Dimension reduction and manifold learning

Introductory lecture notes

Eddie Aamari

Département de mathématiques et applications
CNRS, École normale supérieure – PSL

2024

Contents

CHAPTER 1

High-dimensional geometry, estimation, and hope

Modern statistical learning often deals with high-dimensional data, such as genomic data, images, text, and time series, which all present challenges in terms of computational resources and algorithm design. Even theoretically, algorithm performance should decline as dimensionality increases, a phenomenon known as the curse of dimensionality, caused by the geometric behavior of high-dimensional spaces. As summarized by [Giraud, 2021, p.3] the impact of high dimensionality on statistics is multiple.

- High-dimensional spaces are immense, with data points scattered widely.
- Small fluctuations in various directions can combine to create a significant overall change.
- A rare event resulting from the accumulation of many rare events may actually be common.
- Numerical computations and optimizations in high-dimensional spaces can be highly resource-intensive

Even within each of these categories, the oddities of large dimension are multifaceted. We will tour a few phenomena relevant for statistics. See also [Bar and Pozdnyakov, 2024] for more advanced oddities.

1.1 Geometry of high-dimensional point clouds

1.1.1 Concentration of the norm

Many statistical and machine learning methods rely on local averages based on distances between sample points. We shall exemplify these in Section 1.2 later on. Naturally, these estimators are only meaningful if the sample $X_1, \dots, X_n$ is well-distributed in space. However, in

high dimensions $D \gg 1$, this is generally not the case. We expect an iid sample to behave as follows: all points lie on the same sphere, and are approximately equidistant from one another. The proof relies on Hoeffding's inequality (Theorem 1.1).

Lemma 1.1. (Thing shell phenomenon) Let $X_1, \dots, X_n \sim_{\text{iid}} \text{Unif}([0,1]^D)$. Then:

(i) *(Points nearly lie on a sphere)* For all $\epsilon > 0$,

$$\mathbb{P}\left(\forall i \in \{1,\dots,n\}, \quad (1+\epsilon)\frac{D}{3} \leq \|X_i\|^2 \leq (1+\epsilon)\frac{D}{3}\right) \xrightarrow[D\to\infty]{} 1.$$

(ii) *(Points nearly are equidistant)* For all $\epsilon > 0$,

$$\mathbb{P}\left(\forall i \neq i' \in \{1,\dots,n\}, \quad (1+\epsilon)\frac{D}{6} \leq \|X_i - X_{i'}\|^2 \leq (1+\epsilon)\frac{D}{6}\right) \xrightarrow[D\to\infty]{} 1.$$

The proof makes use of the following measure concentration inequality for sums of independent bounded random variables.

Theorem 1.1. (Hoeffding's inequality) If $Y_1, \dots, Y_N$ are independent random variables such that $a_j \leq Y_j \leq b_j$ for all $j \in \{1,\dots,N\}$, then $S := \sum_{j=1}^N Y_j$ satisfies

$$\mathbb{P}\left(S > \mathbb{E}[S] + t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right).$$

Proof.

See [Boucheron et al., 2013, Section 2.6]

Proof of Lemma 1.1.

First fix $i \in \{1,\dots,n\}$, and write $X_i = (X_{i,j})_{j\in\{1,\dots,D\}}$ with $X_{i,j} \sim_{iid} \text{Unif}[0,1]$. The random variables $X_{i,j}^2$ have mean $1/3$ and take values in $[0,1]$. Therefore, Hoeffding's concentration inequality (Theorem 1.1) yields

$$\mathbb{P}\left(\left|\sum_{j=1}^D X_{i,j}^2 - \frac{D}{3}\right| \geq s\right) \leq 2\exp\left(-\frac{2s^2}{D}\right),$$

for all $s \geq 0$. Picking $s = (D/3)\epsilon$, the right-hand side equals $\exp(-2D/9)$. Then, a union bound over $i \in \{1,\dots,n\}$ gives a lower-bound the probability of the event of (i) equal to $1 - n\exp(-2D/9)$, which goes to 1.

For point (ii), we let $Z_{i,i',j} := (X_{i,j} - X_{i',j})^2$. The random variables $(Z_{i,i',j})_{j\in\{1,\dots,D\}}$ are iid and

take values in $[0,1]$. Furthermore, $\mathbb{E}[Z_{i,i',j}] = 1/6$, so that by Hoeffding's inequality,

$$\mathbb{P}\left(\left|\|X_i - X_{i'}\|^2 - \frac{D}{6}\right| \geq s\right) \leq 2\exp\left(-\frac{2s^2}{D}\right).$$

Taking $s = (D/6)\epsilon$ and using a union bound over $(i,i') \in \binom{D}{2}$ allows to conclude.

Note that from the proof, we can actually strengthen the result by letting $n \to \infty$ as $D \to \infty$, as soon as $D \gg \log n$. In fact, we will see that the (apparently) critical dimension value $D \simeq \log n$ has connection with the so-called *Johnson-Lindenstrauss* lemma. It roughly states that the metric structure of a set of n points $X_1, \dots, X_n \in \mathbb{R}^D$ is essentially characterized by its projection $\mathrm{pr}_V(X_1), \dots, \mathrm{pr}_V(X_n)$ onto a well-chosen $(C \log n)$-dimensional subspace $V \subset \mathbb{R}^D$.

1.1.2 Volume of high-dimensional balls

Another way to account for the weird immensity of $\mathbb{R}^D$ for large D is to count the minimal size N of a sample needed to cover the whole unit-cube $[0,1]^D$, meaning that

$$[0,1]^D \subset \bigcup_{i=1}^{N} \mathrm{B}_2^D(x_i, 1).$$

Given such a covering with minimal cardinality N_{covering}, a volume argument yields

$$\begin{aligned} 1 = \left|[0,1]^D\right| &\leq \left|\bigcup_{i=1}^{N_{\text{covering}}} \mathrm{B}_2^D(x_i, 1)\right| \\ &\leq \sum_{i=1}^{N_{\text{covering}}} \left|\mathrm{B}_2^D(x_i, 1)\right| = N_{\text{covering}}\omega_D, \end{aligned}$$

where $\omega_D := \left|\mathrm{B}_2^D(0,1)\right|$ is the volume of the D-dimensional Euclidean ball. Hence, we necessary have $N_{\text{covering}} \geq \omega_D$. It remains to understand how ω_D behaves as D grows large.

Lemma 1.2. (Volume of the unit Euclidean ball) For all $D \in \mathbb{N}^*$,

$$\omega_D = \frac{\pi^{D/2}}{\Gamma\left(\frac{D}{2}+1\right)} \sim_{D\to\infty} \frac{1}{\sqrt{\pi D}}\left(\frac{2\pi e}{D}\right)^{D/2}.$$

Proof.

The explicit formula for ω_D can be obtained by a change of variable, and by induction on D. The asymptotic equivalent comes from Stirling's formula $\Gamma(z) \sim_{z\to\infty} \sqrt{2\pi/z}\left(\frac{z}{e}\right)^z$.

Numerical application. The number of samples needed to cover $[0,1]^D$ at a precision of 1 behaves thus roughly like $D^{D/2}$, which is of course completely impractical. The bound $N_{\text{covering}} \geq 1/\omega_D$ yields:

- for $D = 20$, $N_{\text{covering}} \geq 39$;
- for $D = 30$, $N_{\text{covering}} \geq 45\ 631$;
- for $D = 40$, $N_{\text{covering}} \geq 277\ 413\ 227$;

which is already a huge size for a dataset. For the MNIST dataset, the dimension of the pictures is $D = 784$, and the resulting bound yields $N_{\text{covering}} \gtrsim 10^{653}$.

Remark 1.1. (Cube VS Euclidean balls in high dimension) The intuition of high dimension is sharpened by considering the cube $\mathrm{B}_\infty^D(0,1) = [-1,1]^D$, for which:

- the ratio of the volume with the smallest Euclidean ball that contains it satisfies

$$\frac{|\mathrm{B}_2^D(0,\sqrt{D})|}{|\mathrm{B}_\infty^D(0,1)|} = \frac{\omega_D D^{\frac{D}{2}}}{2^D} = \frac{\pi^{\frac{D}{2}} D^{\frac{D}{2}}}{2^D \Gamma\left(\frac{D}{2}+1\right)} \sim \frac{1}{\sqrt{\pi D}} \left(\frac{\pi e}{2}\right)^{\frac{D}{2}} \xrightarrow[D\to\infty]{} \infty.$$

- the ratio of the volume with the largest Euclidean ball it contains satisfies

$$\frac{|\mathrm{B}_\infty^D(0,1)|}{|\mathrm{B}_2^D(0,1)|} = \frac{2^D}{\omega_D} = \frac{2^D \Gamma\left(\frac{D}{2}+1\right)}{\pi^{\frac{D}{2}}} \sim \sqrt{\pi D} \left(\frac{2D}{\pi e}\right)^{\frac{D}{2}} \xrightarrow[D\to\infty]{} \infty.$$

1.1.3 Spheres and gaussians

As a final example of a counter-intuitive phenomenon in high dimensions, let us turn towards the uniform distributions over spheres. It appears that they have a representation based on independent gaussian ingredients.

Proposition 1.1. (Gaussians and uniforms on spheres)

- Let $Z = (Z_1, \dots, Z_D)$ be a $\mathbb{R}^D$-valued random variable. Then

$$Z \sim \mathcal{N}(0, I_D) \iff (Z/\|Z\|, \|Z\|^2) \sim \mathrm{Unif}\left(\mathrm{S}_2^{D-1}(0,\sqrt{1})\right) \otimes \chi^2(D).$$

- If $X = (X_1, \dots, X_D) \sim \mathrm{Unif}(\mathrm{S}_2^{D-1}(0,\sqrt{D}))$, then for all fixed $k \geq 1$, the projection of X over the k-dimensional subspace $E_k := \mathbb{R}^k \times \{0\}^{D-k}$ satisfies

$$\mathrm{pr}_{E_k}(X) = (X_1, \dots, X_k) \rightsquigarrow_{D\to\infty} \mathcal{N}(0, I_k)$$

- $X \sim \mathrm{Unif}(\mathrm{S}_2^{D-1}(0,1))$, then for all fixed $\theta \in \mathrm{Unif}(\mathrm{S}_2^{D-1}(0,1))$,

$$\sqrt{D}\langle \theta, X\rangle \rightsquigarrow_{D\to\infty} \mathcal{N}(0,1).$$

Proof.

The first point follows from the use of spherical coordinates $e^{-\|x\|^2/2}\mathrm{d}x = r^{D-1}e^{-r^2/2}\mathrm{d}r\mathrm{d}\theta$. For the second point, use the representation $X = \sqrt{D}Z/\|Z\|$ with $Z \sim \mathcal{N}(0, I_D)$. As $(Z_1, \dots, Z_k) \sim \mathcal{N}(0, I_k)$ and $\sqrt{D}/\|Z\| \xrightarrow[D\to\infty]{a.s.} 1$ by the law of large numbers, Slutsky's lemma yields

$$\mathrm{pr}_{E_k}(X) = (X_1, \dots, X_k) = \frac{\sqrt{D}}{\|Z\|}(Z_1, \dots, Z_k) \rightsquigarrow_{D\to\infty} \mathcal{N}(0, I_k)$$

The third point is an application of the second one with $k = 1$.

The above result has a direct consequence in terms of concentration around hyperplanes.

Proposition 1.2. (Concentration around equator and orthogonality in high dimension)

- Let $X \sim \mathrm{S}_2^{D-1}(0,1)$. Given fixed $\theta \in \mathrm{S}_2^{D-1}(0,1)$, write $H_\theta := \mathrm{span}(\theta)^\perp$ for the hyperplane orthogonal to θ. Then when $D \to \infty$, X is concentrated around the equator $E_\theta := H_\theta \cap \mathrm{S}_2^{D-1}(0,1)$ orthogonal to θ. That is for all $r \geq 0$,

$$\mathbb{P}\left(\mathrm{dist}(X, H_\theta) \geq r/\sqrt{D}\right) \xrightarrow[D\to\infty]{} \mathbb{P}(|Z| \geq r) \leq e^{-r^2/2},$$

  where $Z \sim \mathcal{N}(0,1)$.

- Let $X, Y \sim \mathrm{S}_2^{D-1}(0,1)$ be independent. Then when $D \to \infty$, they are almost orthogonal. That is, for all $r \geq 0$,

$$\mathbb{P}\left(|\langle X, Y\rangle| \geq r/\sqrt{D}\right) \xrightarrow[D\to\infty]{} \mathbb{P}(|Z| \geq r) \leq e^{-r^2/2},$$

  where $Z \sim \mathcal{N}(0,1)$.

The first point can intuitively be understood from the fact that in high dimension, $X \in \mathrm{S}_2^{D-1}(0,1)$ has more and more space to be far from θ, and is hence almost orthogonal to H_θ.

Proof.

Since $\mathrm{dist}(X, H_\theta) = |\langle \theta, X\rangle|$, the first point follows from the third point Proposition 1.1. The second one is deduced after applying the first one conditionally on Y (or equivalently by applying Fubini-Tonelli).

1.2 Statistical curse of dimensionality

1.2.1 Non-parametric regression

Assume that we observe iid couples of random variables $(X_1, Y_1), \dots, (X_n, Y_n) \in [0,1]^D \times \mathbb{R}$ with $Y_i = \eta(X_i) + \varepsilon_i$ and unknown *regression function* $\eta : [0,1]^D \to \mathbb{R}$ to estimate. To estimate $\eta(x)$ for a given $x \in [0,1]^D$, a simple strategy consists in doing a weighted average of the output values Y_i's associated to inputs X_i's nearby x. The notion of "X_i close to x" can, for instance, be measured through a user-defined bandwidth parameter $h > 0$, or a number k of neighbors to take into account. The two associated estimators are:

- The Nadaraya-Watson estimator

$$\hat{\eta}_h^{\mathrm{NW}}(x) := \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{K\left(\frac{x-X_i}{h}\right)},$$

  where $h > 0$ is a bandwidth and $K : \mathbb{R}^D \to \mathbb{R}$ is a kernel.

- The Nearest neighbors estimators

$$\hat{\eta}_k^{(\mathrm{NN})}(x) := \frac{1}{k} \sum_{i \in N_k(x)} Y_i,$$

  where $k \in \mathbb{N}$ is a number of neighbors, and $N_k(x)$ is the index set of the j nearest neighbors of $x \in \mathbb{R}^D$ among $\{X_1, \dots, X_n\}$.

Let us develop on the later strategy by providing an integrated error bound on the nearest-neighbors estimator. See Figure 1.1 for a visual representation of the estimator.

Proposition 1.3. (Risk bound for nearest-neighbors regressor) We observe pairs $(X_i, Y_i)_{i \le n}$ with $Y_i = \eta(X_i) + \epsilon_i$. Assume that:

- *(Design)* $X_1, \dots, X_n \sim_{iid} P_X = f(x)\mathrm{d}x$ with $f : [0,1]^D \to \mathbb{R}_+$ such that $\inf_{[0,1]^D} f \ge a > 0$.
- *(Noise)* $\epsilon_1, \dots, \epsilon_n$ are independent centered random variables, independent of the design points $X_1, \dots, X_n$, and with equal variance $\mathbb{E}[\epsilon_i^2] = \sigma^2$.
- *(Smoothness)* $\eta : [0,1]^D \to \mathbb{R}$ is L-Lipschitz.

For all $x \in [0,1]^D$ and $k \in \mathbb{N}$, let

$$\hat{\eta}_k^{(\mathrm{NN})}(x) := \frac{1}{k} \sum_{i \in N_k(x)} Y_i$$

be the k-nearest neighbor regressor.

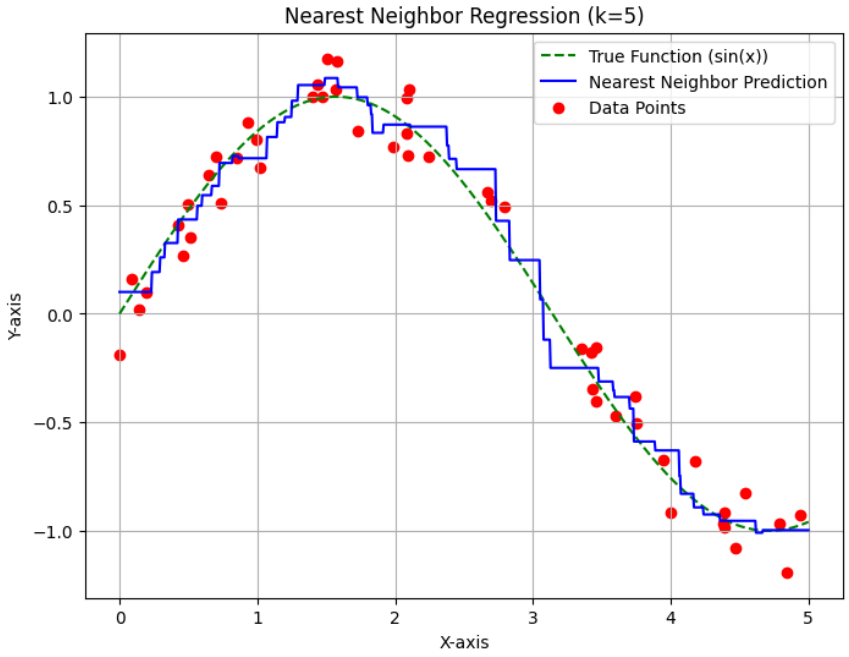

Figure 1.1: A nearest neighbor regressor in dimension $D = 1$.

- For all $n \geq 1$, we have

$$\mathbb{E}\left[(\hat{\eta}_k^{(\mathrm{NN})}(x) - \eta(x))^2\right] \lesssim \frac{DL^2 + \sigma^2}{k} + \left(\frac{k}{a\omega_D n}\right)^{2/D},$$

  up to a numeric constant. Choosing $k = n^{2/(2+D)}$ hence yields

$$\mathbb{E}\left[(\hat{\eta}_k^{(\mathrm{NN})}(x) - \eta(x))^2\right] \lesssim n^{-2/(2+D)},$$

  up to a constant depending on a, L, and D.

- This rate is *minimax-optimal*, in the sense that no estimator can do better simultaneously for all L-Lipschitz regression functions. That is,

$$\inf_{\hat{\eta}_n} \sup_{\eta \in \mathrm{Lip}_L([0,1]^D)} \mathbb{E}\left[(\hat{\eta}_n(x) - \eta(x))^2\right] \gtrsim n^{-2/(2+D)},$$

  where $\hat{\eta}_n$ ranges among all the possible estimators based on a n-sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Proof.

Write $\mathbb{X}_n := (X_1, \ldots, X_n)$. First, apply a bias-variance decomposition conditioned on the

design $\mathbb{X}_n$ to get

$$\begin{aligned}\mathbb{E}(\hat{\eta}_k(x) - \eta(x))^2 &= \mathbb{E}\left[\mathbb{E}\left[(\hat{\eta}_k(x) - \eta(x))^2 \mid \mathbb{X}_n\right]\right]. \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}(\eta(X_{(i)}) - \eta(x)) + \frac{1}{k}\sum_{i=1}^{k}\epsilon_{(i)}\right)^2 \middle| \mathbb{X}_n\right]\right]. \\ &= \mathbb{E}[b_x(\mathbb{X}_n)^2] + \frac{\sigma^2}{k},\end{aligned}$$

where $b_x(\mathbb{X}_n) := \frac{1}{k}\sum_{i=1}^{k}\eta(X_{(i)}) - \eta(x)$. Now analyzing the bias term, we use the Lipschitzness of η to get

$$|b_x(\mathbb{X}_n)| \leq \frac{1}{k}\sum_{i=1}^{k}|\eta(X_{(i)}) - \eta(x)| \leq L\|X_{(k)} - x\|.$$

Let now $\varepsilon > 0$ be such that $p := \mathbb{P}(X_1 \in \mathrm{B}_2^D(x, \varepsilon))$ satisfies $k \leq np/2$. Since $f \geq a$ on $[0,1]^D$, this is always possible as soon as $a\omega_D(\epsilon/2)^D \geq 2k/n$. Using Tchebychev's inequality, we get

$$\begin{aligned}\mathbb{P}\left(\|X_{(k)} - x\| > \varepsilon\right) &= \mathbb{P}(\mathrm{Card}(\mathbb{X}_n \cap \mathrm{B}(x, \varepsilon)) < k) \\ &= \mathbb{P}\left(\mathrm{Bin}(n, p) - np < (k - np/2) - np/2\right) \\ &\leq \mathbb{P}\left(|\mathrm{Bin}(n, p) - np| > np/2\right) \\ &\leq \frac{4}{np} \leq \frac{2}{k}.\end{aligned}$$

Writing $\varepsilon_k := 2(2k/(a\omega_D n))^{1/D}$, we can hence bound the bias term by

$$\begin{aligned}\mathbb{E}\left[\|X_{(k)} - x\|^2\right] &\leq \varepsilon_k^2\mathbb{P}\left(\|X_{(k)} - x\| \leq \varepsilon_k\right) + D\mathbb{P}\left(\|X_{(k)} - x\| > \varepsilon_k\right) \\ &\leq \varepsilon_k + 2D/k,\end{aligned}$$

where we used that $\|X_{(k)} - x\|^2 \leq \mathrm{diam}([0,1]^D)^2 = D$ almost surely, which ends the proof of the upper bound. The minimax-optimality is out of the scope of this class. See [Tsybakov, 2008, Section 2.6.1] for the proof of the last statement, called a *minimax lower bound*.

Remark 1.2. (Numerical application for MNIST dataset) For the MNIST dataset ambient dimension is $D = 28 \times 28 = 784$, and sample size is $n = 60\,000$. For the nearest neighbor or kernel density estimation (KDE), the optimal rate of convergence is given by $n^{-2/(D+2)} = 60\,000^{-2/786} \simeq 0.968$. This result indicates that the convergence rate for high-dimensional data like MNIST is incredibly slow, showing that we are far from the 95% accuracy achieved by modern deep learning techniques like [LeCun et al., 1998].

Smoothness to overcome the curse of dimensionality Statisticians have introduced various approaches to address the curse of dimensionality. One key approach leverages smoothness. The result of Proposition 1.3 can be extended to β-Hölder densities with arbitrary $\beta > 0$,

yielding a convergence rate of order $n^{-2\beta/(2\beta+D)}$. When β scales with D, specifically $\beta = \alpha D$ (i.e., the true regression function is highly smooth), the rate simplifies to $n^{-2\alpha/(2\alpha+1)}$, which is independent of D. Thus, the curse of dimensionality is effectively mitigated, but at a high cost : a very stringent smoothness assumption.

1.2.2 High-dimensional linear regression

Even for parametric problems such as *linear regression*, high dimensions affects the behavior estimators and risk. Consider a linear regression model $Y_i = \langle \beta^*, X_i \rangle + \epsilon_i$, where $\beta^* \in \mathbb{R}^D$ is unknown parameter. Using notation from Section 1.2.1, this amounts to restricting ourselves to regression functions $\eta : \mathbb{R}^D \to \mathbb{R}$ among the parametric family $(\langle \beta, \cdot \rangle)_{\beta \in \mathbb{R}^D}$. The data consists of

- The design matrix $X = (X_1^\top | \cdots | X_n^\top)^\top \in \mathbb{R}^{n \times D}$, where each row corresponds to a sample $X_i \in \mathbb{R}^D$,
- The response vector $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$,
- The noise vector $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^n$, where the ϵ_i's are centered independent real-valued variables with common variance σ^2.

The model writes matricially as

$$Y = X\beta^* + \epsilon.$$

The *least squares estimator*

$$\hat{\beta}_{\mathrm{LS}} \in \arg\min_{\beta \in \mathbb{R}^D} \|Y - X\beta\|^2.$$

On data, the predicted (or denoised) points are $\hat{Y} := X\hat{\beta}$. See Figure 1.2. One can easily give an explicit formula for the error made

Proposition 1.4. (Prediction risk for linear least squares) Adopt the above notation, with X deterministic.

- The least-squares estimator satisfies

$$\mathbb{E}[\|X\hat{\beta}_{\mathrm{LS}} - X\beta^*\|^2] = \sigma^2 \mathrm{rank}(X),$$

  with $\mathrm{rank}(X)$ taking value up to $\min\{n, D\}$.
- This risk is minimax optimal when $\epsilon \sim \mathcal{N}(0, \sigma^2 I_D)$.

Proof.

By construction, $X\hat{\beta}_{\mathrm{LS}}$ corresponds to the orthogonal projection of Y onto the image (or column space) of X, i.e. $X\hat{\beta}_{\mathrm{LS}} = \mathrm{pr}_{\mathrm{Im}X}(Y)$. Hence,

$$X\hat{\beta}_{\mathrm{LS}} - X\beta^* = \mathrm{pr}_{\mathrm{Im}X}(X\beta^* + \epsilon) - X\beta^* = \mathrm{pr}_{\mathrm{Im}X}(\epsilon).$$

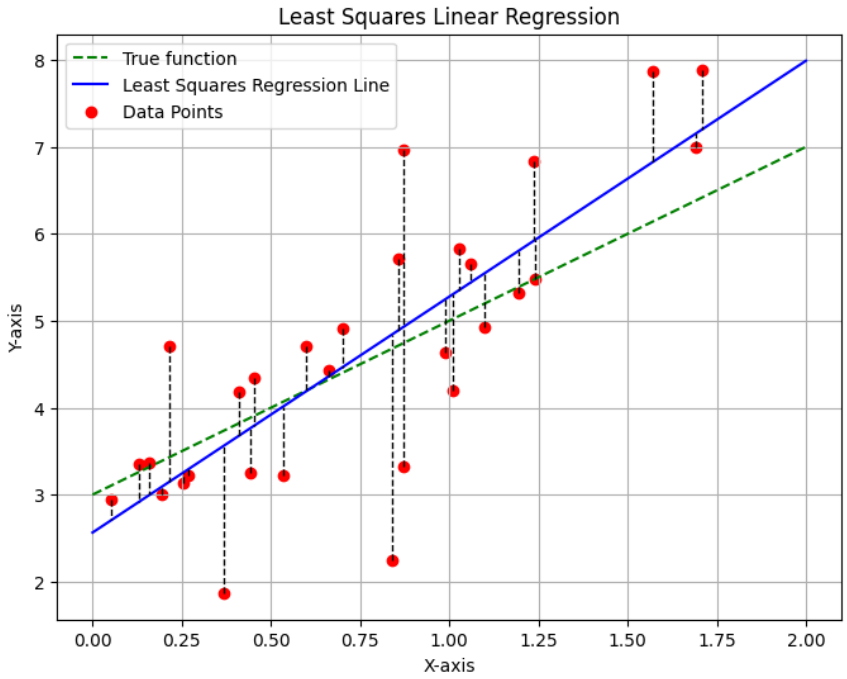

Figure 1.2: Least squares regression in dimension $D = 1$. Vertical lines correspond to the (signed) residuals $\langle X_i, \hat{\beta} - \beta^* \rangle$.

Furthermore, ϵ is as centered random variable with covariance matrix $\sigma^2 I_D$, and hence

$$\begin{aligned}\mathbb{E}[\|X\hat{\beta}_{\text{LS}} - X\beta^*\|^2] &= \mathbb{E}[\|\text{pr}_{\text{Im}X}(\epsilon)\|^2] \\ &= \text{Tr}\left(\text{pr}_{\text{Im}X}(\sigma^2 I_D)\text{pr}_{\text{Im}X}^\top\right) \\ &= \sigma^2 \text{Tr}(\text{pr}_{\text{Im}X}) \\ &= \sigma^2 \text{rank}(X).\end{aligned}$$

The proof of the minimax bound can be found in [Mourtada, 2022]

Here, the risk (in fact, variance) can hence scale linearly with the dimension D. This phenomenon is another example of *curse of dimensionality*: as D increases, the performance of the least-squares estimator degrades, and we need more data to achieve a low error. Even more critically, when $D \geq n$, the model with *overfit* the noise (i.e. $\hat{Y} = Y$), leading to poor generalization on new data.

Sparsity to overcome the curse of dimensionality In high-dimensional settings, methods like *regularization* (e.g., ridge regression or LASSO) constrain the size of the coefficient estimates and controlling overfitting. For instance, one can impose *sparsity* on the model, assuming that only a small subset of the features $(X^{(1)}|\dots|X^{(D)})$ actually influence the response variable Y. In the linear regression model, this means that only a few coefficients $(\beta_1^*, \dots, \beta_D^*)$ are non-zero. One may hence restrict $\beta \in \mathbb{R}^D$ to have at most $d \ll D$ non-zero entries. The sparse estimator

is given by

$$\hat{\beta}_d \in \underset{\|\beta\|_0 \leq d}{\operatorname{argmin}} \|Y - X\beta\|_2^2,$$

where $\|\beta\|_0 := \sum_{i=1}^D \mathbf{1}_{\beta_i \neq 0}$. Theoretical bounds for this estimator are challenging to derive, but adding an ℓ_0-penalty leads to the estimator

$$\hat{\beta}_d \in \underset{\|\beta\|_0 \leq d}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right\},$$

for some $\lambda > 0$. In this case, the following risk bound holds.

Theorem 1.2. ([Giraud, 2021, Theorem 2.2]) There exists a constant C depending on λ, such that if $\|\beta^*\|_0 \leq d$, then

$$\mathbb{E}[\|X\hat{\beta}_d - X\beta^*\|_2^2] \leq C\sigma^2 d \log D.$$

Comparing with Proposition 1.4, the dependence on the dimension is reduced from D to $d \log D$, which is a significant improvement when $d \ll D$. Here, the curse of dimensionality is effectively mitigated, but at two high costs :

- a stringent sparsity assumption on β^*, imposing a very specific structure of $Y \mid X$ in the coordinate system given by X (not invariant by transformations of X);
- a high computational cost, since evaluating $\hat{\beta}_d$ requires computing $\binom{D}{d} \simeq D^d$ least squares estimators, which is in effectively prohibitive. This computational cost can however be reduced drastically, by using a convexified ℓ^1-norm penalisation $\|\beta\|_1$.

1.2.3 Non-parametric density estimation

It appears that the curse of dimensionality is ubiquitous in high-dimensional inference, no matter the quantity to estimate. We conclude this section by considering the *density estimation* problem, with a result similar to Proposition 1.3. See Figure 1.3 for a visual representation of the estimator.

Proposition 1.5. (Kernel density estimation) We observe variables $(X_i)_{i \leq n}$. Assume that:

- *(Design)* $X_1, \dots, X_n \sim_{iid} P_X(\mathrm{d}x) = f(x)\mathrm{d}x$;
- *(Smoothness)* $f : \mathbb{R}^D \to \mathbb{R}$ is L-Lipschitz and bounded by C.

Let $K : \mathbb{R}^D \to \mathbb{R}_+$ be a *kernel* such that $\operatorname{support}(K) \subset \mathrm{B}_2^D(0,1)$, $\int K(z)\mathrm{d}z = 1$, and $v_K := \int K(z)^2 \mathrm{d}z < \infty$. The task is to estimate f at some point $x \in (0,1)^D$. For all $x \in [0,1]^D$ and $h > 0$, let

$$\hat{f}_h^{(\mathrm{KDE})}(x) := \frac{1}{nh^D} \sum_{i \in N_k(x)} K\left(\frac{x - X_i}{h}\right)$$

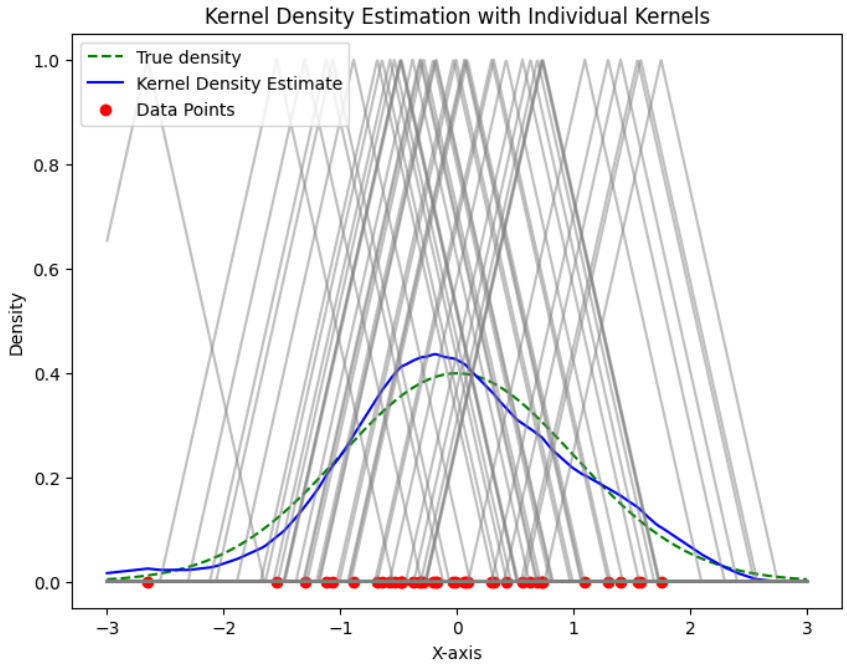

Figure 1.3: A kernel density estimation in dimension $D = 1$, with triangular kernel $K(x) = (1 - |x|)_+$.

be the *kernel density estimator* with kernel K and bandwidth h.

- For all $n \geq 1$, we have

$$\mathbb{E}\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] \leq L^2 h^2 + \frac{C v_K}{n h^D}.$$

  Choosing $h = n^{-1/(2+D)}$ hence yields

$$\mathbb{E}\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] \lesssim n^{-2/(2+D)}.$$

- This rate is *minimax-optimal*, in the sense that no estimator can do better simultaneously for all L-Lipschitz density functions. That is,

$$\inf_{\hat{f}_n} \sup_{f \in \mathrm{Lip}_L([0,1]^D)} \mathbb{E}_f\left[(\hat{f}_n(x) - f(x))^2\right] \gtrsim n^{-2/(2+D)},$$

  where $\hat{f}_n$ ranges among all the possible estimators based on a n-sample $X_1, \ldots, X_n$.

Proof.

Start with the standard bias-variance decomposition

$$\mathbb{E}\left[(\hat{f}_h(x) - f(x))^2\right] = \left(\mathbb{E}[\hat{f}_h(x)] - f(x)\right)^2 + \mathrm{Var}(\hat{f}_h(x)).$$

- *(Variance term)* Since $\hat{f}_h(x)$ is a sum of i.i.d. random variables, its variance is

$$\begin{aligned}\mathrm{Var}(\hat{f}_h(x)) &= \frac{1}{nh^D}\mathrm{Var}\left(K\left(\frac{X-x}{h}\right)\right)\\ &\leq \frac{1}{nh^D}\mathbb{E}\left[K\left(\frac{X-x}{h}\right)^2\right]\\ &\leq \frac{Cv_K}{nh^D},\end{aligned}$$

  where the last line uses the fact that $f \leq C$, and a change of variables, $y = (z-x)/h$.

- *(Bias term)* The same change of variables yields

$$\begin{aligned}\mathbb{E}[\hat{f}_h(x)] - f(x) &= \int_{\mathrm{B}(x,h)} K\left(\frac{z-x}{h}\right)(f(z) - f(x))dz\\ &= \int_{B(0,1)} K(y)(f(x+hy) - f(x))dy.\end{aligned}$$

  Since f is L-Lipschitz, the bias term is bounded by

$$|\mathbb{E}[\hat{f}_h(x)] - f(x)| \leq Lh\int_{B(0,1)} K(y)\|y\|dy \leq Lh.$$

Squaring both sides gives the result.

1.3 Manifold hypothesis

Despite the aforementioned curse of dimensionality, many — sometimes simple and computationally inexpensive — algorithms perform well on high-dimensional data, revealing a clear gap between their empirical success and theoretical expectations.

1.3.1 Differential geometry

Loosely spearking, a submanifold of $\mathbb{R}^D$ is a topological space that locally resembles a Euclidean space of dimension $d \leq D$. It can be thought of as a generalized surface that may be curved, and on which each small local patch looks like a flat regular d-dimensional ball (See Figure 1.4). The formal mathematical definition goes as follows (see Figure 1.5).

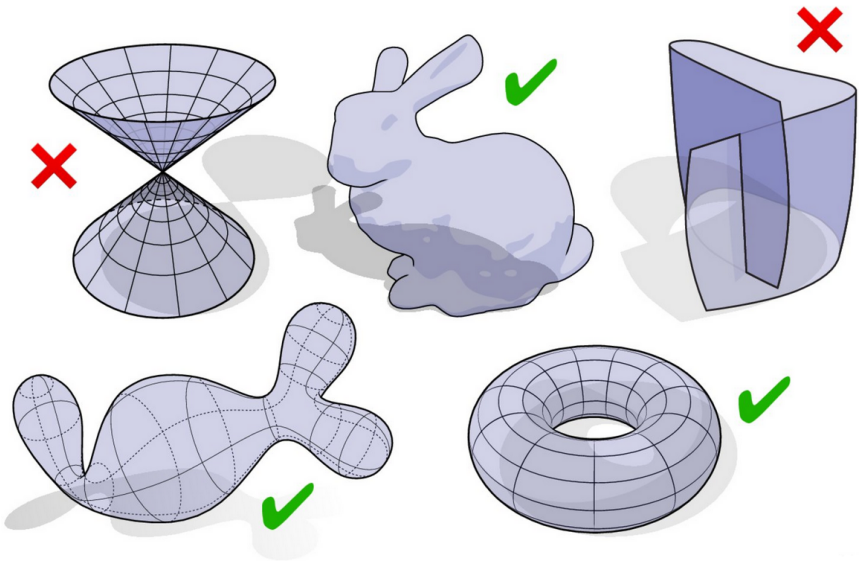

Figure 1.4: Examples and counter-examples of 2-dimensional (sub-)manifolds in ambient dimension $D = 3$. Manifolds cannot self-intersect (upper right), are smooth, and cannot exhibit bifurcating points (upper left). Taken from Keenan Crane's lecture slides.

Definition 1.1. (Submanifold as parametrized spaces) A subset $M \subset \mathbb{R}^D$ is a d-dimensional submanifold of $\mathbb{R}^D$ if for all $p \in M$, there exist an open neighborhood $U \subset \mathbb{R}^D$ of p and a diffeomorphism $\Psi : \mathring{\mathrm{B}}_2^D(0,1) \to U$, such that $\Psi(0) = p$ and

$$M \cap U = \Psi\big(\mathring{\mathrm{B}}_2^d(0,1) \times \{0\}^{D-d}\big).$$

$\dim(M) := d$ is called the *intrinsic dimension* of M, and D its *ambient dimension*.

In the definition, M behaves locally like a d-dimensional Euclidean space within $\mathbb{R}^D$. When restricted to $\mathrm{B}_2^d(0,1) \times \{0\}^{D-d}$, the map Ψ is a *local parametrization* of the set M. It provides local coordinates for the submanifold : its inverse $\Psi^{-1} : U \to \mathrm{B}_2^d(0,1) \times \{0\}^{D-d}$ is called a *chart*. A chart basically straightens the manifold locally to make it fully flat. Statistically speaking, it yields an (unfortunately unknown) local linearization of the dataset. Another equivalent characterization uses implicit equations, which can also be interpreted statistically.

Proposition 1.6. (Submanifolds as implicit equations / level sets) A subset $M \subset \mathbb{R}^D$ is a d-dimensional submanifold of $\mathbb{R}^D$ if and only if for all $p \in M$, there exists an open neighborhood $U \subset \mathbb{R}^D$ containing p and a smooth map $F : U \to \mathbb{R}^{D-d}$ such that $\mathrm{rank}(\nabla_p F) = D - d$ and

$$M \cap U = \{q \in U \mid F(q) = 0\}.$$

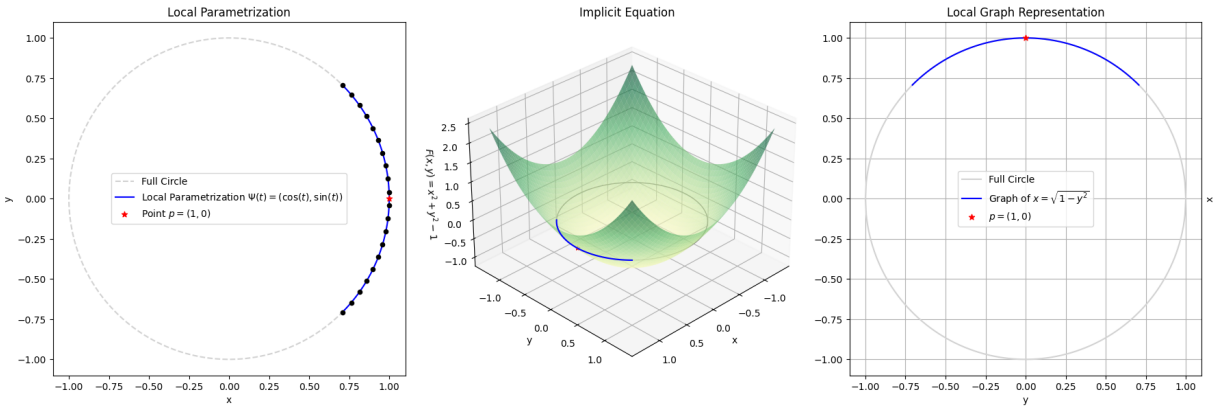

Figure 1.5: The unit circle M is a one-dimensional submanifold of the plane: $d = 1$, $D = 2$. Nearby $p = (1, 0)$: *(left)* M has local parametrization $\Psi(t) = (\cos(4t/\pi), \sin(4t/\pi))$ for $t \in (-1, 1)$; *(middle)* M has implicit equation $F(x, y) = x^2 + y^2 - 1 = 0$; *(right)* M is the graph of $\varphi(y) = \sqrt{1 - y^2}$, y over $y \in (-\sqrt{2}/2, \sqrt{2}/2)$.

The full rank assumption is necessary to guarantee flatness locally. Indeed, the solution set $S := \{((u, v) \in \mathbb{R}^2 \mid uv = 0\} \subset \mathbb{R}^2$ exhibits a crossing at $(0, 0) \in S$. Indeed, the gradient of $F(u, v) := uv$ vanishes at $(u, v) = (0, 0)$, so that the above proposition does not apply.

Thinking about datasets, the function F can be seen as encoding all the non-linear correlations (or implicit dependencies) that M exhibits nearby $p \in M$. The last equivalent way to see submanifolds is through graphs of functions.

Proposition 1.7. (Submanifolds as local rotated graphs) A subset $M \subset \mathbb{R}^D$ is a d-dimensional submanifold of $\mathbb{R}^D$ if and only if for all $p \in M$, there exist an open neighborhood $U \subset \mathbb{R}^D$ containing p, a smooth map $\varphi : \mathring{\mathrm{B}}_2^d(0, 1) \to \mathbb{R}^{D-d}$, and an invertible affine transformation $A : \mathbb{R}^D \to \mathbb{R}^D$ such that

$$M \cap U = \{A(x, \varphi(x)) \mid x \in \mathring{\mathrm{B}}_2^d(0, 1)\}.$$

In this outlook, the affine subspace spanned by $A(\mathring{\mathrm{B}}_2^d(0, 1) \times \{0\}^{D-d}$ explains all the variability of the submanifold M locally. As such, it can be seen as hidden variables parametrizing M. When $\varphi(0_d) = 0_{D-d}$, $\nabla\varphi(0_d) = 0_{(D-d)\times d}$ and $A(0_d, 0_{D-d}) = p$, the affine space $T_pM := A(\mathbb{R}^d \times \{0\}^{D-d})$ is called the *tangent space* at $p \in M$. In practice, this subspace shall not be the first d variables only but a rather complicated subspace which needs to be discovered.

The link between the above three equivalent definition is made via the following two theorems, which are fundamental tools in differential geometry and analysis. For proofs and more, we refer to [Rudin, 1976].

Theorem 1.3. (Implicit Function Theorem) Let U is an open subset of $\mathbb{R}^D$ and $F : U \to \mathbb{R}^{D-d}$ be a smooth map. Let $q = (x_0, y_0) \in U$ where $x_0 \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}^{D-d}$. Assume that $F(q) = 0$, and that the partial differential $\nabla_y F(q) \in \mathbb{R}^{(D-d)\times(D-d)}$ is invertible.

Then there exist open sets $U_1 \subset \mathbb{R}^d$ around x_0, $U_2 \subset \mathbb{R}^{D-d}$ around y_0, and a unique smooth function $\varphi : U_1 \to U_2$ such that $F(x, \varphi(x)) = 0$ for all $x \in U_1$.

In other words, near p, the equation $F(x, y) = 0$ can be solved locally for y as a smooth function of x.

Theorem 1.4. (Inverse Function Theorem) Let U be an open subset of $\mathbb{R}^D$, $p \in U$, and $\Psi : U \to \mathbb{R}^D$ be a smooth map. Assume that the differential $\nabla\Psi(p)$ is invertible.

Then there exists an open subset $U' \subset U$ containing p such that $\Psi : U' \to \Psi(U')$ is a diffeomorphism. (i.e. Ψ is smooth, bijective, with smooth inverse Ψ^{-1})

In other words, if the derivative of a smooth map is invertible at a point, then the map is locally invertible around that point, and the inverse is also smooth.

1.3.2 Non-linear sparsity

When the data lacks a clear low-dimensional parametrization, modeling "sparsity" as in Section 1.2.2 becomes challenging. For example, in the MNIST dataset, it is unreasonable to assume that digit structures depend on a small set of pixels or pixel groups. A natural extension beyond sparsity is to assume the presence of local implicit low-dimensional structures throughout the data. These structures can be described using (sub)manifolds, which generalize curves and surfaces to higher dimensions. The manifold hypothesis posits that:

Observations $X_1, \dots, X_n$ lie on (or close to) an unknown submanifold $M \subset \mathbb{R}^D$. (MH)

Under this hypothesis, *manifold learning* encompasses several tasks.

- *(Standard statistical tasks)* Classification, regression, density estimation, clustering, etc.
- *(Geometric inference)* Estimating features of the unknown manifold M, such as its dimension, topological invariants, intrinsic geodesic distances, etc.
- *(Dimension reduction)* Finding a lower-dimensional representation of the data via mappings $\Psi : \mathbb{R}^D \to \mathbb{R}^{D'}$ with $D' < D$. Ideally the *reduced dimension* D' should be:
  - comparable to $\dim M$ in order to compress information as much as possible, or
  - equal to 2 of 3 to allow for visualization.

  The key goal is that the transformed point cloud $\{\Psi(X_i)\}_{i\le n}$ shares "properties" with the original data $\{X_i\}_{i\le n}$ in a sense to define problem-wise. See Figure 1.6.

Overall, manifold learning seeks to take advantage of the manifold hypothesis, and hopes for performances that depend on the intrinsic dimension d rather than the ambient one $D \gg d$. For example, the rate of estimating an L-Lipschitz density on M improves from $n^{-2/(2+D)}$ to $n^{-2/(2+d)}$. Similarly, the k-nearest neighbor regression rate improves from $n^{-2/(2+D)}$ to $n^{-2/(2+d)}$.

Remark 1.3. (Numerical application for MNIST dataset, but less depressing) Assuming that the MNIST dataset has an intrisic dimension $d < D$ and that the performance of an optimal k-NN classifier is driven by $n^{-2/(2+d)}$, the dimension d corresponding to a 95% accuracy with $n = 60\,000$ observation is $d = 2\log(n)/\log(1/0.05) - 2 \simeq 5 \ll 784$. In fact, some heuristics lean towards a (varying) intrinsic dimension of order 15 for MNIST.

1.3.3 Empirical evidence

The *manifold hypothesis* must not be seen as a universal statistical formalization for describing all modern data. Sometimes, its justification can be handwavy. However, some arguments lean in his camp.

- *(A posteriori reasoning)* As described in Section 1.2, standard statistical tasks become unfeasible in high dimensions. However, common data (image, sound, text) do not fulfill coordinate sparsity whatsoever. Thus, the most plausible explanation for the success of modern machine learning methods on high-dimensional data is the presence of a significantly lower intrinsic dimension. This lower intrinsic dimension is thought to simplify the learning process, making it feasible to achieve effective learning on datasets of manageable size.

- *(Pious wish)* Even though the manifold hypothesis is not fulfilled properly speaking, an approximate version of it (or variants) yield a way out to be able to say something about data. Otherwise, information-theoretic lower-bounds just say no valuable inference can be done. In the same vein, when it comes to visualization of high-dimensional data, where some information loss in unavoidable, but where the constraints on the output are not negotiable.

- *(Numerics)* Certain datasets present invariance properties which actually yield submanifold structures naturally. For instance, consider the "space of images of cats" M_{cat} within the space of images of a given number of pixels $D \gg 1$. One can transform an image by changing its luminosity, balance of colors, orientation, or scale. One can also change the texture of a cat's fur, change slightly its nose location, etc. However, all these transformations leave the *meaning* of an image unchanged.

  As a result, M_{cat} is invariant under certain possibly non-linear transformations of the ambient space. Assuming that all these transformations are smooth, we have just described M_{cat} as (containing) a submanifold of $\mathbb{R}^D$. For sequences of images of the same object, the shooting angle (arrow of time for videos) yields a natural latent variable parametrizing implicitly the dataset. See Figure 1.8.

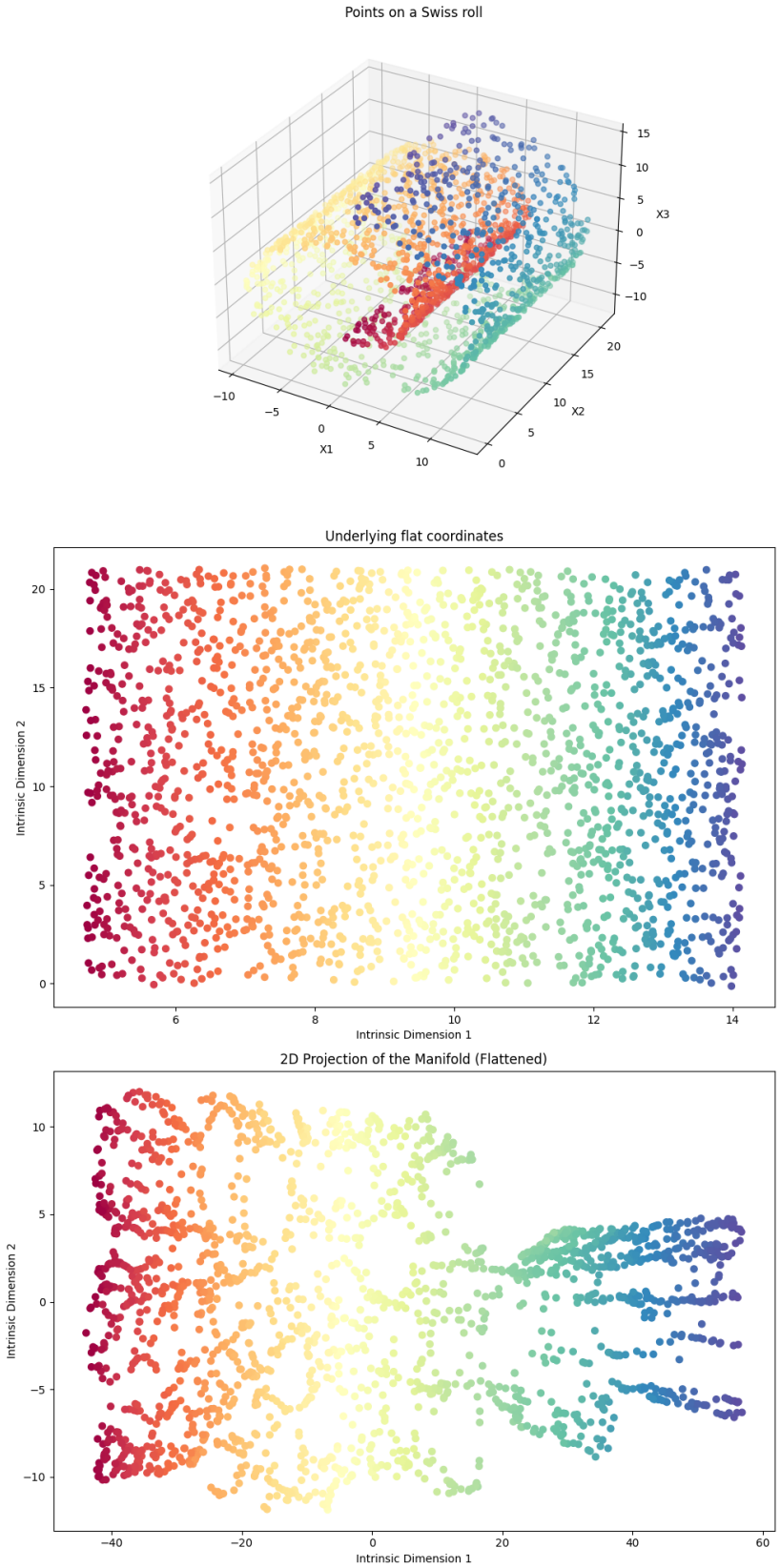

Figure 1.6: The manifold hypothesis assumes that high-dimensional data often lies on a lower-dimensional manifold. The 3D Swiss Roll data actually has an intrinsic 2D structure, which can be recovered using dimensionality reduction techniques like *Isomap*. This algorithm tries to capture the essential metric structure of the data by estimating its geodesic distances.

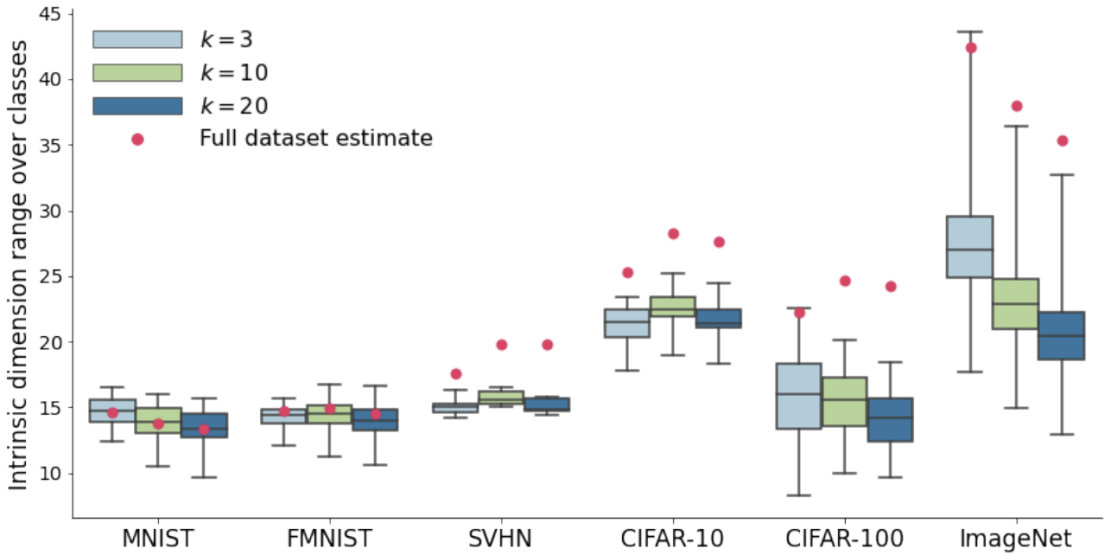

Figure 1.7: Estimated intrinsic dimensionalities of standard image datasets. The estimators used arise from the volume heuristic that $d \sim \log(|\mathrm{B}_2^d(p,r)|)/\log(r)$ for $r \to 0$, with neighborhood measured through k-nearest neighbors. Values appear stable across values of hyperparameter $k \in \{3, 10, 20\}$. Taken from [Brown et al., 2022].

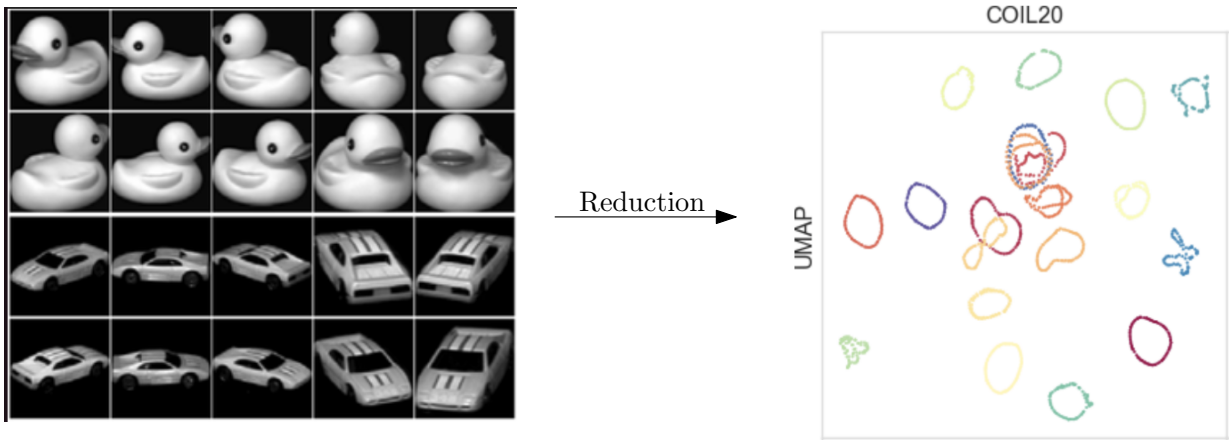

Figure 1.8: The Columbia Object Image Library (Coil20) is a semi-synthetic dataset composed of black & white images of size $D = 128 \times 128$. It contains pictures of 20 different objects, each taken from 72 poses. A non-linear dimension reduction technique (UMAP) reveals a circular structure for each object, corresponding to the shooting angle. Here, UMAP is remarkable in that it does not use any label information, while keeping clusters separated.

CHAPTER 2

Linear algebra refresher

2.1 Orthogonal projectors

We will naturally be required to consider distance-minimizing points ranging over closed subsets K of $\mathbb{R}^D$. That is, for $x \in \mathbb{R}^D$, the *distance from x to K* is

$$\operatorname{dist}(x, K) := \min_{p \in K} \|x - p\|.$$

When K is convex, this minimum is attained at a unique point called the *projection of x onto K*, and defined by

$$\operatorname{pr}_K(x) := \operatorname*{argmin}_{p \in K} \|x - p\|.$$

By construction, the map $\operatorname{pr}_K : \mathbb{R}^D \to K$ clearly satisfies $\operatorname{pr}_K \circ \operatorname{pr}_K = \operatorname{pr}_K$. By convexity of K, it can be characterized geometrically through obtuse angles (see Figure 2.1). That is for all $x \in \mathbb{R}^D$ and $y \in K$,

$$y = \operatorname{pr}_K(x) \iff \text{For all } z \in K, \langle x - y, z - y \rangle \leq 0.$$

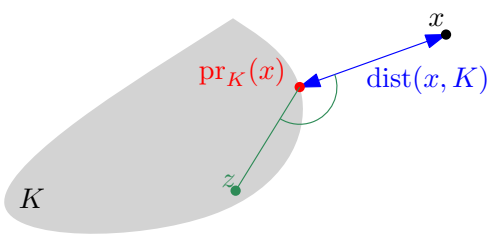

Figure 2.1: Projection onto a closed convex subset in the plane.

When $K = T$ is a linear subspace, the obtuse angle condition (one-sided inequality) becomes an orthogonality condition (equality) by symmetry. Hence, $\mathrm{pr}_T(x)$ is characterized by

$$\text{For all } z \in T, \langle x - \mathrm{pr}_T(x), z - \mathrm{pr}_T(x)\rangle = 0,$$

and by stability of T under subtractions,

$$\text{For all } z \in T, \langle x - \mathrm{pr}_T(x), z\rangle = 0.$$

We hence obtain that pr_T is a linear map, which coincides with the identity on T and is identically equal to zero on $T^\perp$. We shall take these properties as an axiomatic definition of an orthogonal projector.

Definition 2.1. (Orthogonal projector) An *orthogonal projection* onto a linear subspace $T \subset \mathbb{R}^D$ is a linear map $\mathrm{pr}_T : \mathbb{R}^D \to \mathbb{R}^D$ such that:

- $\mathrm{pr}_T \circ \mathrm{pr}_T = \mathrm{pr}_T$;
- $\mathrm{pr}_T(x) = x$ for all $x \in T$;
- $\mathrm{pr}_T(x) = 0$ for all $x \in T^\perp$.

From the definition, we see that $\mathrm{Id} - \mathrm{pr}_T = \mathrm{pr}_{T^\perp}$. Matricially, if $d = \dim(T)$ and $U := (u_1|\dots|u_d) \in \mathbb{R}^{D\times d}$ is an orthonormal basis of T, the orthogonal projection onto T can be written as:

$$\mathrm{pr}_T(x) := \sum_{i=1}^{d} \langle u_i, x\rangle u_i = \sum_{i=1}^{d} u_i u_i^\top x = UU^\top x.$$

Hence, $UU^\top$ is the matrix form of pr_T in the standard basis of $\mathbb{R}^D$. It is a symmetric positive semi-definite matrix of rank d, having eigenvalue 1 with multiplicity d and 0 with multiplicity $D - d$. The corresponding eigenspaces are T and $T^\perp$. For non-orthogonal bases, the projector writes as follows.

Proposition 2.1. (Orthogonal projector onto a column space) If $U = (u_1|\cdots|u_d) \in \mathbb{R}^{D\times d}$ has full rank d, the matrix of the orthogonal projection pr_T is given by $U(U^\top U)^{-1}U^\top$.

Proof.

Since U has full rank $\min\{d, D\} = d$, the matrix $U^\top U \in \mathbb{R}^{d\times d}$ is invertible, so that $H := U(U^\top U)^{-1}U^\top$ is well defined. It is straightforward to verify that $H^2 = H$, meaning that H is a projection. Similarly, $H^\top = H$, meaning that this projection is orthogonal. Finally, we have $\mathrm{Im}(H) \subset \mathrm{Im}(U) = T$, and on the other hand $HU = U$ so that $T \subset \mathrm{Im}(H)$, completing the proof.

2.2 Singular value decomposition

2.2.1 Construction and interpretation

Recall that the *orthogonal group* $\mathcal{O}(\mathbb{R}^k)$ in dimension k is the group of distance-preserving transformations of $\mathbb{R}^k$ letting 0_k fixed. In coordinates, it can be identified to the matrix set

$$\mathcal{O}(\mathbb{R}^k) := \{Q \in \mathbb{R}^{k\times k} \mid Q^\top Q = QQ^\top = I_k\}$$

The *singular value decomposition* asserts that any linear operator is the composition of a rigid transform of the input space, a diagonal operator, and a rigid transform of the output space. In signal processing it is known as *Kosambi–Karhunen–Loève decomposition*, and in image analysis as *Hotelling transform*. It will be at the core of many of the spectral methods to come.

Theorem 2.1. (Singular Value Decomposition (SVD)) Let $A \in \mathbb{R}^{n\times m}$ and $\ell := \min(n, m)$.

- There exist orthogonal matrices $U \in \mathcal{O}(\mathbb{R}^n)$ and $V \in \mathcal{O}(\mathbb{R}^m)$, along with real numbers $s_1 \geq \ldots \geq s_\ell \geq 0$, such that
$$A = USV^\top,$$
where $S = \mathrm{diag}(s_1, \ldots, s_\ell) \in \mathbb{R}^{n\times m}$.
- The reals $s_1, \ldots, s_\ell$ do not depend on the choice of U and V. They are called the *singular values* of A and are denoted by $s_1(A), \ldots, s_\ell(A)$.
- The matrices $U = (u_1|\cdots|u_n)$ and $V = (v_1|\cdots|v_m)$ are not generically unique.
  - Vectors u_i's are called *left-singular vectors* (or *principal components*)
  - Vectors v_j's are called *right-singular vectors* (or *principal axes*)

Remark 2.1. (Reduced SVD) To compress information and storage, some works do not extend the SVD to the null space of A, but work with the so-called *reduced* SVD instead. That is, if $r := \mathrm{rank}(A)$, write $A = USV^\top$ with $U \in \mathbb{R}^{n\times r}$ such that $U^\top U = I_r$, $V \in \mathbb{R}^{m\times r}$ such that $V^\top V = I_r$, and $S = \mathrm{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r\times r}$ with $s_1 \geq \ldots \geq s_r > 0$. Here, $U = (u_1|\cdots|u_r)$ is an orthogonal basis of $\ker(A)^\perp$, and $V = (v_1|\cdots|v_r)$ an orthogonal basis of $\mathrm{Im}(A)$.

Proof.

If $n = 1$ or $m = 1$, the result is trivial. Otherwise, let $v_1 \in \mathbb{R}^m$ be a unit-norm vector such that $v_1 \in \mathrm{argmax}_{\|v\|=1} \|Av\|$, and set $s_1 := \|Av_1\|$.

- If $s_1 = 0$, then $A = 0$, and the result follows.
- If $s_1 > 0$, set $u_1 := s_1^{-1} Av_1 \in \mathbb{R}^n$, which is also a unit-norm vector. Now, complete $(v_1) \in$

$\mathbb{R}^{m\times 1}$ and $(u_1) \in \mathbb{R}^{n\times 1}$ to form orthonormal bases V and U of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Since $Av_1 = s_1u_1$ by construction, we can write $u_1^\top Av_1 = s_1$, and hence by orthogonality of the columns of U,

$$A_1 := U^\top AV = \begin{pmatrix} s_1 & t^\top \\ 0 & B \end{pmatrix},$$

for some $t \in \mathbb{R}^{m-1}$ and $B \in \mathbb{R}^{(n-1)\times(m-1)}$. Letting $w := (s_1, t^\top)^\top \in \mathbb{R}^m$, we have

$$A_1 w = \begin{pmatrix} s_1^2 + \|t\|^2 \\ Bt \end{pmatrix} \in \mathbb{R}^n,$$

so that the unit vector $w_0 := w/\|w\|$ satisfies

$$\|A_1 w_0\|^2 = \frac{(s_1^2 + \|t\|^2)^2 + \|Bt\|^2}{s_1^2 + \|t\|^2} \geq s_1^2 + \|t\|^2.$$

On the other hand,

$$\|A_1 w_0\|^2 = \|U^\top AV w_0\|^2 = \|A(V w_0)\|^2 \leq \max_{\|v\|=1} \|Av\| = s_1^2,$$

since $\|V w_0\|^2 = \|w_0\|^2 = 1$. This implies that $t = 0$, and hence that

$$A = U \begin{pmatrix} s_1 & 0 \\ 0 & B \end{pmatrix} V^\top,$$

for some $B \in \mathbb{R}^{(n-1)\times(m-1)}$. The result then follows by induction.

Remark 2.2. (Practical computation of a SVD) The SVD of a matrix $A \in \mathbb{R}^{m\times n}$ can theoretically be computed by diagonalizing both $AA^\top = US^2U^\top$ and $A^\top A = VS^2V^\top$. However, in practice, this method is prone to numerical instability and loss of precision. Therefore, more robust approaches based on numerical algorithms are commonly used in practice. For instance, a more stable two-step algorithm (see [Golub and Van Loan, 2013, Section 8]) is the following.

- *(Orthogonal bidiagonalization)* Using iterations of *Householder* reflections, first reduce A to a bidiagonal (or tridiagonal) form. That is, compute orthogonal matrices $U_0 \in \mathcal{O}(\mathbb{R}^n)$ and $V_0 \in \mathcal{O}(\mathbb{R}^m)$ such that the matrix $B := U_0^\top AV_0$ is a banded matrix. This pre-processing step simplifies the structure of the matrix, making the computation of singular values more stable and efficient.
- *(Iterative Algorithm for Bidiagonal Matrices)* Compute the SVD of B using an iterative method, often based on a variant of the QR algorithm. For this, Golub and Kahan developed a specific QR-based algorithm that is particularly effective for bidiagonal or tridiagonal matrices. For large ($m, n \gg 1$) or unbalanced matrices ($m \gg n$ or $n \gg m$), a

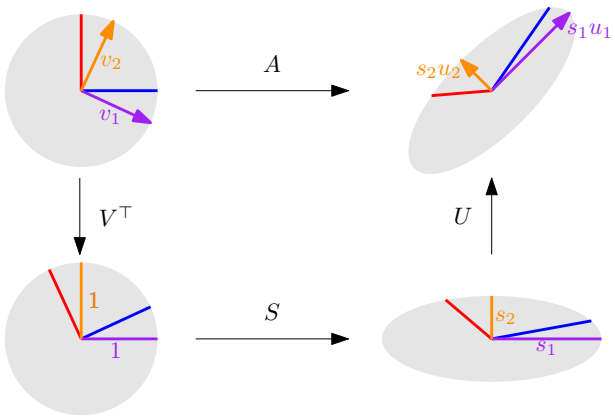

Figure 2.2: A geometric interpretation of the singular value decomposition $USV^\top$ for $A \in \mathbb{R}^{2\times 2}$ through the transformation of the unit ball.

parallelizable divide and conquer strategy can also be used.

These computational methods yield highly stable outputs making the computation of an SVD efficient in practice. Overall, the time complexity if of order $O(nm\min\{n,m\})$ for the full SVD (see [Cline and Dhillon, 2006]), and it reduces to $O(nmk)$ for the computation of the first k singular features only.

From the proof above, it is clear that $s_1(A) = \|A\|_{\text{op}}$. Additionally, we see that

$$\|A\|_F^2 = \text{Tr}(A^\top A) = \text{Tr}(VS^\top SV^\top) = \text{Tr}(S^\top S) = \sum_{k=1}^{\ell} s_k(A)^2.$$

Corollary 2.1. For all $A \in \mathbb{R}^{n\times m}$, $\|A\|_{\text{op}} \le \|A\|_F \le \text{rank}(A)\cdot\|A\|_{\text{op}}$.

A SVD of A provides a geometric interpretation of how A acts as a linear map between $\mathbb{R}^m$ and $\mathbb{R}^n$ (see Figure 2.2). If $u_1,\dots,u_n$ are the columns of U and $v_1,\dots,v_m$ are the columns of V, then for all $k \le \ell$,

$$Av_k = s_k(A)u_k \quad \text{and} \quad A^\top u_k = s_k(A)v_k$$

and these quantities are zero for $k > \ell$. Hence, we have

$$A = \sum_{k=1}^{r} s_k(A)u_k v_k^\top,$$

where $r := \text{rank}(A)$. Similarly, the null space and image of A can be written as:

$$\text{Ker}(A) = \text{Span}\{v_{r+1},\dots,v_m\} \quad \text{and} \quad \text{Im}(A) = \text{Span}\{u_1,\dots,u_r\}.$$

Note that $s_1(A)^2, \dots, s_\ell(A)^2$ are the eigenvalues of $A^\top A$ and $AA^\top$, with corresponding eigenvector bases V and U, respectively:

$$s_k(A)^2 = \lambda_k(AA^\top) = \lambda_k(A^\top A) = s_k(A^\top)^2.$$

Indeed, $A^\top A = V\Lambda V^\top$ and $AA^\top = U\Lambda U^\top$, with $\Lambda := \mathrm{diag}(s_1(A)^2, \dots, s_\ell(A)^2) \in \mathbb{R}^{\ell\times\ell}$.

2.2.2 Variational formulations and eigenstructures

The singular structure of a matrix can be interpreted geometrically through various optimization problems. These interpretations will be important later on when considering various risk minimization heuristics.

Theorem 2.2. (Courant-Fischer Min-Max Theorem) Let $F_{k,m}$ represent the set of all linear subspaces of $\mathbb{R}^m$ with dimension k (called the Grassmanian of order k of $\mathbb{R}^m$). Then the k-th singular value $s_k(A)$ of $A \in \mathbb{R}^{n\times m}$ writes as

$$s_k(A) = \max_{F\in F_{k,m}} \min_{\substack{x\in F\\ \|x\|=1}} \|Ax\| = \min_{F\in F_{m-k+1,m}} \max_{\substack{x\in F\\ \|x\|=1}} \|Ax\|.$$

Proof.

First notice that $s_k(A)^2 = \lambda_k(A^\top A)$, and that for all x, $\|Ax\|^2 = \langle Ax, Ax\rangle = \langle x, A^\top Ax\rangle$. By posing $M := A^\top A$, the proof can be reduced to the Courant-Fischer Min-Max Theorem for *symmetric matrices*, which states that for all symmetric matrix $M \in \mathbb{R}^{m\times m}$,

$$\lambda_k(M) = \max_{F\in F_{k,m}} \min_{\substack{x\in F\\ \|x\|=1}} \langle Mx, x\rangle = \min_{F\in F_{m-k+1,m}} \max_{\substack{x\in F\\ \|x\|=1}} \langle Mx, x\rangle.$$

The second equality is derived from applying the first one to $-M$. Indeed, since the eigenvalues are sorted in decreasing order, we have $\lambda_k(-M) = \lambda_{m-k+1}(-M)$. To establish the first one, consider an eigenvector orthogonal basis $(v_1, \dots, v_m)$ of M, fix $k \in \{1, \dots, m\}$, and write $G_k := \mathrm{span}(v_k, \dots, v_m)$.

- For all $F \in F_{k,m}$, we have $\dim(F) + \dim(G_k) = k + (m - k + 1) > m$, so there exists a unit vector $x_0 \in F \cap G_k$. This x_0 belongs to G_k, so it satisfies $\langle x_0, Mx_0\rangle \leq \| M|_G \|_{\mathrm{op}} = \lambda_k(M)$. As

a result,

$$\begin{aligned}\min_{\substack{x\in F\\ \|x\|=1}}\langle Mx,x\rangle &\le \min_{\substack{x\in F\cap G_k\\ \|x\|=1}}\langle Mx,x\rangle\\ &\le \max_{\substack{x\in F\cap G_k\\ \|x\|=1}}\langle Mx,x\rangle\\ &\le \max_{\substack{x\in G_k\\ \|x\|=1}}\langle Mx,x\rangle\\ &\le \lambda_k.\end{aligned}$$

- Conversely, by selecting $F = \operatorname{span}(v_1,\dots,v_k)$, we obtain $\min_{\substack{x\in F\\ \|x\|=1}}\langle Mx,x\rangle = \lambda_k$, which completes the proof.

A direct consequence of the Courant-Fischer Min-Max Theorem is that the singular values are Lipschitz continuous.

Proposition 2.2. (Weyl's Inequality) For all $A, B \in \mathbb{R}^{n\times m}$ and for all $k \in \{1,\dots,\ell\}$ with $\ell := \min(n,m)$, we have:

$$|s_k(A) - s_k(B)| \le \|A-B\|_{\mathrm{op}}.$$

Proof.

Let $x \in \mathbb{R}^m$ have unit norm. Then

$$\|Ax\| \le \|Bx\| + \|(A-B)x\| \le \|Bx\| + \|A-B\|_{\mathrm{op}},$$

and the result follows by Courant-Fischer theorem.

As we have observed, the first singular value $s_1(A)$ of A is simply $\|A\|_{\mathrm{op}}$. More generally, we have the following characterization by small-rank approximation of A.

Theorem 2.3. (Eckhart-Young Theorem) For all $A \in \mathbb{R}^{n\times m}$ and $t \in \{1,\dots,\ell\}$ with $\ell = \min(n,m)$, we have

$$s_t(A) = \min_{\substack{B\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(B)=t-1}} \|A-B\|_{\mathrm{op}} = \|A - A_{t-1}\|_{\mathrm{op}},$$

where $A_{t-1} := \sum_{k=1}^{t-1} s_k(A)u_kv_k^\top$.

Proof.

Let B be a matrix of rank $t-1$. Then $s_t(B) = 0$. Using Weyl's inequality, we get

$$\|A-B\|_{\mathrm{op}} \ge |s_t(A) - s_t(B)| = s_t(A).$$

Conversely, the equality is achieved when $B = A_{t-1}$, since

$$A - B = \sum_{k=t}^{\ell} s_k(A) u_k v_k^\top$$

has singular values $s_1(A) \geq \ldots \geq s_t(A) \geq 0 = \ldots = 0$. This hence gives

$$\|A - B\|_{\text{op}} = s_1(A - B) = s_t(A).$$

As just shown, A_{t-1} is the best approximation of A of rank $t-1$ in operator norm.

Proposition 2.3. (Frobenius version of Eckhart-Young) For all $A \in \mathbb{R}^{n\times m}$ and $t \in \{1, \ldots, \ell\}$ with $\ell := \min(n, m)$, we have

$$\sum_{k=t}^{\ell} s_k(A)^2 = \min_{\substack{B\in\mathbb{R}^{n\times m} \\ \text{rank}(B)=t-1}} \|A - B\|_F^2 = \|A - A_{q-1}\|_F^2,$$

where $A_{t-1} = \sum_{k=1}^{t-1} s_k(A) u_k v_k^\top$.

In fact, A_{t-1} is also its best approximation in any *Schatten p-norm*, which are defined for all $p \geq 1$ by

$$\|A\|_p := \left(\sum_{k=1}^{\ell} s_k(A)^p \right)^{1/p},$$

with $\|A\|_\infty = \|A\|_{\text{op}}$ being the operator norm, $\|A\|_2 = \|A\|_{\text{F}}$ the Frobenius norm, and $\|A\|_1 = \|A\|_*$ the nuclear norm.

Proof.

The proof follows the lines of that of Eckart-Young, by summing over $k \in \{t+1, \ldots, \ell\}$.

Another way to phrase this proposition using projectors is the following.

Proposition 2.4. (Eckart-Young for Principal Component Analysis) Let $A \in \mathbb{R}^{n\times m}$ have singular value decomposition $A = USV^\top$, and $t \in \{1, \ldots, \ell\}$ with $\ell := \min(n, m)$. Then

$$\sum_{k=t+1}^{\ell} s_k(A)^2 = \min_{\substack{\Phi\in\mathbb{R}^{n\times t} \\ \Phi^\top\Phi=I_t}} \|A - A\Phi\Phi^\top\|_F^2 = \|A - A_t\|_F^2,$$

where $A_t = AV_{*,t}V_{*,t}^\top$ and $V_{*,t} = (v_1|\cdots|v_t) \in \mathbb{R}^{n\times t}$ is composed of the t first columns of V.

Proof.

Simply notice that if $\Phi \in \mathbb{R}^{n\times t}$, $A\Phi\Phi^\top$ has rank at most t. Furthermore, one easily checks that the optimum A_t of Proposition 2.3 writes as $AV_{*,t}V_{*,t}^\top$, which yields the result.

A geometric way to interpret this result is to consider the rows of A and their projections.

Remark 2.3. (Geometric interpretation of Eckart-Young)

- As seen in Section 2.1, if $\Phi = (\phi_1|\cdots|\phi_t) \in \mathbb{R}^{m\times t}$ with $\Phi^\top\Phi = I_t$, then $\Phi\Phi^\top \in \mathbb{R}^{m\times m}$ is the orthogonal projector onto the t-dimensional linear subspace $\mathrm{Im}(\Phi) = \mathrm{span}(\phi_1, \dots, \phi_t) \subset \mathbb{R}^m$.
- Writing $A = (a_1^\top|\cdots|a_n^\top)^\top$ for the row-wise decomposition of A, we recognize

$$\begin{aligned}\|A - A\Phi\Phi^\top\|_F^2 &= \sum_{i=1}^n \|a_i - \mathrm{pr}_{\mathrm{Im}\Phi}(a_i)\|^2 \\ &= \sum_{i=1}^n \mathrm{dist}(a_i, \mathrm{Im}(\Phi))^2\end{aligned}$$

  as the cumulated squared distances of $a_1, \dots, a_n$ to $\mathrm{Im}(\Phi)$

2.3 Eigendecompositions and Rayleigh quotients

2.3.1 Eigenvalue problems

The *eigenvalue problem* for a symmetric matrix $M \in \mathbb{R}^{m\times m}$ involves finding *eigenvectors* ϕ_k and *eigenvalues* λ_k such that

$$M\phi_k = \lambda_k\phi_k, \quad \forall k \in \{1, \dots, m\},$$

where λ_i are the eigenvalues and ϕ_i are the corresponding eigenvectors. In matrix form, this is expressed as

$$M\Phi = \Phi\Lambda,$$

where $\Phi = (\phi_1|\cdots|\phi_m) \in \mathbb{R}^{m\times m}$ is the matrix whose columns are the eigenvectors, and $\Lambda \in \mathbb{R}^{m\times m}$ is the diagonal matrix whose diagonal elements are the eigenvalues $\lambda_1, \dots, \lambda_m$. Symmetry of M ensures that the eigenvalues are real, and that the eigenvectors corresponding to distinct eigenvalues are orthogonal.

By homogeneity of the norm, the Courant-Fischer min-max theorem for $t = 1$ can be rewritten as the *Rayleigh quotient*

$$\lambda_1(M) = \max_{x\in\mathbb{R}^m} \frac{\langle Mx, x\rangle}{\langle x, x\rangle}.$$

We recover the variational formula for the first eigenvector of a symmetric matrix $M \in \mathbb{R}^{m\times m}$, also known as the *spectral theorem*. Furthermore, from the method of Lagrange multipliers, one easily sees the following.

Proposition 2.5. (Orthodiagonalization of symmetric matrices) If $M \in \mathbb{R}^{m\times m}$ is symmetric, then a sequence $v_1, \dots, v_t$ such that

$$v_j \in \underset{\substack{\langle x,x\rangle=1 \\ \forall i<j,\, \langle v_i,x\rangle=0}}{\operatorname{argmax}} \langle Mx, x\rangle$$

yields an orthonormal diagonalization family $(v_1, \dots, v_t)$ associated to $\lambda_1(M) \geq \dots \geq \lambda_t(M)$.

Proof.

We will make use of the following version of Lagrange multipliers, which can be found in [Boyd and Vandenberghe, 2004].

> Theorem 2.4. (Lagrange Multiplier Theorem) Let $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ be continuously differentiable functions, with f *coercive*, i.e. $|f(x)| \xrightarrow[\|x\|\to\infty]{} \infty$. Then for all
>
> $$x_* \in \underset{\substack{x\in\mathbb{R}^m \\ g(x)=0}}{\operatorname{argmax}} f(x),$$
>
> there exists a *vector of Lagrange multipliers* $\lambda_* \in \mathbb{R}^k$ such that
>
> $$\nabla f(x_*) = \lambda_*^\top \nabla g(x_*).$$

To prove the result, note that the result is trivial for $m = 1$. Otherwise, take $f(x) := \langle M^\top x, x\rangle$ and $g(x) := \langle x, x\rangle$. Their respective gradients are $\nabla f(x) = 2x^\top M$ and $\nabla g(x) = 2x^\top$. From Lagrange multipliers, for all $v_1 \in \operatorname{argmax}_{g(x)=1} f(x)$, there exists $\lambda_1 \in \mathbb{R}$ such that $Mv_1 = 2\lambda_1 v_1$. Furthermore, setting $A_1 := M - \lambda_1 v_1 v_1^\top$, we see that $A_1 v_1 = 0$, and for all $x \in \mathbb{R}^m$ such that $\langle v_1, x\rangle = 0$, $A_1 x = Ax$ also satisfies $\langle v_1, Ax\rangle = 0$. One can hence restrict A_1 to the $(m-1)$-dimensional subspace $\operatorname{span}(v_1)^\perp \subset \mathbb{R}^m$ and conclude by induction.

As will become clear later, we will naturally be led to consider sequential optimization problems such as that of Proposition 2.5. In matrix form, one can write $V_{*,t} = (v_1|\cdots|v_t) \in \mathbb{R}^{m\times t}$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_t) \in \mathbb{R}^{t\times t}$, so that the top t eigenstructure of M is summarized as the truncated diagonalization

$$MV_{*,t} = V_{*,t}\Lambda_t.$$

Similar to Proposition 2.4, let us give a variational formulation of the eigenstructure of a symmetric matrix.

Theorem 2.5. (Variational properties of eigenstructures) Let $M \in \mathbb{R}^{m\times m}$ be a symmetric matrix. Write $v_1, \dots, v_m$ for an orthonormal diagonalization family of M, with associated eigenvalues $\lambda_1(M) \geq \dots \geq \lambda_m(M)$. For $t \leq m$, set $V_{*,t} := (v_1|\cdots|v_t) \in \mathbb{R}^{m\times t}$. Then $V_{*,t}$ solves

the quadratic problem

$$\sum_{k=1}^{t} \lambda_k = \max_{\substack{\Phi \in \mathbb{R}^{m\times t} \\ \Phi^\top \Phi = I_t}} \operatorname{Tr}(\Phi^\top M \Phi) = \operatorname{Tr}(V_{*,t}^\top M V_{*,t})$$

2.3.2 Generalized eigenvalue problems

The *generalized eigenvalue problem* for a pair of symmetric matrices $M, N \in \mathbb{R}^{m\times m}$ involves finding *generalized eigenvectors* ϕ_k and *generalized eigenvalues* λ_k such that

$$M\phi_k = \lambda_k N \phi_k, \quad \forall k \in \{1, \dots, m\},$$

where λ_i are the generalized eigenvalues, and ϕ_i are the corresponding generalized eigenvectors. In this formulation, N may be assumed to be positive definite, ensuring certain desirable properties of the problem.

Remark 2.4. (Comparing with classical eigenstructures) A few remarks are in order.

- For vectors $\phi \in \ker(N)^\perp$, the problem likely reduces to a regular eigenvalue problem $N^{-1}M\phi = \lambda\phi$, reducing it to an ordinary eigenvector equation.
- For vectors $\phi \in \ker(M)^\perp$, you can likewise write $\lambda^{-1}\phi = M^{-1}N\phi$, again forming a regular eigenvalue problem.
- Any non-zero vector $\phi_0 \in \ker(M) \cap \ker(N)$ is a generalized eigenvector for any generalized eigenvalue. Furthermore, if ϕ is a generalized eigenvector with eigenvalue λ, then $\phi + \phi_0$ is also a generalized eigenvector with the same eigenvalue.
- If $\phi \in \ker(M) \setminus \ker(N)$, it is a generalized eigenvector with eigenvalue 0. In contrast, if $\phi \in \ker(N) \setminus \ker(M)$ lies in the null space of N but not M, it cannot be a generalized eigenvector.

In matrix form, a generalized eigenvalue problem can be expressed as

$$M\Phi = N\Phi\Lambda,$$

where $\Phi = (\phi_1 | \cdots | \phi_m) \in \mathbb{R}^{m\times m}$ is the matrix whose columns are the generalized eigenvectors, and $\Lambda \in \mathbb{R}^{m\times m}$ is the diagonal matrix whose diagonal elements are the generalized eigenvalues $\lambda_1, \dots, \lambda_m$. If both matrices M and N are symmetric, the generalized eigenvalues are real, and the generalized eigenvectors corresponding to distinct eigenvalues are N-orthogonal, meaning that for distinct $i \neq j \in \{1, \dots, m\}$,

$$\langle N\phi_i, \phi_j \rangle = 0.$$

Extending the Courant-Fischer min-max theorem, the largest generalized eigenvalue $\lambda_1(M,N)$ can be characterized variationally as

$$\lambda_1(M,N) = \max_{x\in\mathbb{R}^m} \frac{\langle Mx,x\rangle}{\langle Nx,x\rangle}.$$

Proposition 2.6. (N-Orthodiagonalization of symmetric matrices) If $M,N\in\mathbb{R}^{m\times m}$ are symmetric, then a sequence $v_1,\dots,v_t$ such that

$$v_j \in \operatorname*{argmax}_{\substack{\|x\|_N=1\\ \forall i<j,\ \langle v_i,Nx\rangle=0}} \langle Mx,x\rangle,$$

yields an N-orthonormal family $(v_1,\dots,v_t)$, which diagonalizes M with respect to N, where $\|x\|_N := \langle Nx,x\rangle^{1/2}$ represents the semi-norm induced by the matrix N.

Proof.

Follow the proof of Proposition 2.5 by induction with To prove the result, note that the result is trivial for $m=1$. Otherwise, take $f(x) := \langle M^\top x,x\rangle$ and $g(x) := \langle N^\top x,x\rangle$. From Lagrange multipliers, for all $v_1\in\operatorname{argmax}_{g(x)=1} f(x)$, there exist $\lambda_1\in\mathbb{R}$ such that $2Mv_1 = 2\lambda Nv_1$. Then set $A_1 := M-\lambda_1(Nv_1)(Nv_1)^\top$ and conclude by induction on $m\geq 1$.

The sequence from Proposition 2.6 gives the truncated generalized eigenvalue decomposition

$$MV_{*,k} = NV_{*,k}\Lambda_k,$$

where $V_{*,k} = (v_1|\cdots|v_k)\in\mathbb{R}^{m\times k}$ and $\Lambda_k = \operatorname{diag}(\lambda_1,\dots,\lambda_k)\in\mathbb{R}^{k\times k}$. Thus, similar to the classical eigenvalue problem, one can derive a variational formulation for the generalized eigenstructure of the symmetric matrix pair (M,N).

Theorem 2.6. (Variational properties of generalized eigenstructures) Let $M,N\in\mathbb{R}^{m\times m}$ be symmetric matrices with N positive definite. Write $v_1,\dots,v_m$ for a N-orthonormal diagonalization family of M, with associated eigenvalues $\lambda_1(M,N)\geq\dots\geq\lambda_m(M,N)$. For $t\leq m$, set $V_{*,t} := (v_1|\cdots|v_t)\in\mathbb{R}^{m\times t}$. Then $V_{*,t}$ solves the quadratic problem

$$\sum_{k=1}^{t}\lambda_k(M,N) = \max_{\substack{\Phi\in\mathbb{R}^{m\times t}\\ \Phi^\top N\Phi=I_t}} \operatorname{Tr}(\Phi^\top M\Phi) = \operatorname{Tr}(V_{*,t}^\top MV_{*,t}),$$

Remark 2.5. (Semi-definite programming) All the above optimization problems (Proposition 2.4, Theorems 2.5 and 2.6) can be cast in a unified family of optimization called *semi-definite programming* (SDP) problems. For instance, from Theorem 2.6 and homogeneity, the largest generalized eigenvalue problem of a symmetric pair (M, N) writes

$$\begin{aligned}\lambda_1(M,N) &= \max_{\substack{x\in\mathbb{R}^m \\ \mathrm{Tr}(x^\top N x)=1}} \mathrm{Tr}(x^\top M x) \\ &= \max_{\substack{x\in\mathbb{R}^m \\ \mathrm{Tr}(N x x^\top)=1}} \mathrm{Tr}(M x x^\top) \\ &= \max_{\substack{X\in\mathbb{R}^{m\times m} \\ \mathrm{Tr}(NX)=1 \\ X\succcurlyeq 0}} \mathrm{Tr}(MX),\end{aligned}$$

where the third equality follows by posing $X = xx^\top$. Note that the objective function is linear, with one constraint being linear and the other one concerns the positive semi-definiteness of the matrix variable. See [Wolkowicz et al., 2012] for an in-depth introduction to SDP.

2.4 Moore-Penrose pseudo-inverse

A matrix $A \in \mathbb{R}^{n\times m}$ defines a linear map from $\mathbb{R}^m$ to $\mathbb{R}^n$, and its restriction from $(\mathrm{Ker}(A))^\perp$ to $\mathrm{Im}(A)$ is an isomorphism. We can hence define its inverse from $\mathrm{Im}(A)$ to $(\mathrm{Ker}(A))^\perp \subset \mathbb{R}^n$ in a classical sense, extend it by 0 on $(\mathrm{Im}(A))^\perp$, and then on $\mathbb{R}^m = \mathrm{Im}(A) \oplus (\mathrm{Im}(A))^\perp$ by linearity. This yields a unique map, denoted by $A^\dagger$, usually referred to as the *Moore-Penrose inverse* of A.

Definition 2.2. (Moore-Penrose pseudo-inverse) If $A = USV^\top$ is the singular value decomposition $A \in \mathbb{R}^{n\times m}$, then the *pseudo-inverse* of A is defined as

$$A^\dagger = VS^\dagger U^\top,$$

where $S^\dagger \in \mathbb{R}^{m\times n}$ is the diagonal matrix with

$$S^\dagger_{i,i} = \begin{cases} s_i(A)^{-1} & \text{if } s_i(A) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Proof.

Straightforward based on the SVD definition.

Proposition 2.7. (Properties of the Moore-Penrose Inverse) For all $A \in \mathbb{R}^{n\times m}$, the following properties hold:

1. $AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$;
2. $AA^\dagger$ is the orthogonal projection onto $\mathrm{Im}(A)$,
3. $A^\dagger A$ is the orthogonal projection onto $(\mathrm{Ker}(A))^\perp$
4. $\mathrm{Im}(A)^\dagger = (\mathrm{Ker}(A))^\perp$ and $\mathrm{Ker}(A)^\dagger = (\mathrm{Im}(A))^\perp$;

Furthermore,

5. If $\mathrm{rank}(A) = m$, then $A^\dagger = (A^\top A)^{-1}A^\top$;
6. If $\mathrm{rank}(A) = n$, then $A^\dagger = A^\top(AA^\top)^{-1}$;
7. If $n = m$ and A is invertible, then $A^\dagger = A^{-1}$.

Proof.

Left as an exercise.

In fact, $A^\dagger$ is the unique matrix of $\mathbb{R}^{m\times n}$ satisfying the first two properties of Proposition 2.7. Note also that point 2. is a generalization of Proposition 2.1, yielding another interpretation of the explicit orthogonal projection matrices.

Example 2.1. (Moore-Penrose is what you need) Let us give a few examples of use of the Moore-Penrose pseudo-inverse.

- *(Solving overdetermined systems)* Given an overdetermined system $Ax = b$ with $A \in \mathbb{R}^{n\times m}$ a tall matrix ($n > m$, more rows than columns), there is no exact solution if $b \notin \mathrm{Im}A$. However, the least-squares solution is given by

$$x_{\mathrm{LS}} := M^\dagger b = \operatorname*{argmin}_{x\in\mathbb{R}^m} \|Ax - b\|^2.$$

  In a linear regression setting, this gives a closed form formula for the least squares estimator.

- *(Solving underdetermined systems)* In an underdetermined system $Ax = b$ with $A \in \mathbb{R}^{n\times m}$ a wide matrix ($m > n$, more columns than rows), there are more unknowns than equations and hence infinitely many solutions. The Moore-Penrose pseudoinverse provides the solution x_{MN} with the smallest norm, that is

$$x_{\mathrm{MN}} := A^\dagger b = \operatorname*{argmin}_{\substack{x\in\mathbb{R}^m \\ Ax=b}} \|x\|.$$

- *(Projection onto a column / null space)* As seen above, the pseudoinverse provides a closed form for the orthogonal projector onto the column space of $a \in \mathbb{R}^{n\times m}$, through the formula

$$\mathrm{pr}_{\mathrm{Im}A} = AA^{\dagger}.$$

  Similarly, the orthogonal projector onto the null space A writes as

$$\mathrm{pr}_{\mathrm{Ker}A} = I_m - A^{\dagger}A.$$

- *(Geometry of generalized eigenvalue problems)* Coming back to the generalized eigenvalue problem $M\phi = \lambda N\phi$ for $\phi \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$, we note that

  i) Any non-zero vector $\phi \in \ker(M) \oplus (\ker(M)^{\perp} \cap \ker(N)) = \ker(M) + \ker(N)$ is an eigenvector associated to the eigenvalue $\lambda = 0$

  ii) For all $\phi \in \ker(M)^{\perp} \cap \ker(N)^{\perp}$,

$$\begin{aligned} M\phi = \lambda N\phi &\Leftrightarrow N^{\dagger}M\phi = \lambda N^{\dagger}N\phi \\ &\Leftrightarrow N^{\dagger}M\phi = \lambda\phi. \end{aligned}$$

  The generalized eigenstructure of (M, N) hence reduces to i) the null space of M and N, and to ii) the non-zero classical eigenstructure of $N^{\dagger}M$.

Bibliography

[Bar and Pozdnyakov, 2024] Bar, H. and Pozdnyakov, V. (2024). High dimensional space oddity.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Brown et al., 2022] Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. (2022). Verifying the union of manifolds hypothesis for image data. *arXiv preprint arXiv:2207.02862*.

[Cline and Dhillon, 2006] Cline, A. K. and Dhillon, I. S. (2006). Computation of the singular value decomposition. In *Handbook of linear algebra*, pages 45–1. Chapman and Hall/CRC.

[Giraud, 2021] Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

[Golub and Van Loan, 2013] Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Mourtada, 2022] Mourtada, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178.

[Rudin, 1976] Rudin, W. (1976). Principles of mathematical analysis. *3rd ed.*

[Tsybakov, 2008] Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.

[Wolkowicz et al., 2012] Wolkowicz, H., Saigal, R., and Vandenberghe, L. (2012). *Handbook of semidefinite programming: theory, algorithms, and applications*, volume 27. Springer Science & Business Media.