenvalues yields

$$\text{Tr}((X^\top X)^{-1}) = \frac{p_n}{n}(1 + o_P(1)).$$

Consequently,

$$\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] = \sigma^2 \text{Tr}((X^\top X)^{-1}) = (1 + o_P(1)) \frac{\sigma^2 p_n}{n}. \quad (4)$$

For $\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X]$, notice that

$$\frac{1}{n} \|x_0\|^2 \lambda_{\min}((X^\top X/n)^{-1}) \leq x_0^\top (X^\top X)^{-1} x_0 \leq \frac{1}{n} \|x_0\|^2 \lambda_{\max}((X^\top X/n)^{-1}).$$

Then

$$\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] = \sigma^2 \mathbb{E}[x_0^\top (X^\top X)^{-1} x_0 | X] = \sigma^2 \text{Tr}((X^\top X)^{-1} \mathbb{E}[x_0 x_0^\top]).$$

Since $\mathbb{E}[x_0 x_0^\top] = I_p$, we use (4) and derive

$$\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] = (1 + o_P(1)) \frac{\sigma^2 p}{n}.$$

□

Prediction and coefficient error. For an independent isotropic test vector, the expected squared prediction error for the regression mean equals the expected squared coefficient error, conditional on X . A fresh noisy response adds its own noise variance.

High priority Practice: Prediction at a random test point

Exercise prompt. For a fixed coefficient error $v = (1, 1)^\top$ and a test vector with $\mathbb{E}[x_0 x_0^\top] = \text{diag}(1, 4)$, compute $\mathbb{E}[(x_0^\top v)^2]$ and compare it with $\|v\|_2^2$. What changes for an isotropic test vector?

See Exercise 2 in Section 6 for the full statement and complete solution.

2.2 Regime 2: high dimensions before the spike

Mid priority *Regime 2: high dimensions before the spike*

When $p_n/n \rightarrow c \in (0, 1)$ and X has i.i.d. standard Normal entries, the edge bounds suggested by Gordon's theorem are asymptotically exact:

$$\begin{aligned}\sigma_{\min}(X/\sqrt{n}) &\xrightarrow{a.s.} 1 - \sqrt{c}, \\ \sigma_{\max}(X/\sqrt{n}) &\xrightarrow{a.s.} 1 + \sqrt{c}.\end{aligned}$$

Put differently, the empirical spectral distribution of $X^\top X/n$ approaches a law supported on the interval $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$. This is known as the **Marchenko–Pastur (MP) distribution**. See Hastie et al., Proposition 2 and Appendix C.1 for the spectral-law calculation giving, almost surely

$$\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] \longrightarrow \sigma^2 \frac{c}{1 - c}, \quad \text{as } n \rightarrow \infty.$$

2.3 Regime 3: high dimensions after the spike

Mid priority *Regime 3: high dimensions after the spike*

Now consider $p > n$ and $\text{rank}(X) = n$, as holds almost surely for a Gaussian design. Every $y \in \mathbb{R}^n$ then has infinitely many solutions to $y = X\beta$, since the null space has dimension $p - n > 0$. Each solution is an **interpolator**. This means that we can find an estimator so that we fit the response variables perfectly; if we interpret $(x_i, y_i)_{i \leq n}$ as a *training dataset*, then we can have zero training error.

Among the set of interpolators, sometimes it is meaningful to study the minimum-norm interpolator:

$$\begin{aligned}\min_{b \in \mathbb{R}^p} \quad & \|b\|_2^2 \\ \text{subject to} \quad & y = Xb,\end{aligned} \tag{5}$$

where the norm is the Euclidean norm $\|\cdot\|_2$.

Claim 1. *Assume that $\text{rank}(X) = n$. The solution from (5) is unique and is given by*

$$\hat{\beta} = X^\top (XX^\top)^{-1} y.$$

Moreover, $\hat{\beta}$ is no longer an unbiased estimator of β , since $\mathbb{E}[\hat{\beta} | X] = H\beta$ generally differs from β , where $H = X^\top (XX^\top)^{-1} X$.

Proof. Note that $X\hat{\beta} = XX^\top (XX^\top)^{-1} y = y$, so $\hat{\beta}$ is a solution to $y = Xb$. Let $X = UDV^\top$ be its singular value decomposition with $U \in \mathbb{R}^{n \times n}$ orthogonal, $D \in \mathbb{R}^{n \times n}$ positive diagonal, and $V \in \mathbb{R}^{p \times n}$ satisfying $V^\top V = I_n$. Thus D is invertible. It is easily seen that any vector in

$$\mathcal{L} = \{u : u = (I_p - VV^\top)a, a \in \mathbb{R}^p\}$$

is in the null space of X . Because $\dim(\mathcal{L}) = \dim(\text{Null}(X)) = p - n$, we conclude that any interpolating coefficient b satisfies

$$b = \hat{\beta} + (I_p - VV^\top)a.$$

But $\hat{\beta}$ and $(I_p - VV^\top)a$ are orthogonal,

$$\|b\|^2 = \|\hat{\beta}\|^2 + \|(I_p - VV^\top)a\|^2 \geq \|\hat{\beta}\|^2.$$

Equality holds iff $(I_p - VV^\top)a = 0$ iff $b = \hat{\beta}$. This shows the uniqueness of the minimum-norm solution. By the SVD,

$$\begin{aligned}\hat{\beta} &= VD^{-1}U^\top y = VV^\top \beta + VD^{-1}U^\top \varepsilon, \\ \mathbb{E}[\hat{\beta} | X] &= VV^\top \beta \quad (\text{generally different from } \beta).\end{aligned}$$

□

The minimum-norm fit lies in the row space of X . For noiseless data it is the orthogonal projection of β onto that space; the null-space component cannot be recovered from $X\beta$. This projection is in parameter space $\mathbb{R}^p$, whereas Lecture 13 projected observations onto the column space in $\mathbb{R}^n$. With noise,

$$\mathbb{E}[\|\hat{\beta} - \beta\|_2^2 | X] = \|(I_p - H)\beta\|_2^2 + \sigma^2 \text{Tr}((XX^\top)^{-1}), \quad H = X^\top(XX^\top)^{-1}X.$$

Exercise 6 gives the complete decomposition. Roughly speaking, the roles of rows and columns are flipped in this regime.

For the following limit, assume $p_n/n \rightarrow c > 1$, i.i.d. standard Gaussian design, independent isotropic test covariates, noise variance σ^2 , and a deterministic signal sequence with $\|\beta_n\|^2 \rightarrow r^2 < \infty$. See Hastie et al., Theorem 1 and Appendix C.2 for the following limit, almost surely

$$\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] \rightarrow \underbrace{r^2(1 - 1/c)}_{\text{bias}^2} + \underbrace{\sigma^2 \frac{1}{c-1}}_{\text{variance}}, \quad \text{as } n \rightarrow \infty.$$

Interpolation trades bias against variance. Minimum-norm interpolation keeps the row-space part of the signal and discards its orthogonal component. Its conditional risk is squared projection bias plus a noise term involving $(XX^\top)^{-1}$; the two terms can move in opposite directions.

Mid priority *Practice: A two-coordinate minimum-norm fit*

Exercise prompt. For $X = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $y = 2$, find the minimum-Euclidean-norm interpolator and exhibit another interpolator with larger norm.

See Exercise 3 in Section 6 for the full statement and complete solution.

Low priority *Prediction risk across the interpolation threshold*

Figure 1 combines Regimes 2 and 3 under the same isotropic Gaussian design and independent-noise assumptions. The signal sequence satisfies $\|\beta_n\|^2 \rightarrow r^2$. The paper writes γ for our limiting aspect ratio $c = \lim p_n/n$; no change in data scaling is needed.

Write $R(c)$ for the limiting prediction risk $\mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X]$. This averages over training noise and an independent test input, keeping X fixed. It predicts the regression mean; predicting a fresh noisy response would add σ^2 . The two limits give

$$R(c) = \begin{cases} \sigma^2 \frac{c}{1-c}, & 0 < c < 1, \\ r^2 \left(1 - \frac{1}{c}\right) + \sigma^2 \frac{1}{c-1}, & c > 1. \end{cases} \quad (6)$$

In the figure, $\sigma^2 = 1$, so each legend value of $\text{SNR} = r^2/\sigma^2$ also specifies r^2 . The solid curves show the asymptotic risk in (6). The dots show finite-sample risks for Gaussian designs with $n = 200$. Each dotted horizontal line is the risk r^2 of predicting zero.

Both branches diverge toward the interpolation threshold $c = 1$, which is excluded from the formula. Before interpolation, this correctly specified model has zero bias, so risk rises with c . After interpolation, falling variance competes with growing squared bias; total risk can turn upward again. Exercise 8 asks you to check this competition.

Instructor's Note *Where the name double descent comes from*

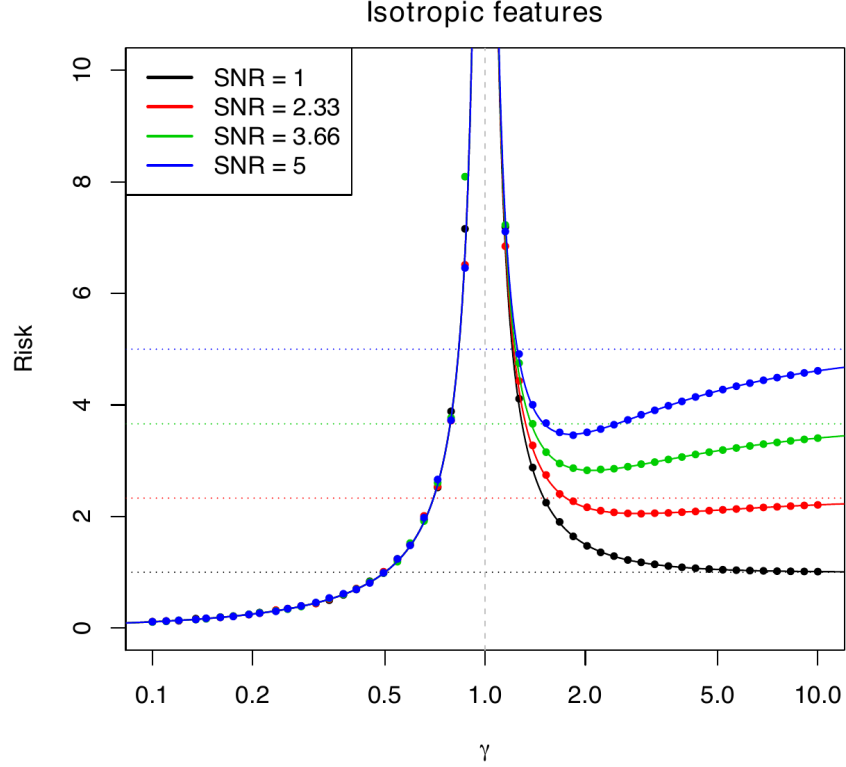


Figure 1: Isotropic prediction risk, with $\gamma = c$, $\sigma^2 = 1$ and $\text{SNR} = r^2/\sigma^2$. The horizontal axis is logarithmic. Source: Hastie et al., Figure 2, p. 10 (CMU author version).

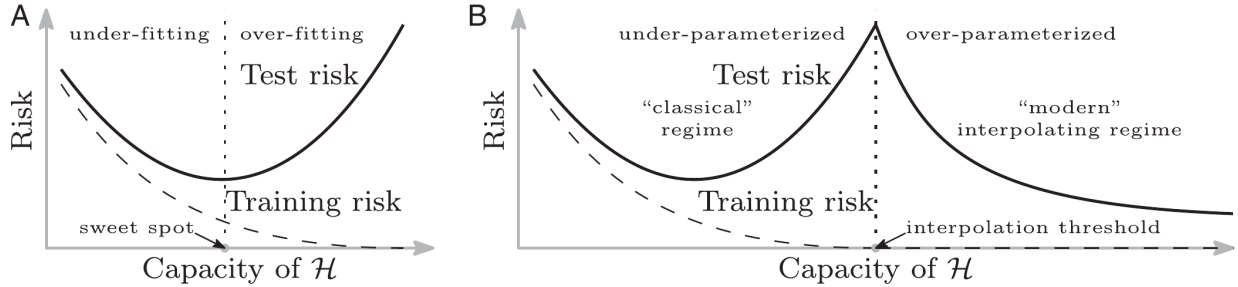


Figure 2: The dashed curve is training risk; the solid curve is test risk. The right panel extends the classical picture past interpolation, where a second decrease can occur. This is a schematic pattern, not a guarantee for every model or dataset. Figure: Belkin, Hsu, Ma and Mandal (2019), Figure 1.

2.4 Regime 4: very high dimensions

Mid priority Regime 4: very high dimensions

We will show a result about the MSE for the regime $p_n \gg n$.

Proposition 2. Suppose that $x_0, x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} N(0, I_p)$, and that $p = p_n$ depends on n and $n/p_n = o(1)$ as $n \rightarrow \infty$. Let $\hat{\beta} = X^\top (XX^\top)^{-1}y$. For $p_n \geq n$, X has full row rank almost surely, and

$$\begin{aligned}\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] &= (1 + o_P(1)) \cdot \left[\left(1 - \frac{n}{p_n}\right) \|\beta\|^2 + \frac{\sigma^2 n}{p_n} \right], \\ \mathbb{E}[(x_0^\top \hat{\beta} - x_0^\top \beta)^2 | X] &= (1 + o_P(1)) \cdot \left[\underbrace{\left(1 - \frac{n}{p_n}\right) \|\beta\|^2}_{\text{bias}^2} + \underbrace{\frac{\sigma^2 n}{p_n}}_{\text{variance}} \right],\end{aligned}$$

Remark 3. The result says that the prediction MSE has a bias component (which depends on $\|\beta\|^2$) and a variance component (which depends on σ^2). Far beyond interpolation, the variance is small, of order n/p_n , even though the training data are fitted exactly. This helps explain the second descent. It does not by itself imply **benign overfitting**, meaning interpolation with vanishing excess prediction risk: for $\|\beta_n\|^2 \rightarrow r^2 > 0$, the isotropic model here has total risk tending to r^2 .

Proof. From the definition we have $\hat{\beta} - \beta = (X^\top (XX^\top)^{-1}X - I_p)\beta + X^\top (XX^\top)^{-1}\varepsilon$. Denote $H = X^\top (XX^\top)^{-1}X \in \mathbb{R}^{p \times p}$ for simplicity. Note that $H^2 = H$ and $H^\top = H$, so H is an orthogonal projection matrix. We have

$$\begin{aligned}\mathbb{E}[\|\hat{\beta} - \beta\|^2 | X] &= \beta^\top (H - I_p)^2 \beta + \mathbb{E}[\varepsilon^\top (XX^\top)^{-1} \varepsilon | X] \\ &= \beta^\top (I_p - H) \beta + \text{Tr}((XX^\top)^{-1} \mathbb{E}[\varepsilon \varepsilon^\top | X]) \\ &= \beta^\top (I_p - H) \beta + \sigma^2 \text{Tr}((XX^\top)^{-1}).\end{aligned}$$

This is exactly a bias—variance decomposition. We only need to show

$$\beta^\top H \beta = (1 + o_P(1)) \cdot \frac{n}{p} \|\beta\|^2, \quad \text{Tr}((XX^\top)^{-1}) = (1 + o_P(1)) \cdot \frac{n}{p}.$$

The second equality is easy: in fact, we just need to switch the role of p and n and then apply the reciprocal-eigenvalue argument in Proposition 1. One way to prove the first equality is to use the orthogonal invariance property of a standard normal random vector, namely, $XR \stackrel{d}{=} X$ for any orthogonal $R \in O(p)$. Hence, we have $H = UU^\top$ where $U \in \mathbb{R}^{p \times n}$ has orthonormal columns ($U^\top U = I_n$) and a rotation-invariant (Haar) distribution,³ so

$$\beta^\top H \beta = \|U^\top \beta\|^2 = (1 + o_P(1))(n/p) \|\beta\|^2.$$

For a nonzero deterministic signal $\beta = \beta_n$, rotational invariance identifies $\beta^\top H \beta / \|\beta\|^2$ with $Q_n / (Q_n + R_{p-n})$, where $Q_n \sim \chi_n^2$ and $R_{p-n} \sim \chi_{p-n}^2$ are independent. Since $n \rightarrow \infty$ and $n/p \rightarrow 0$, the LLN gives the relative approximation above. Indeed, Q_n/n and $(Q_n + R_{p-n})/p$ both converge to one in probability, giving the stated ratio. The zero-signal case has no bias term.

Similarly, we have

$$\begin{aligned}\mathbb{E}[(x_0^\top (\hat{\beta} - \beta))^2 | X] &= \mathbb{E}[(\hat{\beta} - \beta)^\top x_0 x_0^\top (\hat{\beta} - \beta) | X] \\ &= \mathbb{E}[\mathbb{E}[(\hat{\beta} - \beta)^\top x_0 x_0^\top (\hat{\beta} - \beta) | X, \varepsilon] | X] \\ &= \mathbb{E}[(\hat{\beta} - \beta)^\top \mathbb{E}[x_0 x_0^\top] (\hat{\beta} - \beta) | X] \\ &= \mathbb{E}[\|\hat{\beta} - \beta\|^2 | X],\end{aligned}$$

³Here rotation-invariant means that applying any fixed orthogonal transformation leaves the distribution of the random subspace unchanged. It makes no direction preferred before observing the data.

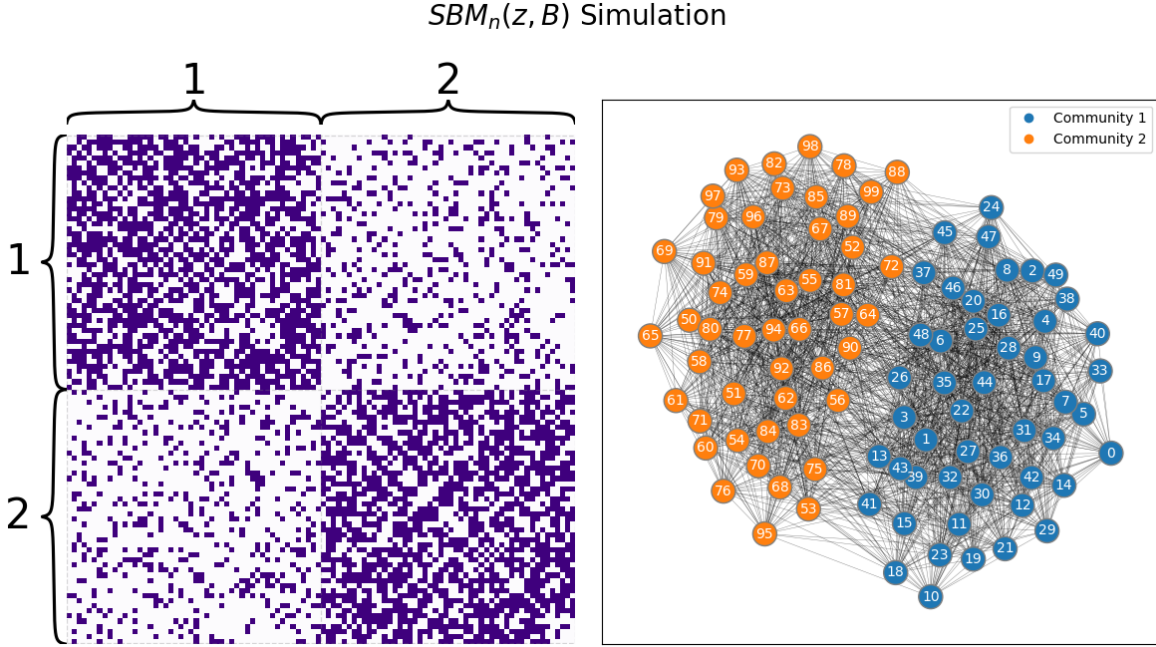


Figure 3: The same two-community network as an adjacency matrix (left) and a graph (right). Dense diagonal blocks show within-community connections; the nodes are grouped by their true labels for illustration. In the source image, B is the matrix of connection probabilities, with within-community probabilities 0.6 and 0.4 and between-community probability 0.2. Our model below simplifies to a common within-community probability p . Figure: NeuroData, Graph Statistics book, stochastic block model chapter.

so the same conclusion follows from the previous calculation.

□

3 Community detection (optional)

This section is an optional application of spectral methods. It is not required for exams. The existing priority labels distinguish the model from its technical extensions.

Mid priority *Community detection*

Let us revisit the problem of community detection. In statistics and machine learning, we often want to infer community structures of n nodes based on the observed network/graph between the nodes. This problem is called *community detection*. To model such community structures, we assume the following statistical model. Suppose there are n nodes of interest (e.g., people on Facebook, students in a club), and each node belongs to one of the two communities. Let $z^* \in \{\pm 1\}^n$ represent the unknown membership vector.

Stochastic block model. Here p_n and q_n denote edge probabilities, not the feature dimension used in Sections 1–2. For $i \leq j$, there is an edge between nodes i and j with probability given by

$$P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = \begin{cases} p_n, & \text{if } z_i^* = z_j^* \\ q_n, & \text{if } z_i^* = -z_j^* \end{cases} \quad (7)$$

where $p_n > q_n$. Here $(A_{ij})_{i \leq j}$ are jointly independent (but not identically distributed) Bernoulli random variables. Intuitively, this model captures the basic structure that the connection probability of two nodes in the same community is higher than across the two communities.

For the dense-network result, take even n and (i) assume $p_n \equiv p$ and $q_n \equiv q$ with fixed $0 < q < p < 1$, (ii) assume $\langle z^*, \mathbf{1}_n \rangle = 0$ namely two communities have equal sizes, and (iii) set $A_{ij} = A_{ji}$ so that $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, usually called the **adjacency matrix**. We also allow self-loops⁴ namely $A_{ii} \sim \text{Bernoulli}(p)$ independently in the graph.

From the statistical perspective, we have a parametric family where the unknown parameter $z^* \in \{\pm 1\}^n$ is a high-dimensional vector taking discrete values. The data are a random graph represented by the adjacency matrix A generated according to the model (7).

Spectral method. Can we use the adjacency matrix A to reveal the hidden community structure, namely estimating z^* ? To convince ourselves that this is plausible, consider the expected version $A^* = \mathbb{E} A$, and we find (see Exercise 5)

$$A^* = \frac{p-q}{2} z^* (z^*)^\top + \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top.$$

Note that the above expression gives exactly the eigendecomposition of A^* : it has only two nonzero eigenvalues $\lambda_1^* = \frac{(p+q)n}{2}$ and $\lambda_2^* = \frac{(p-q)n}{2}$ with corresponding eigenvectors $u_1^* = \mathbf{1}_n / \sqrt{n}$ and $u_2^* = z^* / \sqrt{n}$.

We hope that the eigendecomposition of A produces eigenvectors close to u_1^* and u_2^* . This inspires a natural algorithm below.

1. **Step 1.** Calculate a unit eigenvector $u := u_2$ corresponding to the second-largest eigenvalue of A (ordered by value).
2. **Step 1'.** Alternatively, if p and q are known, calculate a unit eigenvector u for the largest eigenvalue of $A - \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top$.
3. **Step 2.** Form estimator $z \in \{\pm 1\}^n$ by discretizing real values to binary values: setting $z_i = 1$ if $u_i \geq 0$ and $z_i = -1$ if $u_i < 0$.

This algorithm is known as the (vanilla) spectral method, as it relies on eigen-decomposition to form the estimator.

3.1 Theoretical guarantee for spectral method

Low priority *Theoretical guarantee for spectral method*

Theorem 5. Consider the above algorithm Step 1 + Step 2 or Step 1' + Step 2. For any constant $\varepsilon > 0$, we have

$$P \left(\min_{s \in \{\pm 1\}} \frac{1}{\sqrt{n}} \|sz - z^*\|_2 \leq \varepsilon \right) = 1 - o(1), \quad \text{as } n \rightarrow \infty.$$

Because of the identity

$$\frac{1}{n} \|sz - z^*\|_2^2 = \frac{1}{n} (2n - 2\langle sz, z^* \rangle) = \frac{4}{n} |\{i : sz_i \neq z_i^*\}|,$$

the theorem equivalently says that the mismatch ratio between z and z^* is arbitrarily small as $n \rightarrow \infty$ up to sign.

⁴In graph theory, simple graphs don't allow self-loops. But this is only a technical assumption and removing self-loops won't change the main results.

Remark 4 (identifiability). *Our statistical model gives the same joint probability distribution for our observations if we replace z^* by $-z^*$. We have the freedom to label the two communities of nodes by either $-1, 1$ or $1, -1$. In other words, this model is only identifiable up to sign. So the up to sign in the theorem statement is a way of resolving the identifiability issue.*

Proof. We shall give a proof for Step 1' + Step 2. The proof for Step 1 + Step 2 uses the corresponding isolated-second-eigenvalue perturbation bound, with gap $\min(qn, (p-q)n/2)$ for fixed $0 < q < p < 1$; see Yu, Wang and Samworth, Corollary 3. Denote $\tilde{A} = A - \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^\top$.

First, we note that we can write A as a sum of two parts $A = A^1 + A^2$ where A^1 is an upper triangular matrix and A^2 is a lower triangular matrix, namely

$$A_{ij}^1 = A_{ij} \mathbf{1}\{i \leq j\}, \quad A_{ij}^2 = A_{ij} \mathbf{1}\{i > j\}.$$

Note that A^1 has jointly independent entries (since constants are independent of any random variables). Since Bernoulli random variables are subgaussian, we have $\max_{i,j} \|A_{ij}^1\|_{\psi_2} \leq C$. By Lecture 16, Theorem 1, with probability at least $1 - 2e^{-n}$, we have

$$\|A^1 - \mathbb{E}[A^1]\|_{\text{op}} \leq C\sqrt{n}.$$

We have a similar bound on $\|A^2 - \mathbb{E}[A^2]\|_{\text{op}}$. The two triangular parts need not be independent of one another; the triangle inequality and union bound suffice. So together, they imply

$$\|\tilde{A} - \mathbb{E}[\tilde{A}]\|_{\text{op}} = \|A - \mathbb{E}[A]\|_{\text{op}} \leq C\sqrt{n}.$$

holds with probability at least $1 - 4e^{-n}$ by a union bound. Since the first (which is the only nonzero) eigenvalue of $\mathbb{E}[\tilde{A}]$ is $(p-q)n/2$, by applying Davis–Kahan’s theorem,

$$\min_{s \in \{\pm 1\}} \|su - u_2^*\|_2 \leq \frac{C\sqrt{n}}{(p-q)n} = \frac{C_{p,q}}{\sqrt{n}}.$$

Next, to study the discretization step, observe that if a mismatch between sz_i and z_i^* happens, the i -th entry of the two eigenvectors su and u_2^* have opposite signs or one is zero. With the stated tie rule at zero, we still have $|[su - u_2^*]_i| \geq 1/\sqrt{n}$. Thus,

$$\frac{1}{\sqrt{n}} \mathbf{1}\{sz_i \neq z_i^*\} \leq |[su - u_2^*]_i|$$

By squaring the above inequality and summing over i ,

$$\frac{1}{n} |\{i : sz_i \neq z_i^*\}| \leq \|su - u_2^*\|_2^2.$$

The mismatch fraction is therefore at most $C_{p,q}^2/n$ on this event, and the normalized label error is at most $2C_{p,q}/\sqrt{n}$. This proves the theorem, but this bound alone does not force the number of errors to be zero. $\square$

Low priority Practice: Count sign errors from vector error

Exercise prompt. If $v = z^*/\sqrt{n}$, $z^* \in \{\pm 1\}^n$, and $\|su - v\| \leq r$, bound the fraction of sign mismatches after thresholding u with $z_i = 1$ for $u_i \geq 0$ and $z_i = -1$ otherwise. Evaluate the bound for $r = 0.05$ and $n = 10000$.

See Exercise 4 in Section 6 for the full statement and complete solution.

3.2 Phase transition phenomenon in community detection*

Low priority *Phase transition phenomenon in community detection*

In fact, our spectral algorithm works in a much more challenging (sparser) regime with stronger results. We say that an estimator z achieves *exact recovery* if

$$\mathbb{P}\{z = z^* \text{ or } z = -z^*\} = 1 - o(1), \quad n \rightarrow \infty.$$

Exact recovery is a stronger guarantee since it requires that the membership of every node be estimated correctly up to sign, namely $\|sz - z^*\|_2 = 0$ instead of allowing a small fraction of estimation mismatch. The following theorem considers vanishing p_n and q_n instead of constant p and q .

Theorem 6. *Suppose that*

$$p_n = \frac{a \log n}{n}, \quad q_n = \frac{b \log n}{n}$$

where $a > b > 0$ are constants. Then

- *If $\sqrt{a} - \sqrt{b} < \sqrt{2}$, no estimator achieves exact recovery uniformly over the balanced community assignments.*
- *If $\sqrt{a} - \sqrt{b} > \sqrt{2}$, the uncentered spectral algorithm (Step 1 + Step 2) achieves exact recovery; see [1, Theorem 3.2].*

An important matrix operator norm bound is $\|A - A^*\|_{\text{op}} \ll \|A^*\|_{\text{op}}$ with high probability in the $\log n/n$ regime. The dense bound $C\sqrt{n}$ is too large when the signal eigenvalue is only of order $\log n$. Even an improved operator-norm bound alone does not establish exact recovery: one must also control individual eigenvector coordinates so every sign is correct. This is the additional step in [1, Theorem 3.2]. The boundary case $\sqrt{a} - \sqrt{b} = \sqrt{2}$ is not asserted here.

Comparison with the MLE. In an earlier lecture, we derived the MLE for estimating z^* , which involves a complicated optimization problem. In comparison, the spectral method is fast, with an $O(n^2)$ cost per dense matrix-vector iteration. Computing a full eigendecomposition generally costs $O(n^3)$; the iteration count and accuracy matter.

For exact recovery, the MLE cannot do better than the spectral method in our setting. It cannot break the information limit $\sqrt{a} - \sqrt{b} \sim \sqrt{2}$ in Theorem 6.

4 Further reading

Further reading *Further reading*

For the Davis–Kahan perturbation bound, the original reference is [3] and a recent reference is [6]. The result about high-dimensional PCA is from [5]. The double descent phenomenon was popularized by [2], and a theoretical analysis based on linear models can be found in [4]. The spectral method for community detection is analyzed in [1].

Key ideas

- Covariance noise need not vanish when dimension grows proportionally to sample size.
- A PCA eigenvector can have positive limiting overlap without being consistent.
- The risk of a minimum-norm fit changes across the interpolation threshold.
- Spectral community detection is an optional application of the matrix tools.

5 Detailed reference

High priority *An eigengap stabilizes direction*

A small operator perturbation gives a small top-eigenvector error, up to sign, when divided by a positive population eigengap. A shrinking gap can defeat consistency even if the covariance error tends to zero. Return to the earlier explanation.

High priority *Prediction and coefficient error*

For an independent isotropic test vector, the expected squared prediction error for the regression mean equals the expected squared coefficient error, conditional on X . A fresh noisy response adds its own noise variance. Return to the earlier explanation.

Mid priority *Interpolation trades bias against variance*

Minimum-norm interpolation keeps the row-space part of the signal and discards its orthogonal component. Its conditional risk is squared projection bias plus a noise term involving $(XX^\top)^{-1}$; the two terms can move in opposite directions. Return to the earlier explanation.

Looking ahead

The optional Part VI returns to shrinkage and sparsity: can a deliberately biased rule improve on an otherwise optimal unbiased estimator? The risk comparisons and dimensional effects here motivate that next question. Explore Part VI (optional).

You have finished Part V—check yourself. Preview Part VI.

6 Exercises and complete solutions

High priority *What does an eigengap control?*

Exercise 1 (What does an eigengap control?).

For $\Sigma = \text{diag}(3, 1)$ and a symmetric perturbation E with $\|E\|_{\text{op}} \leq 0.1$, bound the top eigenvector error up to sign.

Solution.

The population gap is 2. Davis–Kahan gives $\min_s \|\widehat{u}_1 - u_1\| \leq 2\sqrt{2}(0.1)/2 = 0.1\sqrt{2}$. Weyl also gives eigenvalue errors at most 0.1, keeping the two perturbed eigenvalues separated. □

High priority *Prediction at a random test point*

Exercise 2 (Prediction at a random test point).

For a fixed coefficient error $v = (1, 1)^\top$ and a test vector with $\mathbb{E}[x_0 x_0^\top] = \text{diag}(1, 4)$, compute $\mathbb{E}[(x_0^\top v)^2]$ and compare it with $\|v\|_2^2$. What changes for an isotropic test vector?

Solution.

For the stated vector, $\mathbb{E}[(x_0^\top v)^2] = v^\top \text{diag}(1, 4)v = 5$, whereas $\|v\|_2^2 = 2$. Isotropy makes the two equal. More generally, the following calculation gives the risk identity.

Condition on the training data and training noise, and write $v = \widehat{\beta} - \beta$. Then $\mathbb{E}_{x_0}(x_0^\top v)^2 = v^\top \mathbb{E}(x_0 x_0^\top)v = \|v\|^2$. Average over the training noise conditional on X . This is prediction of the noiseless regression mean; a new noisy response adds σ^2 . □

Mid priority *A two-coordinate minimum-norm fit*

Exercise 3 (A two-coordinate minimum-norm fit).

For $X = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $y = 2$, find the minimum-Euclidean-norm interpolator and exhibit another interpolator with larger norm.

Solution.

The constraint is $b_1 + b_2 = 2$. Write $(b_1, b_2) = (1+t, 1-t)$; its squared norm is $2+2t^2$, uniquely minimized at $(1, 1)$. The fit $(2, 0)$ also interpolates but has squared norm 4 rather than 2. □

Low priority *Count sign errors from vector error*

Exercise 4 (Count sign errors from vector error).

Optional practice for Section 3; not tested in exams.

If $v = z^*/\sqrt{n}$, $z^* \in \{\pm 1\}^n$, and $\|su - v\| \leq r$, bound the fraction of sign mismatches after thresholding u with $z_i = 1$ for $u_i \geq 0$ and $z_i = -1$ otherwise. Evaluate the bound for $r = 0.05$ and $n = 10000$.

Solution.

At a mismatched coordinate, $|su_i - v_i| \geq 1/\sqrt{n}$. If M coordinates mismatch, summing squared coordinate errors gives $M/n \leq \|su - v\|^2 \leq r^2$. Equivalently the squared normalized label-vector error is $4M/n \leq 4r^2$. The mismatch fraction is at most 0.0025, hence at most 25 of 10000 vertices. □

Mid priority *The population community eigenvectors*

Exercise 5 (The population community eigenvectors).

Optional practice for Section 3; not tested in exams.

Derive the mean adjacency matrix and its two nonzero eigenpairs in the balanced two-community model with Bernoulli self-loops.

Solution.

For each pair, $I\{z_i = z_j\} = (1 + z_i z_j)/2$, so $\mathbb{E} A_{ij} = (p+q)/2 + (p-q)z_i z_j/2$. Balance gives $\mathbf{1}^\top z = 0$ and both vectors have squared norm n . Multiplying the resulting rank-two matrix by $\mathbf{1}/\sqrt{n}$ and $z/\sqrt{n}$ yields eigenvalues $(p+q)n/2$ and $(p-q)n/2$. □

Mid priority *Conditional risk of an interpolator*

Exercise 6 (Conditional risk of an interpolator).

For full-row-rank X , prove the min-norm conditional risk is $\|(I - H)\beta\|^2 + \sigma^2 \text{Tr}((XX^\top)^{-1})$, where $H = X^\top (XX^\top)^{-1} X$.

Solution.

The error is $-(I - H)\beta + X^\top (XX^\top)^{-1} \varepsilon$. The mean-zero noise kills the cross term. The squared bias is the first term. Cyclicity of trace gives noise risk $\sigma^2 \text{Tr}[X^\top (XX^\top)^{-2} X] = \sigma^2 \text{Tr}((XX^\top)^{-1})$. □

Mid priority *Inspect the spiked-PCA limit*

Exercise 7 (Inspect the spiked-PCA limit).

For fixed $\gamma \in (0, 1)$, check that the squared-overlap formula is zero at $\theta = \sqrt{\gamma}$, lies in $(0, 1)$ above the threshold, and tends to one as $\theta \rightarrow \infty$.

Solution.

At the threshold the numerator $1 - \gamma/\theta^2$ is zero. Above it, this numerator lies between zero and one, while the denominator $1 + \gamma/\theta$ exceeds one. Both tend to one as θ grows, so the ratio tends to one. These checks expose the distinction between the signal strength θ and its square. □

Low priority *Separate the two parts of double descent*

Exercise 8 (Separate the two parts of double descent).

For $c > 1$, differentiate $B(c) = r^2(1 - 1/c)$ and $V(c) = \sigma^2/(c - 1)$. Does the total risk always decrease?

Solution.

$B'(c) = r^2/c^2 \geq 0$ and $V'(c) = -\sigma^2/(c - 1)^2 \leq 0$. The derivative of the total is their sum, which can have either sign depending on c and the signal-to-noise ratio. The variance decrease alone does not establish a decreasing total risk. □

References

- [1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.
- [2] Mikhail Belkin, Daniel J. Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849 – 15854, 2019.
- [3] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
- [4] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.
- [5] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.
- [6] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

What changed

October 19, 2026 (instructor follow-up). Marked all of Section 3 (community detection) and its linked exercises optional, while preserving their reading priorities.

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.

October 14, 2026 (figure follow-up). Restored the credited NeuroData community-network figure and its explanation of the different model parameters.