Principal Component Analysis

1 Overview

This chapter describes how to analyze a dataset from a geometric perspective using tools inspired by probability theory. Section 2 explains how to characterize the center of a dataset and its fluctuations around it from a probabilistic viewpoint. It also introduces the concept of correlation, which captures to what extent two quantities are linearly related. Section 3 applies these ideas to finite datasets, providing some geometric intuition. Section 4 introduces the covariance matrix, and shows that it captures the average variation of the data in every direction. This motivates Section 5, dedicated to the spectral theorem, which shows that the eigendecomposition of symmetric matrices has a very intuitive geometric interpretation. This interpretation motivates principal component analysis, described in Section 5. Section 7 applies this technique to dimensionality reduction. Finally, Section 7 establishes a connection to Gaussian random vectors.

2 Mean, variance, and correlation

Let us consider real-valued data corresponding to a single quantity or feature. We model such data as a scalar continuous random variable. This is of course an abstraction, in reality we usually have access to a finite number of data points, not to a continuous distribution. The mean of a random variable is the point that minimizes the expected distance to the random variable. Intuitively, it is the center of mass of the probability density, and hence of the dataset.

Lemma 2.1. *For any random variable $\tilde{a}$ with mean $\mathrm{E}(\tilde{a})$,*

$$\mathrm{E}\left(\tilde{a}\right) = \arg\min_{c\in\mathbb{R}} \mathrm{E}\left(\left(c-\tilde{a}\right)^2\right). \tag{1}$$

Proof. Let $g(c) := \mathrm{E}\left((c-\tilde{a})^2\right) = c^2 - 2c\mathrm{E}\left(\tilde{a}\right) + \mathrm{E}\left(\tilde{a}^2\right)$, we have

$$f'(c) = 2(c - \mathrm{E}(\tilde{a})), \tag{2}$$
$$f''(c) = 2. \tag{3}$$

The function is strictly convex and has a minimum where the derivative equals zero, i.e. when c is equal to the mean. □

The variance of a random variable $\tilde{a}$

$$\mathrm{Var}(\tilde{a}) := \mathrm{E}\left(\left(\tilde{a} - \mathrm{E}(\tilde{a})\right)^2\right) \tag{4}$$

quantifies how much it fluctuates around its mean. The standard deviation, defined as the square root of the variance, is therefore a measure of how spread out the dataset is around its center.

Let us consider data containing two features, each represented by a random variable. The covariance of two random variables $\tilde{a}$ and $\tilde{b}$ $\mathrm{Cov}(\tilde{a}, \tilde{b}) := \mathrm{E}\left[(\tilde{a} - \mathrm{E}(\tilde{a})(\tilde{b} - \mathrm{E}(\tilde{b}))\right]$ quantifies their joint fluctuations around their respective means. If we center the random variables, so that their mean is zero, then the covariance is equal to the expected dot product, $\mathrm{Cov}(\tilde{a}, \tilde{b}) = \mathrm{E}(\tilde{a}\tilde{b})$. In fact, the covariance itself is a valid inner product if we interpret the centered random variables as vectors in a vector space. Zero-mean random variables form a vector space because linear combinations of zero-mean random variables are also zero mean. The origin of the vector space (the zero vector) is the random variable equal to zero with probability one. The covariance is a valid inner product because it is (1) symmetric, (2) linear in its first argument, i.e. for any $\alpha \in \mathbb{R}$

$$\mathrm{E}(\alpha\tilde{a}\tilde{b}) = \alpha\mathrm{E}(\tilde{a}\tilde{b}), \tag{5}$$

and (3) positive definite, i.e.

$$\mathrm{E}(\tilde{a}^2) = 0, \tag{6}$$

if and only if $\tilde{a} = 0$ with probability one. To prove this last property, we use a fundamental inequality in probability theory.

Theorem 2.2 (Markov's inequality)**.** *Let $\tilde{r}$ be a nonnegative random variable. For any positive constant $c > 0$,*

$$\mathrm{P}(\tilde{r} \geq c) \leq \frac{\mathrm{E}(\tilde{r})}{c}. \tag{7}$$

Proof. Consider the indicator variable $1_{\tilde{r} \geq c}$. We have

$$\tilde{r} - c\, 1_{\tilde{r} \geq c} \geq 0, \tag{8}$$

which implies that its expectation is nonnegative (it is the sum or integral of a nonnegative quantity). By linearity of expectation and the fact that $1_{\tilde{r} \geq c}$ is a Bernoulli random variable with expectation $\mathrm{P}(\tilde{r} \geq c)$ we have

$$\mathrm{E}(\tilde{r}) \geq c\,\mathrm{E}\left(1_{\tilde{r} \geq c}\right) = c\,\mathrm{P}(\tilde{r} \geq c). \tag{9}$$

□

Corollary 2.3. *If the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then*

$$\mathrm{P}(\tilde{a} \neq 0) = 0. \tag{10}$$

Proof. If $\mathrm{P}(\tilde{a} \neq 0) \neq 0$ then there exists an ϵ such that $\mathrm{P}(\tilde{a}^2 \geq \epsilon) \neq 0$. This is impossible. Applying Markov's inequality to the nonnegative random variable $\tilde{a}^2$ we have

$$\mathrm{P}(\tilde{a}^2 \geq \epsilon) \leq \frac{\mathrm{E}\left(\tilde{a}^2\right)}{\epsilon} \tag{11}$$

$$= 0. \tag{12}$$

□

The standard deviation is the norm induced by this inner product since $\mathrm{Cov}(\tilde{a}, \tilde{a}) = \mathrm{Var}\,(\tilde{a})$.

When comparing two vectors, a natural measure of their similarity is the cosine of the angle between them, which ranges from 1 when they are collinear, through 0 when they are orthogonal, to -1 when they are collinear but point in opposite directions. The cosine equals the inner product between the vectors normalized by their norms. In the vector space of zero-mean random variables this quantity is called the correlation coefficient,

$$\rho_{\tilde{a},\tilde{b}} := \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\sqrt{\mathrm{Var}(\tilde{a})\mathrm{Var}(\tilde{b})}}, \tag{13}$$

and it also ranges between -1 and 1.

Theorem 2.4 (Cauchy-Schwarz inequality for random variables)**.** *Let $\tilde{a}$ and $\tilde{b}$ be two random variables. Their correlation coefficient satisfies*

$$-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1 \tag{14}$$

with equality if and only if $\tilde{b}$ is a linear function of $\tilde{a}$ with probability one.

Proof. Consider the standardized random variables (centered and normalized),

$$\mathrm{s}(\tilde{a}) := \frac{\tilde{a} - \mathrm{E}(\tilde{a})}{\sqrt{\mathrm{Var}(\tilde{a})}}, \qquad \mathrm{s}(\tilde{b}) := \frac{\tilde{b} - \mathrm{E}(\tilde{b})}{\sqrt{\mathrm{Var}(\tilde{b})}}. \tag{15}$$

The mean square distance between them equals

$$\mathrm{E}\left[(\mathrm{s}(\tilde{b}) - \mathrm{s}(\tilde{a}))^2\right] = \mathrm{E}\left(\mathrm{s}(\tilde{a})^2\right) + \mathrm{E}(\mathrm{s}(\tilde{b})^2) - 2\mathrm{E}(\mathrm{s}(\tilde{a})\,\mathrm{s}(\tilde{b})) \tag{16}$$

$$= 2(1 - \mathrm{E}(\mathrm{s}(\tilde{a})\,\mathrm{s}(\tilde{b}))) \tag{17}$$

$$= 2(1 - \rho_{\tilde{a},\tilde{b}}) \tag{18}$$

because $\mathrm{E}\,(\mathrm{s}(\tilde{a})^2) = \mathrm{E}((\mathrm{s}(\tilde{b})^2) = 1$. This directly implies $\rho_{\tilde{a},\tilde{b}} \leq 1$. Otherwise the right hand side would be negative, which is impossible because the left hand side is the expectation of a nonnegative quantity. By the same argument,

$$\mathrm{E}\left[(\mathrm{s}(\tilde{b}) - (-\mathrm{s}(\tilde{a})))^2\right] = 2(1 + \rho_{\tilde{a},\tilde{b}}). \tag{19}$$

implies $\rho_{\tilde{a},\tilde{b}} \geq -1$. When $\rho_{\tilde{a},\tilde{b}}$ equals 1, the left hand side of Eq. (18) equals zero. By Lemma 2.3, this means that $\mathrm{s}(\tilde{a}) = \mathrm{s}(\tilde{b})$ with probability one, which implies the linear relationship. The same argument applies to Eq. (19) when $\rho_{\tilde{a},\tilde{b}} = -1$. □

The correlation coefficient quantifies to what extent two quantities have a linear relationship. Consider the standardized variables $\mathrm{s}(\tilde{a})$ and $\mathrm{s}(\tilde{b})$ defined in Eq. (15). We can decompose $\mathrm{s}(\tilde{b})$ into two components: one collinear with $\mathrm{s}(\tilde{a})$, the other orthogonal to $\mathrm{s}(\tilde{a})$. This is illustrated in

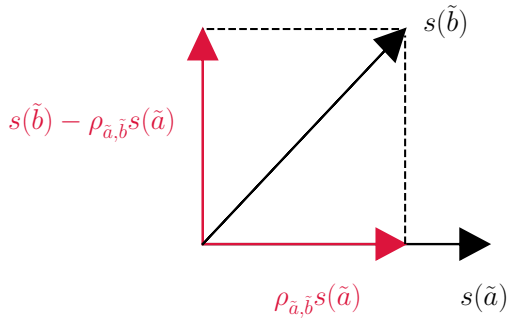

Figure 1: Geometric interpretation of the correlation coefficient in terms of the inner product between two unit-norm vectors representing two standardized random variables $\tilde{a}$ and $\tilde{b}$.

Figure 1. The covariance of the standardized variables equals the correlation coefficient, so the decomposition equals

$$\mathrm{s}(\tilde{b}) = \langle \mathrm{s}(\tilde{a}), \mathrm{s}(\tilde{b}) \rangle \, \mathrm{s}(\tilde{a}) + \left(\mathrm{s}(\tilde{b}) - \langle \mathrm{s}(\tilde{a}), \mathrm{s}(\tilde{b}) \rangle \, \mathrm{s}(\tilde{a}) \right) \tag{20}$$

$$= \rho_{\tilde{a},\tilde{b}} \, \mathrm{s}(\tilde{a}) + \left(\mathrm{s}(\tilde{b}) - \rho_{\tilde{a},\tilde{b}} \, \mathrm{s}(\tilde{a}) \right). \tag{21}$$

The second component is indeed orthogonal to $\mathrm{s}(\tilde{a})$,

$$\langle \mathrm{s}(\tilde{a}), \mathrm{s}(\tilde{b}) - \rho_{\tilde{a},\tilde{b}} \, \mathrm{s}(\tilde{a}) \rangle = \rho_{\tilde{a},\tilde{b}} - \rho_{\tilde{a},\tilde{b}} \tag{22}$$

$$= 0. \tag{23}$$

By Pythagoras' theorem

$$\| \mathrm{s}(\tilde{b}) - \rho_{\tilde{a},\tilde{b}} \, \mathrm{s}(\tilde{a}) \|^2 = \| \mathrm{s}(\tilde{b}) \|^2 - \left| \left| \rho_{\tilde{a},\tilde{b}} \, \mathrm{s}(\tilde{a}) \right| \right|^2 + \tag{24}$$

$$= 1 - \rho_{\tilde{a},\tilde{b}}^2. \tag{25}$$

The collinear component equals $\rho_{\tilde{a},\tilde{b}}$ and the magnitude of the orthogonal component equals $\sqrt{1 - \rho_{\tilde{a},\tilde{b}}^2}$. When $\rho_{\tilde{a},\tilde{b}} = \pm 1$ then there is no orthogonal component and the relationship is purely linear. Otherwise, the value $\rho_{\tilde{a},\tilde{b}}$ represents the relative importance of the collinear component. When it is positive, we say that the two random variables are positively correlated. When it is negative, we say they are negatively correlated (or anticorrelated). When it equals zero, the random variables are uncorrelated, which indicates that the variables are not linearly dependent. Note that two uncorrelated features may be highly dependent, just not linearly.

3 Sample mean, variance and correlation

When analyzing data we do not have access to a probability distribution, but rather to a set of points. In this section, we show how to adapt the analysis described in the previous section to this setting. In an nutshell, we approximate expectations by averaging over the data. Consider a dataset containing n real-valued data with two real valued features $(a_1, b_1), \ldots, (a_n, b_n)$. Let

$\mathcal{A} := \{a_1, \ldots, a_n\}$ and $\mathcal{B} := \{b_1, \ldots, b_n\}$. If we average instead of taking expectations, the resulting estimators for the mean, the covariance, and the variance are the sample mean

$$\text{av}\left(\mathcal{A}\right) := \frac{1}{n}\sum_{i=1}^{n} a_i, \tag{26}$$

the sample covariance

$$\text{cov}(\mathcal{A}, \mathcal{B}) := \frac{1}{n}\sum_{i=1}^{n} (a_i - \text{av}(\mathcal{A}))(b_i - \text{av}(\mathcal{B}), \tag{27}$$

and the sample variance,

$$\text{var}\left(\mathcal{A}\right) := \frac{1}{n}\sum_{i=1}^{n} \left(a_i - \text{av}\left(\mathcal{A}\right)\right)^2. \tag{28}$$

These estimates converge in mean square to the correct values if the data are independent samples from a distribution with a finite higher-order moments.

Theorem 3.1 (Sample mean converges to true mean)**.** *Let $\tilde{\mathcal{A}}_n$ contain n iid copies $\tilde{a}_1, \ldots, \tilde{a}_n$ of a random variable $\tilde{a}$ with finite variance. Then,*

$$\lim_n \text{E}\left(\left(\text{av}(\tilde{\mathcal{A}}_n) - \text{E}(\tilde{a})\right)^2\right) = 0. \tag{29}$$

Proof. By linearity of expection

$$\text{E}\left(\text{av}(\tilde{\mathcal{A}}_n)\right) = \frac{1}{n}\sum_{i=1}^{n} \text{E}(\tilde{a}_i) \tag{30}$$

$$= \text{E}(\tilde{a}), \tag{31}$$

which implies

$$\text{E}\left(\left(\text{av}(\tilde{\mathcal{A}}_n) - \text{E}(\tilde{a})\right)^2\right) = \text{Var}\left(\text{av}(\tilde{\mathcal{A}}_n)\right) \tag{32}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(\tilde{a}_i) \quad \text{by independence} \tag{33}$$

$$= \frac{\text{Var}(\tilde{a})}{n}. \tag{34}$$

□

The same proof can be applied to the sample variance and the sample covariance, under the assumption that higher-order moments of the distribution are bounded.

The sample mean, covariance and variance have geometric interpretations in their own right. The sample mean is the center of the dataset, if we use the square difference as a metric.

Lemma 3.2 (The sample mean is the center)**.** *For any set of real numbers* $\mathcal{A} := \{a_1, \ldots, a_n\}$*,*

$$\operatorname{av}(\mathcal{A}) = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} (c - a_i)^2. \tag{35}$$

Proof. Let $f(c) := \sum_{i=1}^{n} (c - a_i)^2$, we have

$$f'(c) = 2 \sum_{i=1}^{n} (c - a_i) \tag{36}$$

$$= 2 \left(nc - \sum_{i=1}^{n} a_i \right), \tag{37}$$

$$f''(c) = 2n. \tag{38}$$

The function is strictly convex and has a minimum where the derivative equals zero, i.e. when c is equal to the sample mean. □

Note that the proof is essentially the same as the one of Lemma 2.1. The reason is that the expectation and averaging operators are both linear. In fact, analogously to the probabilistic setting, we can show that the sample covariance is a valid inner product between centered sets of samples (it is a scaled dot product between the vectorized sets), and the sample standard deviation– defined as the square root of the sample variance– is its corresponding norm. We can therefore interpret the sample correlation coefficient

$$\rho_{\mathcal{A},\mathcal{B}} := \frac{\operatorname{cov}(\mathcal{A}, \mathcal{B})}{\sqrt{\operatorname{var}(\mathcal{A}) \operatorname{var}(\mathcal{B})}} \tag{39}$$

as a measure of collinearity between the samples of the two quantities. The same argument used to establish Theorem 2.4, shows that the coefficient is indeed restricted between -1 and 1.

Theorem 3.3 (Cauchy-Schwarz inequality)**.** *Let* $\mathcal{A} := \{a_1, \ldots, a_n\}$ *and* $\mathcal{B} := \{b_1, \ldots, b_n\}$ *be real-valued sets of features. The sample correlation coefficient satisfies*

$$-1 \leq \rho_{\mathcal{A},\mathcal{B}} \leq 1 \tag{40}$$

with equality if and only if b_i *is a linear function of* a_i *for all* $1 \leq i \leq n$*.*

Proof. The standardized data (also called z-scores) equal

$$\mathrm{s}(a)_i := \frac{a_i - \operatorname{av}(\mathcal{A})}{\sqrt{\operatorname{var}(\mathcal{A})}}, \qquad \mathrm{s}(b)_i := \frac{b_i - \operatorname{av}(\mathcal{B})}{\sqrt{\operatorname{var}(\mathcal{B})}}, \qquad 1 \leq i \leq n. \tag{41}$$

We have

$$\frac{1}{n} \sum_{i=1}^{n} (\mathrm{s}(b)_i - \mathrm{s}(a)_i)^2 = 2 \left(1 - \rho_{\mathcal{A},\mathcal{B}}\right), \tag{42}$$

$$\frac{1}{n} \sum_{i=1}^{n} (\mathrm{s}(b)_i + \mathrm{s}(a)_i)^2 = 2 \left(1 + \rho_{\mathcal{A},\mathcal{B}}\right), . \tag{43}$$

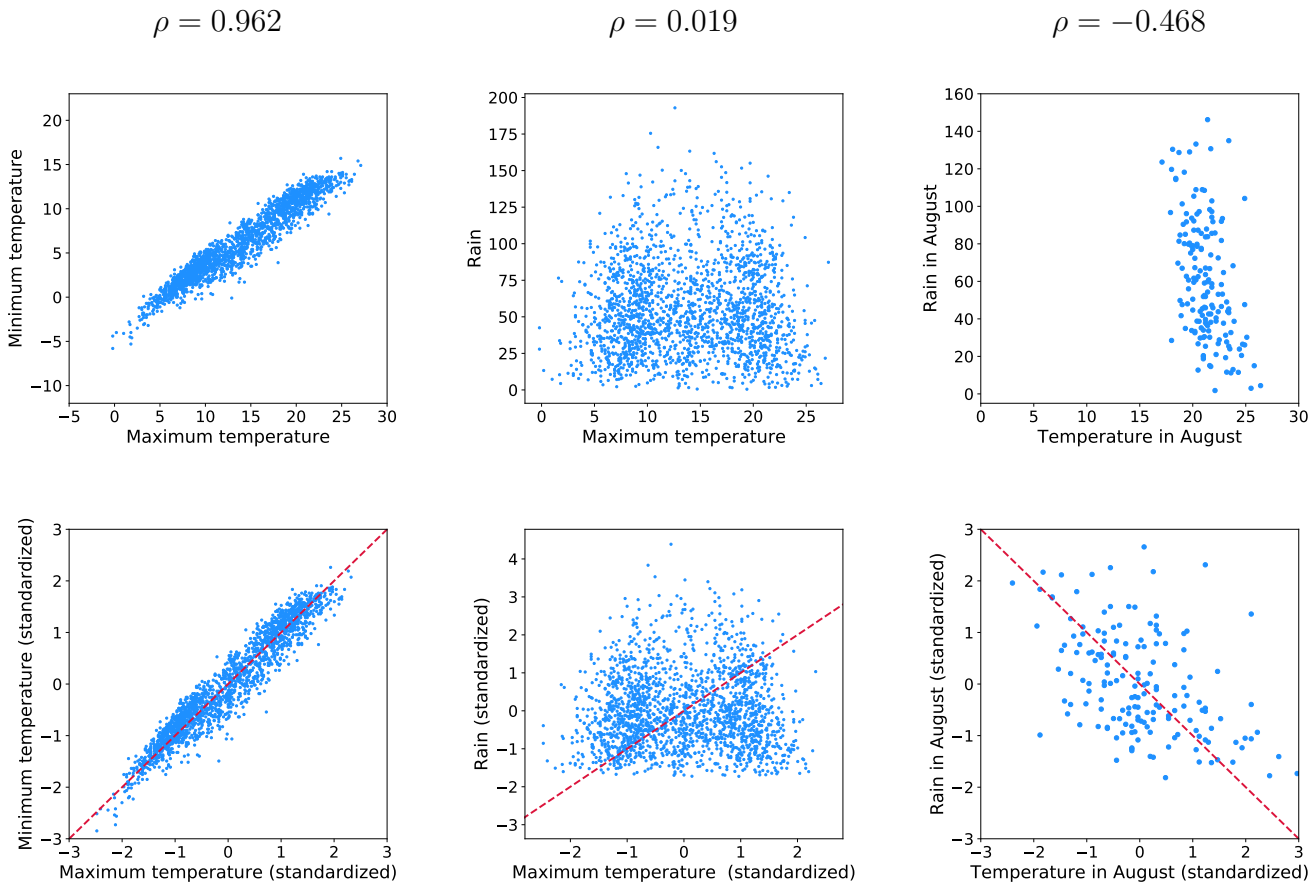

Figure 2: The top row shows scatterplots of the monthly maximum and minimum temperatures (left column), and the monthly rain and maximum temperature (middle column) in Oxford over 150 years. The bottom row shows the scatterplots of the same quantities after standardizing. If the relationship between each pair of features were perfectly linearly then they would lie on the dashed red diagonal lines.

Geometrically these quantities equal the average squared deviation of the standardized data from the lines with slopes +1 and -1. Eq. (42) directly implies $\rho_{\mathcal{A},\mathcal{B}} \leq 1$. Otherwise the right hand side is negative, which is impossible because the left hand side is clearly nonnegative. By the same argument Eq. (43) implies $\rho_{\mathcal{A},\mathcal{B}} \geq -1$. When $\rho_{\mathcal{A},\mathcal{B}}$ equals 1 or -1, the left hand side of Eq. (42) or Eq. (43) respectively is zero, which immediately implies the linear relationship. □

Example 3.4 (Oxford weather)**.** Figure 2 shows the sample correlation coefficient between different weather measurements gathered at a weather station in Oxford over 150 years.[1] Each data point corresponds to a different month. The maximum temperature is highly correlated with the minimum temperature ($\rho = 0.962$). Rainfall is almost uncorrelated with the maximum temperature ($\rho = 0.019$), but this does not mean that the two quantities are not related; the relation is just not linear. When we only consider the rain and temperature in August, then the two quantities

[1]The data are available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

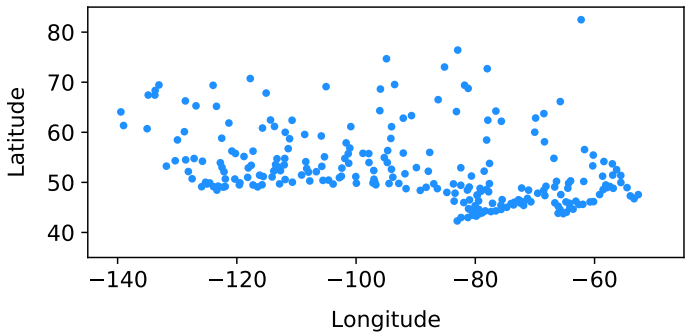

Figure 3: Scatterplot of the latitude and longitude of the main 248 cities in Canada.

are linearly related to some extent. Their correlation is negative ($\rho = -0.468$): when it is warmer it tends to rain less. △

4 The covariance matrix

We now consider the analysis of datasets containing multiple features. We represent the data as a set of n d-dimensional vectors $\mathcal{X} := \{x_1, \ldots, x_n\}$. This embeds the points in the Euclidean vector space $\mathbb{R}^d$. We begin from a probabilistic perspective, interpreting the data as samples from a d-dimensional random vector $\tilde{x}$. Our overall strategy is the same as in one dimension: first we find the center of the dataset, and then we characterize the variation around the center. To illustrate the different concepts we use the data shown in Figure 3. Each data point consists of the latitude and longitude of a city in Canada[2], so $d = 2$ in this case.

The mean of a random vector is the center of its distribution if we use the expected Euclidean distance as a metric.

Lemma 4.1. *For any d-dimensional random vector $\tilde{x}$ with finite mean,*

$$\mathrm{E}(\tilde{x}) := \arg\min_{w \in \mathbb{R}^d} \mathrm{E}\left(||\tilde{x} - w||_2^2\right). \tag{44}$$

Proof. The cost function decouples into d separate terms

$$\mathrm{E}\left(||\tilde{x} - w||_2^2\right) = \sum_{j=1}^{d} \mathrm{E}\left((\tilde{x}[j] - w[j])^2\right), \tag{45}$$

so the entry-wise mean achieves the minimum by Lemma 2.1. □

Similarly, the sample mean

$$\mathrm{av}(\mathcal{X}) := \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{46}$$

[2]The data are available at http://https://simplemaps.com/data/ca-cities

which equals the entry-wise sample mean of each feature, is the center of the dataset under the Euclidean norm, by essentially the same argument.

Lemma 4.2 (The sample mean is the center)**.** *For any set of n d-dimensional real-valued vectors $\mathcal{X} := \{x_1, \ldots, x_n\}$,*

$$\text{av}(\mathcal{X}) = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} ||x_i - w||_2^2 . \tag{47}$$

Proof. The result follows from Lemma 3.2 because the cost function decouples into

$$\sum_{i=1}^{n} ||x_i - w||_2^2 = \sum_{j=1}^{d} \sum_{i=1}^{n} (x_i[j] - w[j])^2. \tag{48}$$

□

From the probabilistic viewpoint, a reasonable measure of the variation of the random vector around the center of the distribution is its variance in different directions of space. Let v be an arbitrary unit-norm vector. The component of $\tilde{x}$ in the direction of v is given by $v^T\tilde{x}$. The variance of this random variable consequently quantifies the variance in that direction. By linearity of expectation,

$$\begin{aligned} \text{Var}\left(v^T\tilde{x}\right) &= \text{E}\left((v^T\tilde{x} - \text{E}(v^T\tilde{x}))^2\right) && (49)\\ &= \text{E}\left((v^T c(\tilde{x}))^2\right) && (50)\\ &= v^T \text{E}\left(c(\tilde{x})c(\tilde{x})^T\right) v, && (51) \end{aligned}$$

where $c(\tilde{x}) := \tilde{x} - \text{E}(\tilde{x})$ is the centered random vector. This motivates defining the covariance matrix of the random vector as follows.

Definition 4.3 (Covariance matrix)**.** *The covariance matrix of a d-dimensional random vector $\tilde{x}$ is the $d \times d$ matrix*

$$\begin{aligned} \Sigma_{\tilde{x}} &:= \text{E}\left(c(\tilde{x})c(\tilde{x})^T\right) && (52)\\ &= \begin{bmatrix} \text{Var}\left(\tilde{x}[1]\right) & \text{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \cdots & \text{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) \\ \text{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \text{Var}\left(\tilde{x}[2]\right) & \cdots & \text{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) & \text{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) & \cdots & \text{Var}\left(\tilde{x}[d]\right) \end{bmatrix}. && (53) \end{aligned}$$

The covariance matrix encodes the variance of the random vector in *every direction of space*.

Lemma 4.4. *For any random vector $\tilde{x}$ with covariance matrix $\Sigma_{\tilde{x}}$, and any vector v with unit ℓ_2-norm*

$$\text{Var}\left(v^T\tilde{x}\right) = v^T\Sigma_{\tilde{x}}v. \tag{54}$$

Proof. This follows immediately from Eq. (51). □

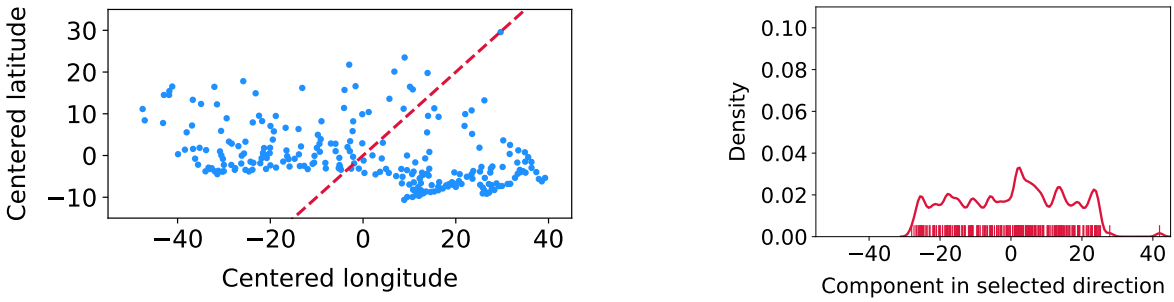

Figure 4: The left scatterplot shows the centered data from Figure 3, and a fixed direction of the two-dimensional space represented by a line going through the origin. The right plot shows the components of each data point in the direction of the line and their density. The variance of the components equals 229, so their standard deviation is 15.1.

For a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$, a natural estimator for the covariance matrix is the sample covariance matrix

$$\Sigma_{\mathcal{X}} := \frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T \tag{55}$$

$$= \begin{bmatrix} \text{var}\left(\mathcal{X}[1]\right) & \text{cov}\left(\mathcal{X}[1], \mathcal{X}[2]\right) & \cdots & \text{cov}\left(\mathcal{X}[1], \mathcal{X}[d]\right) \\ \text{cov}\left(\mathcal{X}[1], \mathcal{X}[2]\right) & \text{var}\left(\mathcal{X}[2]\right) & \cdots & \text{cov}\left(\mathcal{X}[2], \mathcal{X}[d]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}\left(\mathcal{X}[1], \mathcal{X}[d]\right) & \text{cov}\left(\mathcal{X}[2], \mathcal{X}[d]\right) & \cdots & \text{var}\left(\mathcal{X}[d]\right) \end{bmatrix}, \tag{56}$$

where $\mathcal{X}[j] := \{x_1[j], \ldots, x_n[j]\}$ for $1 \leq j \leq d$ and $c(x_i) := x_i - \text{av}\left(\mathcal{X}\right)$ for $1 \leq i \leq n$. The entries of the sample covariance matrix converge to the entries of the covariance matrix if the data are sampled from a random vector such that the higher moments of the entries and their products are bounded. However, beyond this probabilistic viewpoint, the sample covariance matrix has a meaningful geometric interpretation in its own right. Let v again be a unit-norm vector in a fixed direction of space. The component of each point in that direction equals $v^T x_i$. Consider the set of components $\mathcal{P}_v \mathcal{X} := \left\{v^T x_1, \ldots, v^T x_n\right\}$. We quantify the variation of the dataset around its sample mean using the sample variance of $\mathcal{P}_v \mathcal{X}$. Figure 4 illustrates this using the data in Figure 3. Analogously to the probabilistic setting, the sample covariance matrix encodes the sample variance in every direction.

Lemma 4.5. *For any dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ and any vector v with unit ℓ_2 norm*

$$\text{var}\left(\mathcal{P}_v \mathcal{X}\right) = v^T \Sigma_{\mathcal{X}} v. \tag{57}$$

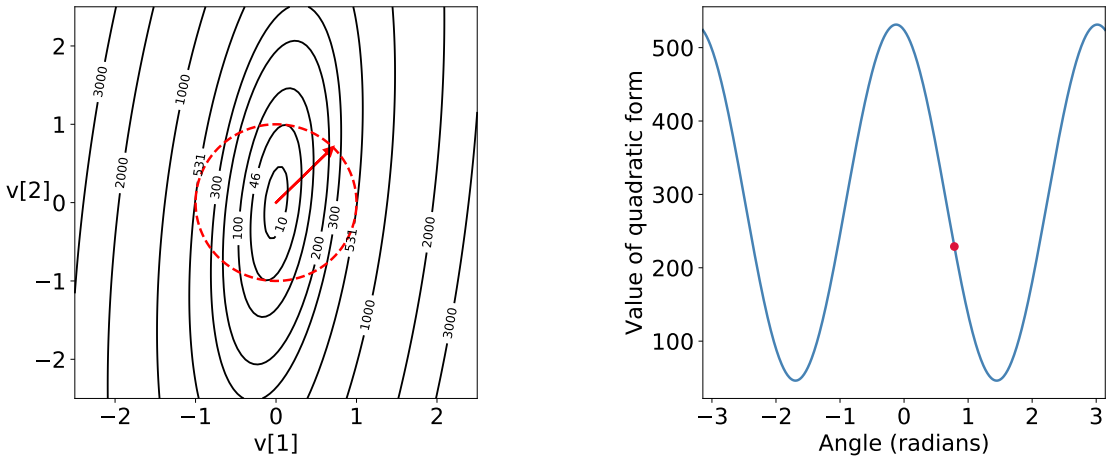

Figure 5: The left plot shows the contours of the quadratic form $v^T \Sigma_{\mathcal{X}} v$, where $\Sigma_{\mathcal{X}}$ is the sample covariance matrix of the data in Figure 3. The unit circle, where $||v||_2 = 1$, is drawn in red. The red arrow is a unit vector collinear with the dashed red line on the left plot of Figure 4. The right plot shows the value of the quadratic function when restricted to the unit circle. The red dot marks the value of the function corresponding to the unit vector represented by the red arrow on the left plot. This value is the sample variance of the data in that direction.

Proof.

$$\begin{aligned}
\text{var}\left(\mathcal{P}_v \, \mathcal{X}\right) &= \frac{1}{n} \sum_{i=1}^{n} \left(v^T x_i - \text{av}\left(\mathcal{P}_v \, \mathcal{X}\right)\right)^2 && (58)\\
&= \frac{1}{n} \sum_{i=1}^{n} \left(v^T \left(x_i - \text{av}\left(\mathcal{X}\right)\right)\right)^2 && (59)\\
&= v^T \left(\frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T\right) v \\
&= v^T \Sigma_{\mathcal{X}} v. && (60)
\end{aligned}$$

□

Figure 5 illustrates the result using the quadratic form corresponding to the sample covariance matrix of the data in Figure 3.

5 The spectral theorem

In this section, we study the properties of functions of the form

$$f(x) := x^T A x, \tag{61}$$

where A is a $d \times d$ symmetric matrix, and $f : \mathbb{R}^d \to \mathbb{R}$. Such functions are called quadratic forms, because they are multidimensional extensions of quadratic functions. In particular, we are interested in the value of $f(x)$ for vectors lying in the unit sphere, i.e. such that $||x||_2 = 1$. The motivation is that for covariance matrices and sample covariance matrices, which are symmetric by definition, the quadratic form on the sphere is equal to the variance in the direction of the vector x.

We begin by investigating whether the quadratic function reaches a maximum value on the unit sphere (in the case of covariance matrices, this would be the direction of maximum variance). The following lemma establishes that this indeed the case. The exact same argument can be used to establish that the function attains a minimum value on the unit sphere.

Lemma 5.1. *For any symmetric matrix $A \in \mathbb{R}^{d \times d}$, there exists a vector $u_1 \in \mathbb{R}^d$ such that*

$$u_1 = \arg \max_{||x||_2 = 1} x^T A x, \tag{62}$$

and a vector $u_d \in \mathbb{R}^d$ such that

$$u_d = \arg \min_{||x||_2 = 1} x^T A x. \tag{63}$$

Proof. We prove the existence of u_1, the existence of u_d follows by the same argument applied to $-A$. The unit sphere is closed and bounded, and the quadratic function is continuous (it is a second-order polynomial). The result follows from the extreme value theorem, which states that a continuous function on a closed and bounded set attains its extreme values. Proving the extreme value theorem is beyond the scope of these notes, but the idea is the following: If a set is closed and bounded, a continuous function maps it to a set of values (called its image) that is also closed and bounded. This means that the image contains all its limit points, and in particular cannot grow indefinitely towards a limit that it does not contain. □

Now we would like to characterize the direction that attains the maximum. The quadratic function is differentiable because it is a second-order polynomial. Consider the gradient of the quadratic function

$$\nabla f(x) = 2Ax. \tag{64}$$

Figure 6 shows the direction of the gradient on the unit circle for the quadratic form associated to the sample covariance matrix of the data in Figure 3. The projection of the gradient at a point x onto a unit vector v is equal to the directional derivative in that direction. If the derivative is positive, $v^T \nabla f(x) > 0$, then the function increases in that direction, i.e. for a small enough $\epsilon > 0$ $f(x + \epsilon v) > f(x)$. If u_1 is the point at which the maximum is attained, this *cannot happen* for directions that stay in our set of interest, which is the unit sphere. Here we hit a minor difficulty: since u_1 is on the sphere, $u_1 + \epsilon v$ is never on the sphere, because the sphere is a curved surface. However, $u_1 + \epsilon v$ is arbitrarily close to the sphere for small ϵ if v belongs to the *tangent plane* of the sphere at u_1.

The unit sphere is a level surface of the function $g(x) := x^T x$ (it contains every point x such that $g(x) = 1$). The tangent plane $\mathcal{T}$ of the level surface of a differentiable function g at a point x is

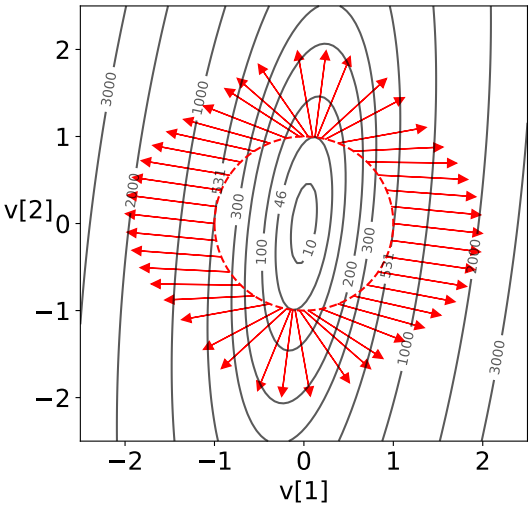

Figure 6: Each red arrow indicates the direction of the gradient of the quadratic form associated to the sample covariance matrix of the data in Figure 3 for a different point of the unit circle.

the set of vectors orthogonal to the gradient of g. A point y belongs to $\mathcal{T}$ if

$$\nabla g(x)^T(y-x) = 0. \tag{65}$$

For such points, if $y - x$ is small, then $g(y) \approx g(x) + \nabla g(x)^T(y-x) = g(x)$, so y is *almost* on the level surface.

If f attains its maximum at u_1 then there cannot be any points y in the tangent plane of the sphere at u_1 such that $\nabla f(u_1)^T(y-u_1) > 0$. If that is the case, then for small enough ϵ we can find a point y' on the sphere that is close enough to $u_1 + \epsilon(y-u_1)$, so that $f(y') \approx f(u_1 + \epsilon(y-u_1)) > f(u_1)$. To avoid this, $\nabla f(u_1)$ must be orthogonal to the tangent plane, and therefore collinear with the gradient of g. Figure 7 illustrates this: the tangent plane is drawn in purple, the direction of the gradient of g is drawn in green, and the direction of the gradient of f is represented by a red arrow. On the left, the gradients of f and g are not collinear, so we can find a direction in which the quadratic form increases on the unit circle. On the right, the gradients are collinear for all four points: these points correspond to the maxima and minima (the argument for the minima is exactly the same) of the quadratic form. Collinearity of the gradients implies that there exists a constant $\lambda_1 \in \mathbb{R}$ such that $\nabla f(u_1) = \lambda_1 \nabla g(u_1)$, so that $Au_1 = \lambda_1 u_1$. In words, the maximum is attained at an eigenvector of the matrix A. The following lemma makes this mathematically precise.

Lemma 5.2 (The maximum and minimum are attained at eigenvectors)**.** *For any symmetric matrix $A \in \mathbb{R}^{d\times d}$, a vector $u_1 \in \mathbb{R}^d$ such that*

$$u_1 = \arg\max_{||x||_2=1} x^T A x \tag{66}$$

is an eigenvector of A. There exists $\lambda_1 \in \mathbb{R}$ such that

$$Au_1 = \lambda_1 u_1, \tag{67}$$

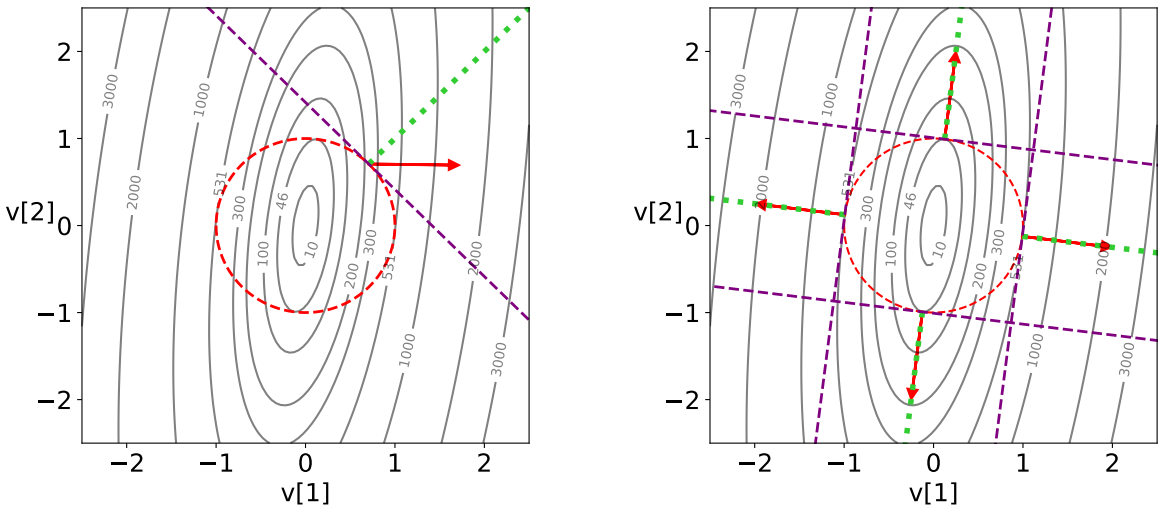

Figure 7: The tangent plane to the unit circle at several points is drawn in purple. The direction of the gradient of the function $g(x) := x^T x$ at those same points is drawn in green. The direction of the gradient of the quadratic form f associated to the sample covariance matrix of the data in Figure 3 is represented by a red arrow. On the left, the gradients of f and g are not collinear, so we can find a direction in which the quadratic form increases on the unit circle. On the right, the gradients are collinear for all four points: these points correspond to the maxima and minima of the quadratic form.

so that

$$\lambda_1 = \max_{||x||_2=1} x^T A x. \tag{68}$$

Similarly, a vector $u_d \in \mathbb{R}^d$ *such that*

$$u_d = \arg\min_{||x||_2=1} x^T A x \tag{69}$$

is an eigenvector of A*, and its corresponding eigenvalue* $\lambda_d \in \mathbb{R}$ *satisfies*

$$\lambda_d = \min_{||x||_2=1} x^T A x. \tag{70}$$

Proof. We prove the statement for the maximum. The statement for the minimum follows from the same argument applied to $-A$. We decompose Au_1 into two orthogonal components, one in the direction of u_1 and the other one in the tangent plane to the sphere,

$$Au_1 = u_1^T A u_1 u_1 + x_\perp, \tag{71}$$

$$x_\perp := Au_1 - u_1^T A u_1 u_1. \tag{72}$$

Our goal is to show that if u_1 attains the maximum then $x_\perp$ must be zero. Let

$$y := u_1 + \epsilon x_\perp \tag{73}$$

for some $\epsilon > 0$ to be chosen later. The vector $x_\perp$ is orthogonal to u_1 so by Pythagoras' theorem

$$||y||_2^2 = ||u_1||_2^2 + \epsilon^2 ||x_\perp||_2^2 \tag{74}$$

$$= 1 + \epsilon^2 ||x_\perp||_2^2 . \tag{75}$$

We now show that unless $x_\perp$ is zero, normalizing y yields a unit-norm vector that achieves a value larger than $f(u_1)$:

$$\left(\frac{y}{||y||_2}\right)^T A \frac{y}{||y||_2} = \frac{y^T A y}{||y||_2^2} \tag{76}$$

$$= \frac{u_1^T A u_1 + 2\epsilon x_\perp^T A u_1 + \epsilon^2 x_\perp^T A x_\perp}{1 + \epsilon^2 ||x_\perp||_2^2} \tag{77}$$

$$= \frac{u_1^T A u_1 + 2\epsilon ||x_\perp||_2^2 + \epsilon^2 x_\perp^T A x_\perp}{1 + \epsilon^2 ||x_\perp||_2^2} \tag{78}$$

$$= u_1^T A u_1 + \frac{2\epsilon ||x_\perp||_2^2 + \epsilon^2 (x_\perp^T A x_\perp - ||x_\perp||_2^2 u_1^T A u_1)}{1 + \epsilon^2 ||x_\perp||_2^2} \tag{79}$$

$$> u_1^T A u_1 \tag{80}$$

if we choose

$$\epsilon < \frac{2 ||x_\perp||_2^2}{|x_\perp^T A x_\perp| + ||x_\perp||_2^2 |u_1^T A u_1|}. \tag{81}$$

□

We can now analyze the landscape of the quadratic function f in directions orthogonal to the eigenvector u_1, by applying essentially the same argument to f restricted to the orthogonal complement of the span of u_1. The result is that the maximum of f on that subspace must be attained at another eigenvector of the symmetric matrix A. Repeating this argument d times yields the spectral theorem for symmetric matrices (the spectrum is the set of eigenvalues of a linear operator), which is a fundamental result in linear algebra: every symmetric $d \times d$ matrix A has d orthogonal eigenvectors.

Theorem 5.3 (Spectral theorem for symmetric matrices)**.** *If $A \in \mathbb{R}^{d \times d}$ is symmetric, then it has an eigendecomposition of the form*

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T , \tag{82}$$

where the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are real and the eigenvectors u_1, u_2, ..., u_n are real

and orthogonal. In addition,

$$\lambda_1 = \max_{||x||_2=1} x^T A x, \tag{83}$$

$$u_1 = \arg\max_{||x||_2=1} x^T A x, \tag{84}$$

$$\lambda_k = \max_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T A x, \quad 2 \leq k \leq d-1, \tag{85}$$

$$u_k = \arg\max_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T A x, \quad 2 \leq k \leq d-1, \tag{86}$$

$$\lambda_d = \min_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T A x, \tag{87}$$

$$u_d = \arg\min_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T A x. \tag{88}$$

Proof. The statements for u_d and λ_d follow directly from Lemmas 5.1 and 5.2. For the rest, apply a proof by induction on the dimension d of the matrix. If $d = 1$ the result is trivial: $A = a \in \mathbb{R}$, so we can set $u_1 := 1$ and $\lambda_1 := a$.

Assume that the theorem holds for $d-1$, and let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Lemmas 5.1 and 5.2 provide an eigenvector u_1 that achieves the maximum, equal to λ_1. Consider the matrix

$$A - \lambda_1 u_1 u_1^T. \tag{89}$$

Its column space is orthogonal to u_1,

$$(A - \lambda_1 u_1 u_1^T) u_1 = A u_1 - \lambda_1 u_1 \tag{90}$$

$$= 0, \tag{91}$$

where 0 denotes the n-dimensional zero vector. Its row space is also orthogonal by the same argument. Both subspaces are therefore contained in $\text{span}(u_1)^\perp$, the orthogonal complement of the span of u_1, which is a subspace of dimension $d-1$. Let $V_\perp$ be a $d \times d-1$ orthogonal matrix whose columns are an orthonormal basis of $\text{span}(u_1)^\perp$. $V_\perp V_\perp^T$ is a projection matrix that projects onto $\text{span}(u_1)^\perp$, so that

$$A - \lambda_1 u_1 u_1^T = V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T. \tag{92}$$

We define $B := V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp$, which is a $d-1 \times d-1$ symmetric matrix. By the induction hypothesis there exist $\gamma_1, \ldots, \gamma_{d-1}$ and $w_1, \ldots, w_{d-1}$ such that

$$\gamma_1 = \max_{||y||_2=1} y^T B y, \tag{93}$$

$$w_1 = \arg\max_{||y||_2=1} y^T B y, \tag{94}$$

$$\gamma_k = \max_{||y||_2=1, y\perp w_1,\ldots,w_{k-1}} y^T B y, \quad 2 \leq k \leq d-2, \tag{95}$$

$$w_k = \arg\max_{||y||_2=1, y\perp w_1,\ldots,w_{k-1}} y^T B y, \quad 2 \leq k \leq d-2. \tag{96}$$

For any $x \in \text{span}(u_1)^\perp$, we have

$$x^T (A - \lambda_1 u_1 u_1^T) x = x^T A x \tag{97}$$

and $x = V_\perp y$ for a vector $y \in \mathbb{R}^{d-1}$ such that $||x||_2^2 = y^T V_\perp^T V_\perp y = ||y||_2^2$. This implies

$$\max_{||x||_2=1, x \perp u_1} x^T A x = \max_{||x||_2=1, x \perp u_1} x^T V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T x \tag{98}$$

$$= \max_{||y||_2=1} y^T B y \tag{99}$$

$$= \gamma_1. \tag{100}$$

Inspired by this, we set $u_k := V_\perp w_{k-1}$ for $k = 2, \ldots, d$. Each u_k is an eigenvector of A with eigenvalue $\lambda_k := \gamma_{k-1}$:

$$A u_k = V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1} \tag{101}$$

$$= V_\perp B w_{k-1} \tag{102}$$

$$= \gamma_{k-1} V_\perp w_{k-1} \tag{103}$$

$$= \lambda_k u_k. \tag{104}$$

Note that these vectors are all orthogonal and unit norm. Also, we can express any $x \in \text{span}(u_1)^\perp$ orthogonal to $u_{k'}$, where $2 \leq k' \leq d$, as $x = V_\perp y$, $y \in \mathbb{R}^{d-1}$, where $y \perp w_{k'-1}$, because $u_{k'}^T x = w_{k'-1}^T V_\perp^T V_\perp y = w_{k'-1}^T y$. This implies

$$\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x = \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T x \tag{105}$$

$$= \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T x \tag{106}$$

$$= \max_{||y||_2=1, y \perp w_1, \ldots, w_{k-2}} y^T B y \tag{107}$$

$$= \gamma_{k-1} \tag{108}$$

$$= \lambda_k. \tag{109}$$

□

6 Principal component analysis

We now reap the rewards for our hard work in the previous section. By the spectral theorem (Theorem 5.3) combined with Lemma 4.4, in order to characterize the variance of a random vector in different directions of space, we just need to perform an eigendecomposition of its covariance matrix. The first eigenvector u_1 is the direction of highest variance, which is equal to the corresponding eigenvalue. In directions orthogonal to u_1 the maximum variance is attained by the second eigenvector u_2, and equals the corresponding eigenvalue λ_2. In general, when restricted to the orthogonal complement of the span of $u_1, \ldots, u_k$ for $1 \leq k \leq d-1$, the variance is highest in the direction of the $k+1$th eigenvector u_{k+1}.

Theorem 6.1. *Let $\tilde{x}$ be a random vector d-dimensional with covariance matrix $\Sigma_{\tilde{x}}$, and let u_1,*

..., u_d, and $\lambda_1 > \ldots > \lambda_d$ denote the eigenvectors and corresponding eigenvalues of $\Sigma_{\tilde{x}}$. We have

$$\lambda_1 = \max_{||v||_2=1} \text{Var}(v^T \tilde{x}), \tag{110}$$

$$u_1 = \arg\max_{||v||_2=1} \text{Var}(v^T \tilde{x}), \tag{111}$$

$$\lambda_k = \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d, \tag{112}$$

$$u_k = \arg\max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d. \tag{113}$$

We call the directions of the eigenvectors *principal directions*. The component of the centered random vector $c(\tilde{x}) := \tilde{x} - \text{E}(\tilde{x})$ in each principal direction is called a *principal component*,

$$\widetilde{pc}[i] := u_i^T c(\tilde{x}), \quad 1 \leq i \leq d \tag{114}$$

By Theorem 6.1 the variance of each principal component is the corresponding eigenvalue of the covariance matrix. Geometrically, we are rotating the random vector to make the axes align with the principal directions. The principal components are uncorrelated,

$$\begin{aligned} \text{E}(\widetilde{pc}[i]\widetilde{pc}[j]) &= \text{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) && (115)\\ &= u_i^T \text{E}(c(\tilde{x}) c(\tilde{x})^T) u_j && (116)\\ &= u_i^T \Sigma_{\tilde{x}} u_j && (117)\\ &= \lambda_j u_i^T u_j && (118)\\ &= 0, && (119) \end{aligned}$$

so there is no linear relationship between them.

In practice, the principal directions and principal components are computed by performing an eigendecomposition of the sample covariance matrix of the data.

Algorithm 6.2 (Principal component analysis (PCA))**.** *Given a dataset $\mathcal{X}$ containing n vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with d features each, where $n > d$.*

1. *Compute the sample covariance matrix of the data $\Sigma_{\mathcal{X}}$.*
2. *Compute the eigendecomposition of $\Sigma_{\mathcal{X}}$, to find the principal directions u_1, ..., u_d.*
3. *Center the data and compute the principal components*

$$pc_i[j] := u_j^T c(x_i), \quad 1 \leq i \leq n, \; 1 \leq j \leq d, \tag{120}$$

where $c(x_i) := x_i - \text{av}(\mathcal{X})$

As in the case of the sample mean, variance and covariance, when we perform PCA on a dataset, the result has a geometric interpretation which does *not* require the existence of any underlying distribution. This again follows from the spectral theorem (Theorem 5.3), in this case combined with Lemma 4.5.

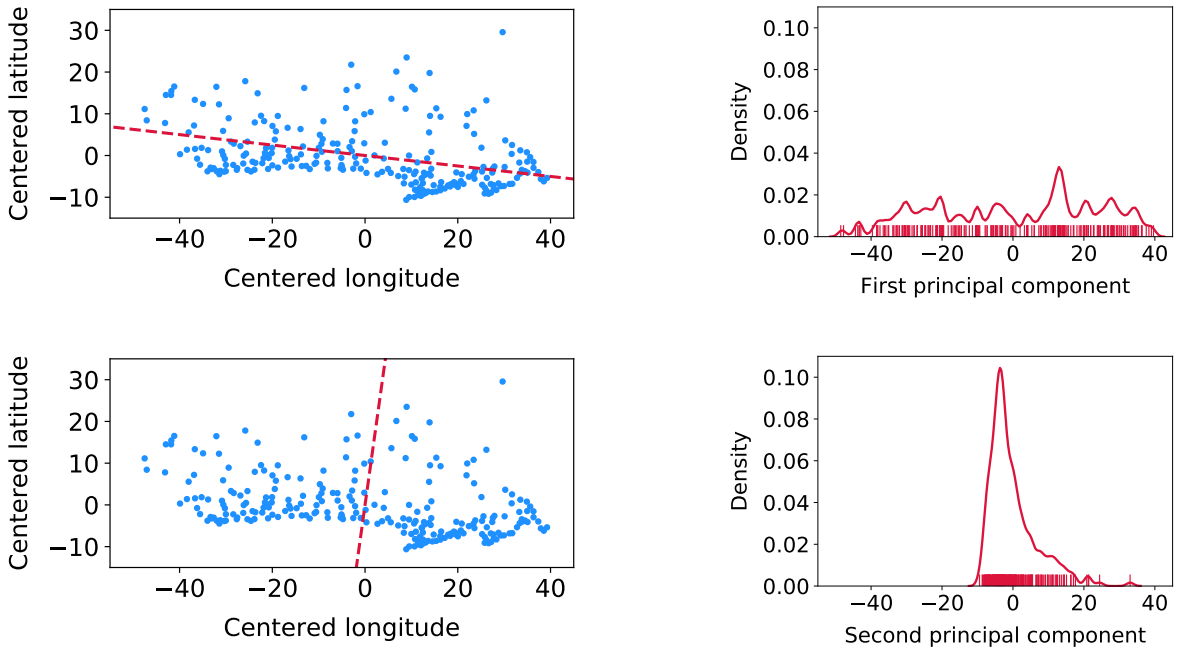

Figure 8: The scatterplots in the left column show the centered data from Figure 3, and the first (top) and second (bottom) principal directions of the data represented by lines going through the origin. The right column shows the first (top) and second (bottom) principal components of each data point and their density. The sample variance of the first component equals 531 (standard deviation: 23.1). For the second it equals 46.2 (standard deviation: 6.80)

Theorem 6.3. *Let $\mathcal{X}$ contain n vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with sample covariance matrix $\Sigma_{\mathcal{X}}$, and let u_1, ..., u_d, and $\lambda_1 > \ldots > \lambda_d$ denote the eigenvectors and corresponding eigenvalues of $\Sigma_{\mathcal{X}}$. We have*

$$\lambda_1 = \max_{||v||_2=1} \text{var}\left(\mathcal{P}_v \mathcal{X}\right), \tag{121}$$

$$u_1 = \arg\max_{||v||_2=1} \text{var}\left(\mathcal{P}_v \mathcal{X}\right), \tag{122}$$

$$\lambda_k = \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{var}\left(\mathcal{P}_v \mathcal{X}\right), \quad 2 \leq k \leq d, \tag{123}$$

$$u_k = \arg\max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{var}\left(\mathcal{P}_v \mathcal{X}\right), \quad 2 \leq k \leq d. \tag{124}$$

In words, u_1 is the direction of maximum sample variance, u_2 is the direction of maximum sample variance orthogonal to u_1, and in general u_k is the direction of maximum variation that is orthogonal to u_1, u_2, ..., u_{k-1}. The sample variances in each of these directions are given by the eigenvalues. Figure 8 shows the principal directions and the principal components for the data in Figure 3. Comparing the principal components to the component in the direction shown in Figure 4, we confirm that the first principal component has larger sample variance, and the second principal component has smaller sample variance.

Example 6.4 (PCA of faces)**.** The Olivetti Faces dataset[3] contains 400 64×64 images taken from

[3]Available at http://www.cs.nyu.edu/~roweis/data.html

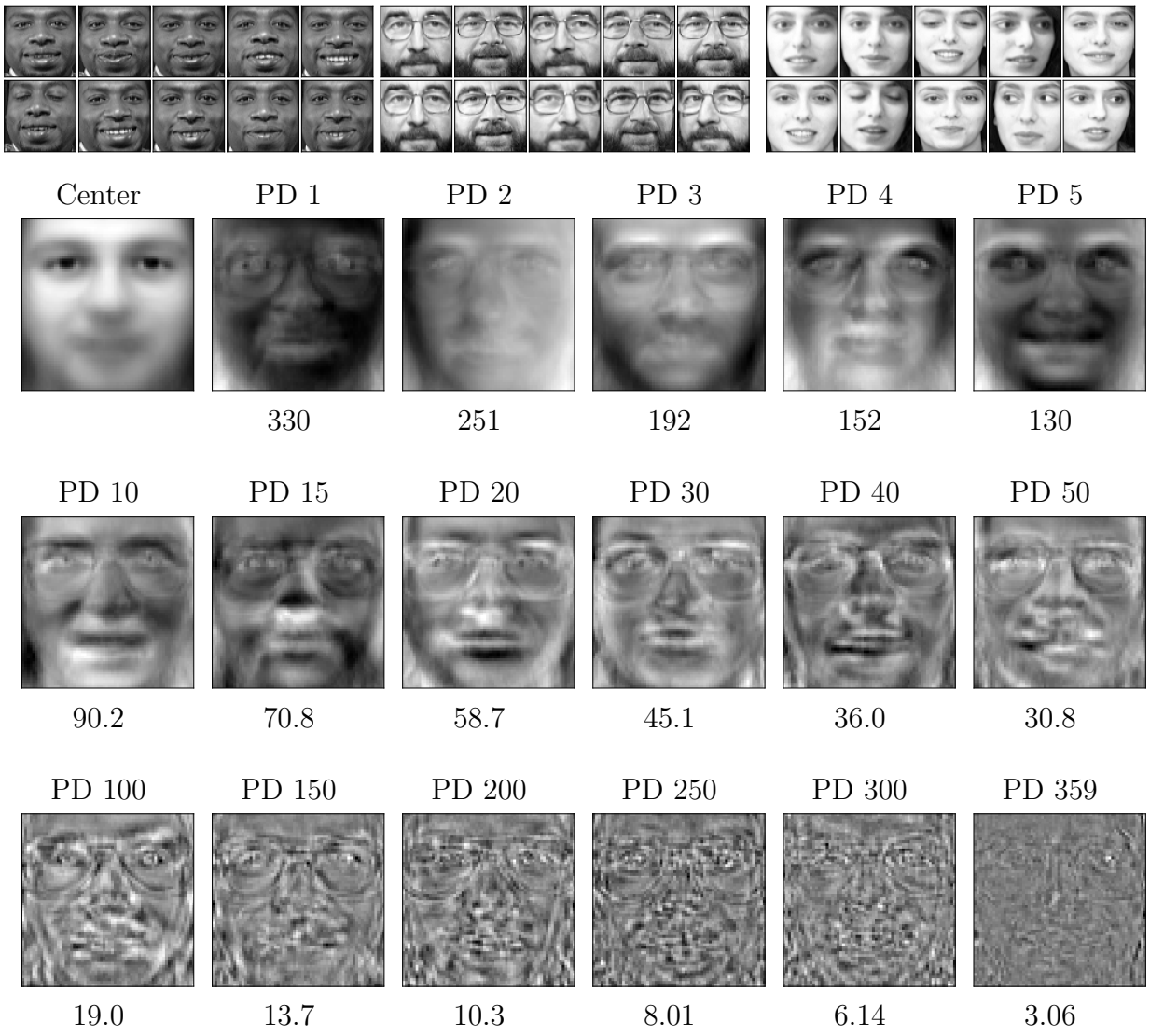

Figure 9: The top row shows the data corresponding to three different individuals in the Olivetti dataset. The sample mean and the principal directions (PD) obtained by applying PCA to the centered data are depicted below. The sample standard deviation of each principal component is listed below the corresponding principal direction.

40 different subjects (10 per subject). We vectorize each image so that each pixel is interpreted as a different feature. Figure 9 shows the center of the data and several principal directions, together with the standard deviations of the corresponding principal components. The first principal components seem to capture low-resolution structure, which account for more sample variance, whereas the last incorporate more intricate details. △

7 Dimensionality reduction via PCA

Data containing a large number of features can be difficult to analyze and process. The goal of dimensionality-reduction techniques is to embed the data points in a low-dimensional space where they can be described with a small number of variables. This is a crucial preprocessing step in many applications. A popular choice is to perform *linear* dimensionality reduction, where the lower-dimensional representation is obtained by computing the inner products of each data point with a small number of basis vectors. Let $\mathcal{X} := \{x_1, \ldots, x_n\}$ be a dataset containing n d-dimensional vectors, and let $v_1, \ldots, v_k$ be a set of $k < d$ orthonormal vectors. The k-dimensional representation of the ith data point consists of $v_1^T x_i, \ldots, v_k^T x_i$. This preserves the component of the data point contained in the k-dimensional subspace spanned by $v_1, \ldots, v_k$. An important question is how to choose these vectors, or equivalently how to choose the low-dimensional subspace that will be preserved. This is a difficult question, as it depends on what we want to do with the data. However, PCA provides a compelling option: it uncovers subspaces that are optimal in the sense that they capture the maximum possible sample variance in the data. This is a direct consequence of the following theorem.

Theorem 7.1. *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix, and $u_1, \ldots, u_k$ be its first k eigenvectors. For any set of k orthonormal vectors $v_1, \ldots, v_k$*

$$\sum_{i=1}^{k} u_i^T A u_i \geq \sum_{i=1}^{k} v_i^T A v_i. \tag{125}$$

Proof. We prove the result by induction on k. The base case $k = 1$ follows immediately from (84). To complete the proof we show that if the result is true for $k - 1 \geq 1$ (the induction hypothesis) then it also holds for k. Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$. For any orthonormal basis $b_1, \ldots, b_k$ of $\mathcal{S}$ the projection matrix VV^T, whose columns are $v_1, \ldots, v_k$, is equal to the projection matrix BB^T, whose columns are $b_1, \ldots, b_k$. This implies

$$\begin{aligned} \textstyle\sum_{i=1}^{k} v_i^T A v_i &= \text{trace}\left(V^T A V\right) && (126)\\ &= \text{trace}\left(A V V^T\right) && (127)\\ &= \text{trace}\left(A B B^T\right) && (128)\\ &= \textstyle\sum_{i=1}^{k} b_i^T A b_i, && (129) \end{aligned}$$

so we are free to choose an arbitrary basis for $\mathcal{S}$. Because it has dimension k, $\mathcal{S}$ contains at least one vector b that is orthogonal to $u_1, u_2, \ldots, u_{k-1}$. By (86),

$$u_k^T A u_k \geq b^T A b. \tag{130}$$

We now build an orthonormal basis $b_1, b_2, \ldots, b_k$ for $\mathcal{S}$ such that $b_k := b$ (we can construct such a basis by Gram-Schmidt, starting with b). By the induction hypothesis,

$$\sum_{i=1}^{k-1} u_i^T A u_i \geq \sum_{i=1}^{k-1} b_i^T A b_i. \tag{131}$$

Combining (131) and (130) yields the desired result. □

Corollary 7.2. *For any dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ of dimension d, any dimension $k < d$, and any set of k orthonormal vectors $v_1, \ldots, v_k$*

$$\sum_{i=1}^{k} \text{var}(\text{pc}[i]) \geq \sum_{i=1}^{k} \text{var}(\mathcal{P}_{v_i} \mathcal{X}), \tag{132}$$

where $\text{pc}[i] := \{\text{pc}_1[i], \ldots, \text{pc}_n[i]\} = \mathcal{P}_{u_i} \mathcal{X}$ denotes the ith principal component of the data.

Proof. Let $u_1, \ldots, u_k$ be the first k eigenvectors of $\Sigma_{\mathcal{X}}$

$$\sum_{i=1}^{k} \text{var}(\text{pc}[i]) = \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{n} u_i^T c(x_j) c(x_j)^T u_i \tag{133}$$

$$= \sum_{i=1}^{k} u_i^T \Sigma_{\mathcal{X}} u_i \tag{134}$$

$$\geq \sum_{i=1}^{k} v_i^T \Sigma_{\mathcal{X}} v_i \tag{135}$$

$$= \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{n} v_i^T c(x_j) c(x_j)^T v_i. \tag{136}$$

□

Example 7.3 (PCA of faces (continued))**.** Figure 10 shows the result of representing one of the faces in the dataset from Example 6.4 using its first 7 principal components. To visualize the result, we project the representation onto the image space using the principal directions,

$$x_i^{\text{reduced}} := \text{av}(\mathcal{X}) + \sum_{j=1}^{7} \text{pc}_i[j] u_j, \tag{137}$$

where x_i is the chosen face, and $\mathcal{X}$ is the set of faces. Figure 11 shows representations of increasing dimensionality of the same face. As suggested by the visualization of the principal directions in Figure 9, the lower-dimensional projections produce blurry images. △

Example 7.4 (Nearest neighbors in principal-component space)**.** To illustrate a possible use of PCA-based dimensionality reduction, we consider the problem of face classification.The nearest-neighbor algorithm is a classical method to perform classification. Assume that we have access to a training set of n pairs of data encoded as vectors in $\mathbb{R}^d$ along with their corresponding labels:

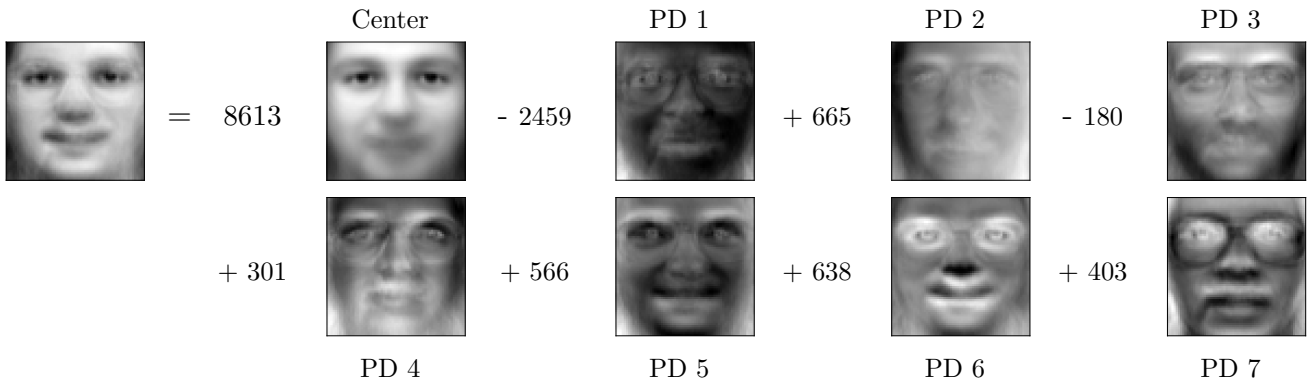

Figure 10: Representation of one of the faces in the dataset from Example 6.4 using the first 7 principal components. We visualize the representation by projecting onto the image space using the corresponding principal directions.

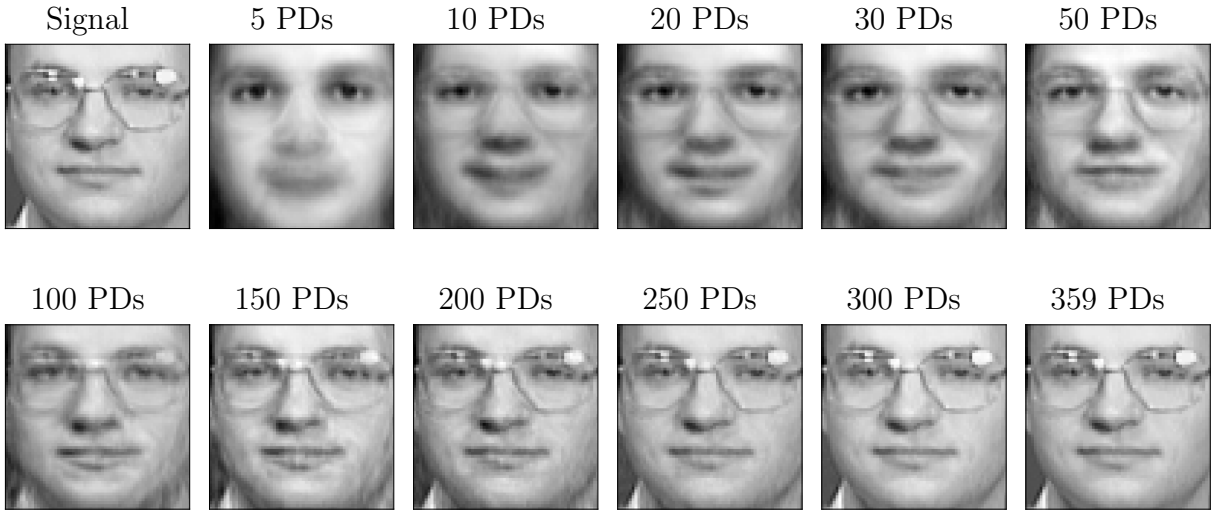

Figure 11: Representation of one of the faces in the dataset from Example 6.4 using different numbers of principal components. We visualize the representation by projecting onto the image space using the corresponding principal directions.

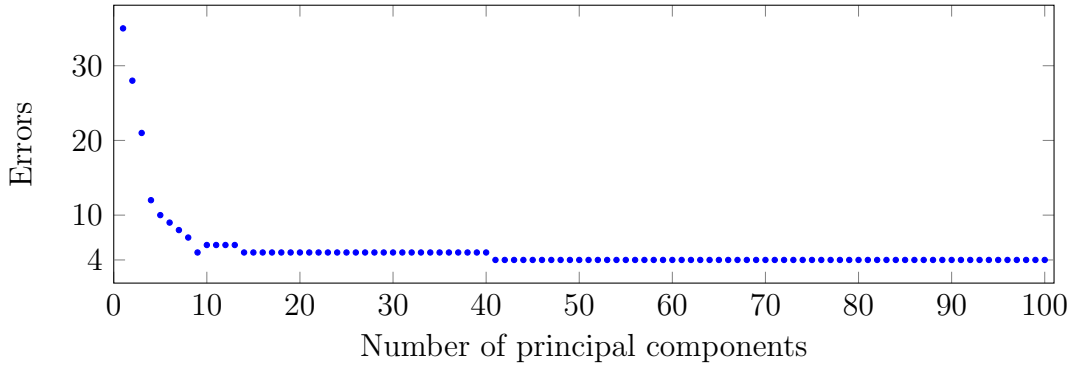

Figure 12: Errors for nearest-neighbor classification combined with PCA-based dimensionality reduction for different dimensions.

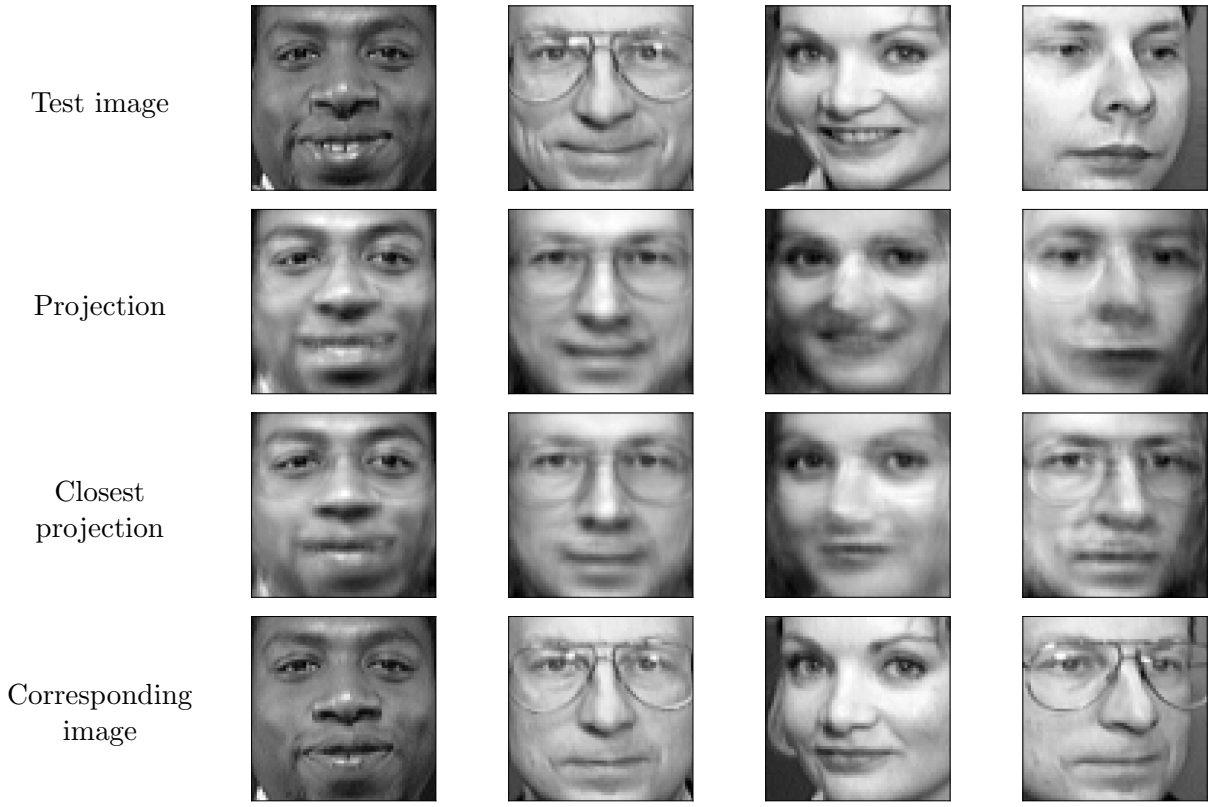

Figure 13: Results of nearest-neighbor classification combined with PCA-based dimensionality reduction of order 41 for four of the people in Example 7.4. The assignments of the first three examples are correct, but the fourth is wrong.

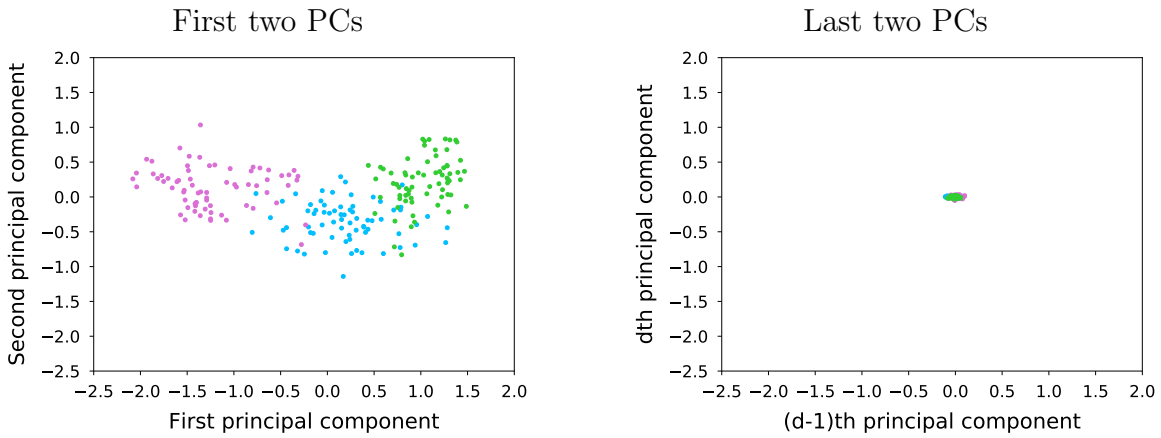

Figure 14: Projection of 7-dimensional vectors describing different wheat seeds onto the first two (left) and the last two (right) principal dimensions of the dataset. Each color represents a variety of wheat.

$\{x_1, l_1\}, \ldots, \{x_n, l_n\}$. To classify a new data point y we find the closest element of the training set,

$$i^* := \arg \min_{1 \leq i \leq n} ||y - x_i||_2 \,, \tag{138}$$

and assign the corresponding label l_{i^*} to y. Every time we classify a new point, we need to compute n distances in a d-dimensional space. The computational cost is $\mathcal{O}(nd)$. To alleviate the cost, we can perform PCA and apply the algorithm in a space of reduced dimensionality k, so that the cost is now $\mathcal{O}(nk)$. Applying PCA to the training data is costly, but only needs to be done once.

In this example we explore this idea using the faces dataset from Example 6.4. The training set consists of 360 64×64 images taken from 40 different subjects (9 per subject). The test set consists of an image of each subject, which is different from the ones in the training set. We apply the nearest-neighbor algorithm to classify the faces in the test set, modeling each image as a vector in $\mathbb{R}^{4096}$ and using the ℓ_2-norm distance. The algorithm classifies 36 of the 40 subjects correctly.

Figure 12 shows the accuracy of the algorithm when we compute the distance using k principal components, obtained by applying PCA to the training set, for different values of k. The accuracy increases with the dimension at which the algorithm operates. Interestingly, this is not necessarily always the case because projections may actually be helpful for tasks such as classification (for example, factoring out small shifts and deformations). The same precision as in the ambient dimension (4 errors out of 40 test images) is achieved using just $k = 41$ principal components (in this example $n = 360$ and $d = 4096$). Figure 13 shows some examples of the projected data represented in the original d-dimensional space along with their nearest neighbors in the k-dimensional space. △

Example 7.5 (Dimensionality reduction for visualization)**.** Dimensionality reduction is very useful for visualization. When visualizing data the objective is usually to project it down to 2D or 3D in a way that preserves its structure as much as possible. In this example, we consider a dataset where each data point corresponds to a seed with seven features: area, perimeter, compactness,

length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian.[4]

Figure 14 shows the data represented by the first two and the last two principal components. In the latter case, there is almost no discernible variation. As predicted by our theoretical analysis of PCA, the structure in the data is much better conserved by the two first principal components, which allow to clearly visualize the difference between the three types of seeds. Note that using the first few principal components only ensures that we preserve as much variation as possible; this does not necessarily mean that these are the best low-dimensional features for tasks such as clustering or classification. △

8 Gaussian random vectors

The Gaussian or normal random variable is arguably the most popular random variable in statistical modeling and signal processing. The reason is that sums of independent random variables often converge to Gaussian distributions, a phenomenon characterized by the central limit theorem. As a result any quantity that results from the additive combination of several unrelated factors will tend to have a Gaussian distribution. For example, in signal processing and engineering, noise is often modeled as Gaussian.

Definition 8.1 (Gaussian)**.** *The pdf of a Gaussian or normal random variable $\tilde{a}$ with mean μ and standard deviation σ is given by*

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}. \tag{139}$$

One can verify that the parameters μ and σ correspond to the mean and standard deviation of the distribution, i.e.

$$\mu = \int_{a=-\infty}^{\infty} a f_{\tilde{a}}(a)\, \mathrm{d}a, \tag{140}$$

$$\sigma^2 = \int_{a=-\infty}^{\infty} (a-\mu)^2 f_{\tilde{a}}(a)\, \mathrm{d}a. \tag{141}$$

An important property of Gaussian random variables is that scaling and shifting Gaussians preserves their distribution.

Lemma 8.2. *If $\tilde{a}$ is a Gaussian random variable with mean μ and standard deviation σ, then for any $\alpha, \beta \in \mathbb{R}$*

$$\tilde{b} := \alpha\tilde{a} + \beta \tag{142}$$

is a Gaussian random variable with mean $\alpha\mu + \beta$ and standard deviation $|\alpha|\,\sigma$.

[4]The data can be found at https://archive.ics.uci.edu/ml/datasets/seeds.

Proof. We assume $\alpha > 0$ (the argument for $\alpha < 0$ is very similar), to obtain

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{b} \leq b\right) \tag{143}$$

$$= \mathrm{P}\left(\alpha\tilde{a} + \beta \leq b\right) \tag{144}$$

$$= \mathrm{P}\left(\tilde{a} \leq \frac{b-\beta}{\alpha}\right) \tag{145}$$

$$= \int_{-\infty}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} \, \mathrm{d}a \tag{146}$$

$$= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(w-\alpha\mu-\beta)^2}{2\alpha^2\sigma^2}} \, \mathrm{d}w \qquad \text{by the change of variables } w = \alpha a + \beta. \tag{147}$$

Differentiating with respect to b yields

$$f_{\tilde{b}}(b) = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(b-\alpha\mu-\beta)^2}{2\alpha^2\sigma^2}} \tag{148}$$

so $\tilde{b}$ is indeed a standard Gaussian random variable with mean $\alpha\mu + \beta$ and standard deviation $|\alpha|\,\sigma$. □

Gaussian random vectors are a multidimensional generalization of Gaussian random variables. They are parametrized by a vector and a matrix that are equal to their mean and covariance matrix (this can be verified by computing the corresponding integrals).

Definition 8.3 (Gaussian random vector)**.** *A Gaussian random vector $\tilde{x}$ of dimension d is a random vector with joint pdf*

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right), \tag{149}$$

where $|\Sigma|$ denotes the determinant of Σ. The mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, which is symmetric and positive definite (all eigenvalues are positive), parametrize the distribution.

In order to better understand the geometry of the pdf of Gaussian random vectors, we analyze its contour surfaces. The contour surfaces are sets of points where the density is constant. The spectral theorem (Theorem 5.3) ensures that $\Sigma = U\Lambda U^T$, where U is an orthogonal matrix and Λ is diagonal, and therefore $\Sigma^{-1} = U\Lambda^{-1}U^T$. Let c be a fixed constant. We can express the contour surfaces as

$$c = x^T\Sigma^{-1}x \tag{150}$$

$$= x^T U\Lambda^{-1}Ux \tag{151}$$

$$= \sum_{i=1}^{d} \frac{(u_i^T x)^2}{\lambda_i}. \tag{152}$$

The equation corresponds to an ellipsoid with axes aligned with the directions of the eigenvectors. The length of the ith axis is proportional to $\sqrt{\lambda_i}$. We have assumed that the distribution is centered around the origin (μ is zero). If μ is nonzero then the ellipsoid is centered around μ.

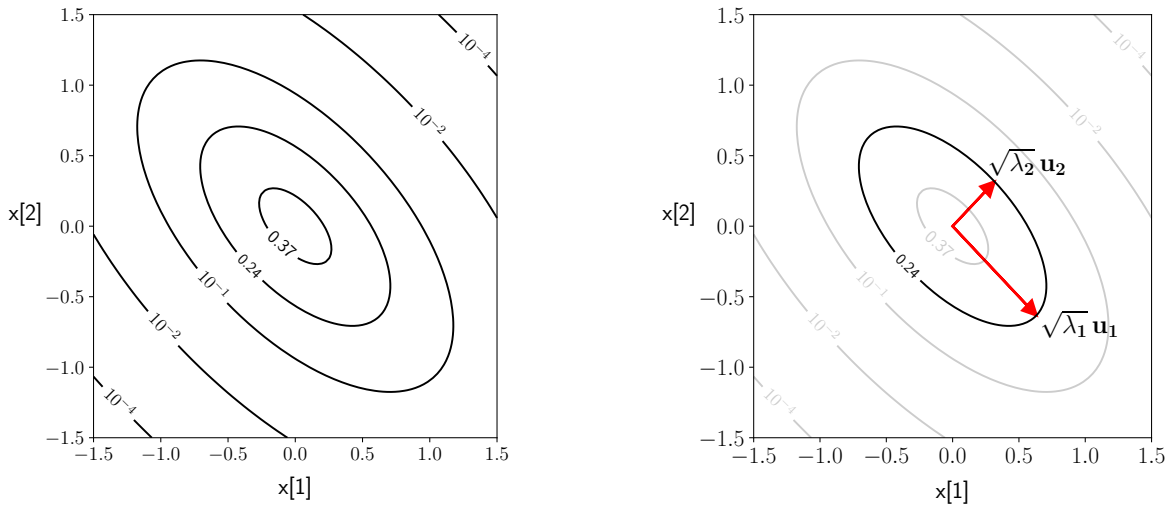

Figure 15: The left image shows a contour plot of the probability density function of the two-dimensional Gaussian random vector defined in Example 8.4. The axes align with the eigenvectors of the covariance matrix, and are proportional to the square root of the eigenvalues, as shown on the right image for a specific contour.

Example 8.4 (Two-dimensional Gaussian)**.** We illustrate the geometry of the Gaussian probability distribution function with a two-dimensional example where μ is zero and

$$\Sigma = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}. \tag{153}$$

The eigendecomposition of Σ yields $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, and

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}. \tag{154}$$

The left plot of Figure 15 shows several contours of the density. The right plot shows the axes for the contour line

$$\frac{(u_1^T x)^2}{\lambda_1} + \frac{(u_2^T x)^2}{\lambda_2} = 1, \tag{155}$$

where the density equals 0.24. △

When the entries of a Gaussian random vector are uncorrelated, then they are also independent. The relationship between the entries is purely linear. This is *not* the case for most other random distributions,

Lemma 8.5 (Uncorrelation implies mutual independence for Gaussian random variables)**.** *If all the components of a Gaussian random vector $\tilde{x}$ are uncorrelated, then they are also mutually independent.*

Proof. If all the components are uncorrelated then the covariance matrix is diagonal

$$\Sigma_{\tilde{x}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}, \tag{156}$$

where σ_i is the standard deviation of the ith component. Now, the inverse of this diagonal matrix is just

$$\Sigma_{\tilde{x}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{bmatrix}, \tag{157}$$

and its determinant is $|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$ so that

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{158}$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{159}$$

$$= \prod_{i=1}^{d} f_{\tilde{x}_i}(x_i). \tag{160}$$

Since the joint pdf factors into the product of the marginals, the entries are all mutually independent. □

A fundamental property of Gaussian random vectors is that performing linear transformations on them always yields vectors with joint distributions that are also Gaussian. This is a multidimensional generalization of Lemma 8.2. We omit the proof, which is very similar.

Theorem 8.6 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let $\tilde{x}$ be a Gaussian random vector of dimension d with mean $\mu_{\tilde{x}}$ and covariance matrix $\Sigma_{\tilde{x}}$. For any matrix $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$, $\tilde{y} = A\tilde{x} + b$ is a Gaussian random vector with mean $\mu_{\tilde{x}} := A\mu_{\tilde{x}} + b$ and covariance matrix $\Sigma_{\tilde{y}} := A\Sigma_{\tilde{x}}A^T$, as long as $\Sigma_{\tilde{y}}$ is full rank.*

By Theorem 8.6 and Lemma 8.5, the principal components of a Gaussian random vector are independent. Let $\Sigma := U\Lambda U^T$ be the eigendecomposition of the covariance matrix of a Gaussian vector $\tilde{x}$. The vector containing the principal components

$$\widetilde{pc} := U^T \tilde{x} \tag{161}$$

has covariance matrix $U^T \Sigma U = \Lambda$, so the principal components are all independent. It is important to emphasize that this is the case because $\tilde{x}$ is Gaussian. In most cases, there will be nonlinear dependencies between the principal components (see Figures 8 and 14 for example).

In order to fit a Gaussian distribution to a dataset $\mathcal{X} := \{x_1, \ldots, x_n\}$ of d-dimensional points, a common approach is to maximize the log-likelihood of the data with respect to the mean and covariance parameters assuming independent samples,

$$(\mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}}) := \arg \max_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \log \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right) \tag{162}$$

$$= \arg \min_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \frac{n}{2} \log |\Sigma| . \tag{163}$$

The optimal parameters turn out to be the sample mean and the sample covariance matrix (we omit the proof, which relies heavily on matrix calculus). One can therefore interpret the analysis described in this chapter as fitting a Gaussian distribution to the data, but– as we hopefully have made clear– the analysis is meaningful even if the data are not Gaussian.