Covariance matrix

1 The covariance matrix

We have learned how to summarize datasets consisting of a single feature using the mean, median and variance, and datasets containing two features using the covariance and the correlation coefficient. Here we consider datasets containing multiple features, where each data point is modeled as a real-valued d-dimensional vector.

If we model the data as a d-dimensional random vector, its mean is defined as the vector formed by the means of its components.

Definition 1.1 (Mean of a random vector)**.** *The mean of a d-dimensional random vector $\tilde{x}$ is*

$$\mathrm{E}(\tilde{x}) := \begin{bmatrix} \mathrm{E}\left(\tilde{x}[1]\right) \\ \mathrm{E}\left(\tilde{x}[2]\right) \\ \cdots \\ \mathrm{E}\left(\tilde{x}[d]\right) \end{bmatrix}. \tag{1}$$

Similarly, we define the mean of a matrix with random entries as the matrix of entrywise means.

Definition 1.2 (Mean of a random matrix)**.** *The mean of a $d_1 \times d_2$ matrix with random entries $\tilde{X}$ is*

$$\mathrm{E}(\tilde{X}) := \begin{bmatrix} \mathrm{E}\left(\tilde{X}[1,1]\right) & \mathrm{E}\left(\tilde{X}[1,2]\right) & \cdots & \mathrm{E}\left(\tilde{X}[1,d_2]\right) \\ \mathrm{E}\left(\tilde{X}[2,1]\right) & \mathrm{E}\left(\tilde{X}[2,2]\right) & \cdots & \mathrm{E}\left(\tilde{X}[2,d_2]\right) \\ & & \cdots & \\ \mathrm{E}\left(\tilde{X}[d_1,1]\right) & \mathrm{E}\left(\tilde{X}[d_1,2]\right) & \cdots & \mathrm{E}\left(\tilde{X}[d_1,d_2]\right) \end{bmatrix}. \tag{2}$$

Linearity of expectation holds also for random vectors and random matrices.

Lemma 1.3 (Linearity of expectation for random vectors and matrices)**.** *Let $\tilde{x}$ a d-dimensional random vector, and let $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times d}$ for some positive integer m, then*

$$\mathrm{E}(A\tilde{x} + b) = A\mathrm{E}(\tilde{x}) + b. \tag{3}$$

Similarly let, $\tilde{X}$ be a $d_1 \times d_2$ random matrix, and let $B \in \mathbb{R}^{m \times d_2}$ and $A \in \mathbb{R}^{m \times d_1}$ for some positive integer m, then

$$\mathrm{E}(A\tilde{X} + B) = A\mathrm{E}(\tilde{X}) + B. \tag{4}$$

Proof. We prove the result for vectors, the proof for matrices is the same. The ith entry of $\mathrm{E}(A\tilde{x} + b)$ equals

$$\begin{aligned}
\mathrm{E}(A\tilde{x} + b)[i] &= \mathrm{E}\left((A\tilde{x} + b)[i]\right) \qquad \text{by definition of the mean for random vectors} && (5)\\
&= \mathrm{E}\left(\sum_{j=1}^{d} A[i,j]\tilde{x}[j] + b[i]\right) && (6)\\
&= \sum_{j=1}^{d} A[i,j]\mathrm{E}\left(\tilde{x}[j]\right) + b[i] \qquad \text{by linearity of expectation for scalars} && (7)\\
&= (A\mathrm{E}(\tilde{x}) + b)[i]. && (8)
\end{aligned}$$

□

We usually estimate the mean of random vectors by computing their sample mean, which equals the vector of sample means of the entries.

Definition 1.4 (Sample mean of multivariate data)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a set of d-dimensional vectors of real-valued data. The sample mean is the entry-wise average*

$$\mu_X := \frac{\sum_{i=1}^{n} x_i}{n}. \tag{9}$$

When manipulating a random vector within a probabilistic model, it may be useful to know the variance of linear combinations of its entries, i.e. the variance of the random variable $\langle v, \tilde{x} \rangle$ for some deterministic vector $v \in \mathbb{R}^d$. By linearity of expectation, this is given by

$$\begin{aligned}
\mathrm{Var}\left(v^T\tilde{x}\right) &= \mathrm{E}\left((v^T\tilde{x} - \mathrm{E}(v^T\tilde{x}))^2\right) && (10)\\
&= \mathrm{E}\left((v^T c(\tilde{x}))^2\right) && (11)\\
&= v^T \mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right) v, && (12)
\end{aligned}$$

where $c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$ is the centered random vector. For an example where $d = 2$ and the mean of $\tilde{x}$ is zero we have,

$$\begin{aligned}
\mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right) &= \mathrm{E}\left(\tilde{x}\tilde{x}^T\right) && (13)\\
&= \mathrm{E}\left(\begin{bmatrix}\tilde{x}[1]\\ \tilde{x}[2]\end{bmatrix}\begin{bmatrix}\tilde{x}[1] & \tilde{x}[2]\end{bmatrix}\right) && (14)\\
&= \mathrm{E}\left(\begin{bmatrix}\tilde{x}[1]^2 & \tilde{x}[1]\tilde{x}[2]\\ \tilde{x}[1]\tilde{x}[2] & \tilde{x}[2]^2\end{bmatrix}\right) && (15)\\
&= \begin{bmatrix}\mathrm{E}(\tilde{x}[1]^2) & \mathrm{E}(\tilde{x}[1]\tilde{x}[2])\\ \mathrm{E}(\tilde{x}[1]\tilde{x}[2]) & \mathrm{E}(\tilde{x}[2]^2)\end{bmatrix} && (16)\\
&= \begin{bmatrix}\mathrm{Var}\left(\tilde{x}[1]\right) & \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right)\\ \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \mathrm{Var}\left(\tilde{x}[2]\right)\end{bmatrix}. && (17)
\end{aligned}$$

This motivates defining the covariance matrix of the random vector as follows.

Definition 1.5 (Covariance matrix)**.** *The covariance matrix of a d-dimensional random vector* $\tilde{x}$ *is the* $d \times d$ *matrix*

$$\Sigma_{\tilde{x}} := \mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right) \tag{18}$$

$$= \begin{bmatrix} \mathrm{Var}\left(\tilde{x}[1]\right) & \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \cdots & \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) \\ \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \mathrm{Var}\left(\tilde{x}[2]\right) & \cdots & \mathrm{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) & \mathrm{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) & \cdots & \mathrm{Var}\left(\tilde{x}[d]\right) \end{bmatrix}, \tag{19}$$

where $c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$.

The covariance matrix encodes the variance of *any linear combination* of the entries of a random vector.

Lemma 1.6. *For any random vector* $\tilde{x}$ *with covariance matrix* $\Sigma_{\tilde{x}}$*, and any vector* v

$$\mathrm{Var}\left(v^T\tilde{x}\right) = v^T\Sigma_{\tilde{x}}v. \tag{20}$$

Proof. This follows immediately from Eq. (12). □

Example 1.7 (Cheese sandwich)**.** A deli in New York is worried about the fluctuations in the cost of their signature cheese sandwich. The ingredients of the sandwich are bread, a local cheese, and an imported cheese. They model the price in cents per gram of each ingredient as an entry in a three dimensional random vector $\tilde{x}$. $\tilde{x}[1]$, $\tilde{x}[2]$, and $\tilde{x}[3]$ represent the price of the bread, the local cheese and the imported cheese respectively. From past data they determine that the covariance matrix of $\tilde{x}$ is

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1.2 \end{bmatrix}. \tag{21}$$

They consider two recipes; one that uses 100g of bread, 50g of local cheese, and 50g of imported cheese, and another that uses 100g of bread, 100g of local cheese, and no imported cheese. By Lemma 1.6 the standard deviation in the price of the first recipe equals

$$\sigma_{100\tilde{x}[1]+50\tilde{x}[2]+50\tilde{x}[3]} = \sqrt{\begin{bmatrix} 100 & 50 & 50 \end{bmatrix} \Sigma_{\tilde{x}} \begin{bmatrix} 100 \\ 50 \\ 50 \end{bmatrix}} \tag{22}$$

$$= 153 \text{ cents}. \tag{23}$$

The standard deviation in the price of the second recipe equals

$$\sigma_{100\tilde{x}[1]+100\tilde{x}[2]} = \sqrt{\begin{bmatrix} 100 & 100 & 0 \end{bmatrix} \Sigma_{\tilde{x}} \begin{bmatrix} 100 \\ 100 \\ 0 \end{bmatrix}} \tag{24}$$

$$= 164 \text{ cents}. \tag{25}$$

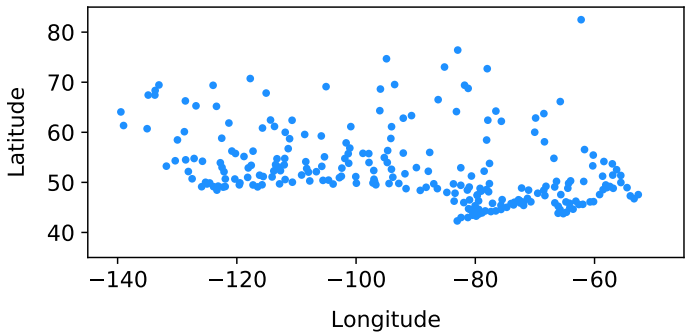

Figure 1: **Canadian cities.** Scatterplot of the latitude and longitude of the main 248 cities in Canada.

Even though the price of the imported cheese is more volatile than that of the local cheese, adding it to the recipe lowers the variance of the cost because it is uncorrelated with the other ingredients. △

A natural way to estimate the covariance matrix from data is to compute the sample covariance matrix.

Definition 1.8 (Sample covariance matrix)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a set of d-dimensional vectors of real-valued data. The sample covariance matrix equals*

$$\Sigma_X := \frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T \tag{26}$$

$$= \begin{bmatrix} \sigma^2_{X[1]} & \sigma_{X[1],X[2]} & \cdots & \sigma_{X[1],X[d]} \\ \sigma_{X[1],X[2]} & \sigma^2_{X[2]} & \cdots & \sigma_{X[2],X[d]} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X[1],X[d]} & \sigma_{X[2],X[d]} & \cdots & \sigma^2_{X[d]} \end{bmatrix}, \tag{27}$$

where $c(x_i) := x_i - \mu_X$ for $1 \leq i \leq n$, $X[j] := \{x_1[j], \ldots, x_n[j]\}$ for $1 \leq j \leq d$, $\sigma^2_{X[i]}$ is the sample variance of $X[i]$, and $\sigma_{X[i],X[j]}$ is the sample covariance of the entries of $X[i]$ and $X[j]$.

Example 1.9 (Canadian cities)**.** We consider a dataset which contains the locations (latitude and longitude) of major cities in Canada[1] (so $d = 2$ in this case). Figure 1 shows a scatterplot of the data. The sample covariance matrix is

$$\Sigma_X = \begin{bmatrix} 524.9 & -59.8 \\ -59.8 & 53.7 \end{bmatrix}. \tag{28}$$

The latitudes have much higher variance than the longitudes. Latitude and longitude are negatively correlated because people at higher longitudes (in the east) tend to live at lower latitudes (in the south). △

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

It turns out that just like the covariance matrix encodes the variance of any linear combination of a random vector, the sample covariance matrix encodes the sample variance of any linear combination of the data.

Lemma 1.10. *For any dataset $X = \{x_1, \ldots, x_n\}$ of d-dimensional data and any vector $v \in \mathbb{R}^d$, let*

$$X_v := \{\langle v, x_1 \rangle, \ldots, \langle v, x_n \rangle\} \tag{29}$$

be the set of inner products between v and the elements in X. Then

$$\sigma_{X_v}^2 = v^T \Sigma_X v. \tag{30}$$

Proof.

$$\begin{aligned}
\sigma_{X_v}^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(v^T x_i - \mu_{X_v} \right)^2 && (31)\\
&= \frac{1}{n} \sum_{i=1}^{n} \left(v^T x_i - \frac{1}{n} \sum_{j=1}^{n} v^T x_j \right)^2 && (32)\\
&= \frac{1}{n} \sum_{i=1}^{n} \left(v^T \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right) \right)^2 && (33)\\
&= \frac{1}{n} \sum_{i=1}^{n} (v^T c(x_i))^2 && (34)\\
&= \frac{1}{n} \sum_{i=1}^{n} v^T c(x_i) c(x_i)^T v && (35)\\
&= v^T \left(\frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T \right) v && (36)\\
&= v^T \Sigma_X v. && (37)
\end{aligned}$$

□

The component of a random vector lying in a specific direction can be computed by taking their inner products with a unit-norm vector u pointing in that direction. As a result, by Lemma 1.6 the covariance matrix describes the variance of a random vector in any direction of its ambient space. Similarly, the sample covariance matrix describes the sample variance of the data in any direction by Lemma 1.10, as illustrated in the following example.

Example 1.11 (Variance in a specific direction)**.** We consider the question of how the distribution of Canadian cities varies in specific directions. This can be computed from the sample covariance matrix. Let us consider a southwest-northeast direction. The positions of the cities in that direction are given by the inner product of their locations with the unit-norm vector

$$v := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{38}$$

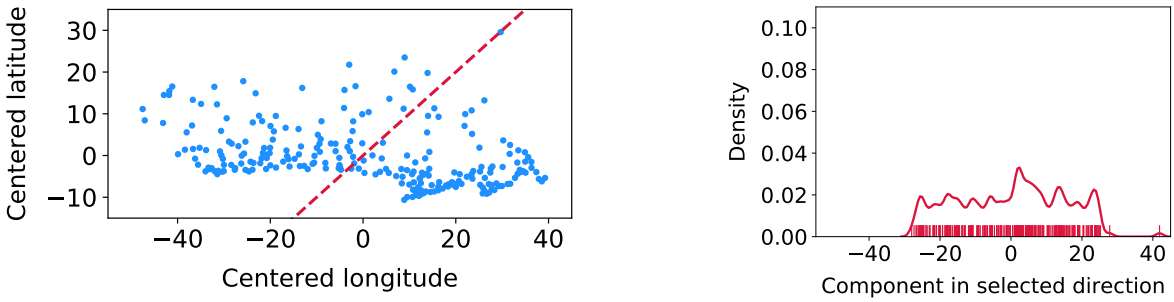

Figure 2: **Sample variance in southwest-northeast direction.** The left scatterplot shows the centered data from Figure 1, and a fixed direction of the two-dimensional space represented by a line going through the origin from southwest to northeast. The right plot shows the components of each data point in the direction of the line and a kernel density estimate. The sample standard deviation of the components is 15.1.

By Lemma 1.10 we have

$$\sigma^2_{X_u} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \Sigma_X \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{39}$$

$$= 229, \tag{40}$$

so the standard deviation is 15.1. Figure 2 shows the direction of interesting on the scatterplot, as well as a kernel density estimate of the components of the positions in that direction. Figure 3 shows the sample variance in every possible direction, given by the quadratic form

$$q(v) := v^T \Sigma_X v, \tag{41}$$

for all possible unit-norm vectors v. △

2 Principal component analysis

As explained at the end of the last section, the covariance matrix $\Sigma_{\tilde{x}}$ of a random vector $\tilde{x}$ encodes the variance of the vector in every possible direction of space. In this section, we consider the question of finding the directions of maximum and minimum variance. The variance in the direction of a vector v is given by the quadratic form $v^T \Sigma_{\tilde{x}} v$. By the following fundamental theorem in linear algebra, quadratic forms are best understood in terms of the eigendecomposition of the corresponding matrix.

Theorem 2.1 (Spectral theorem for symmetric matrices)**.** *If $A \in \mathbb{R}^{d \times d}$ is symmetric, then it has an eigendecomposition of the form*

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T, \tag{42}$$

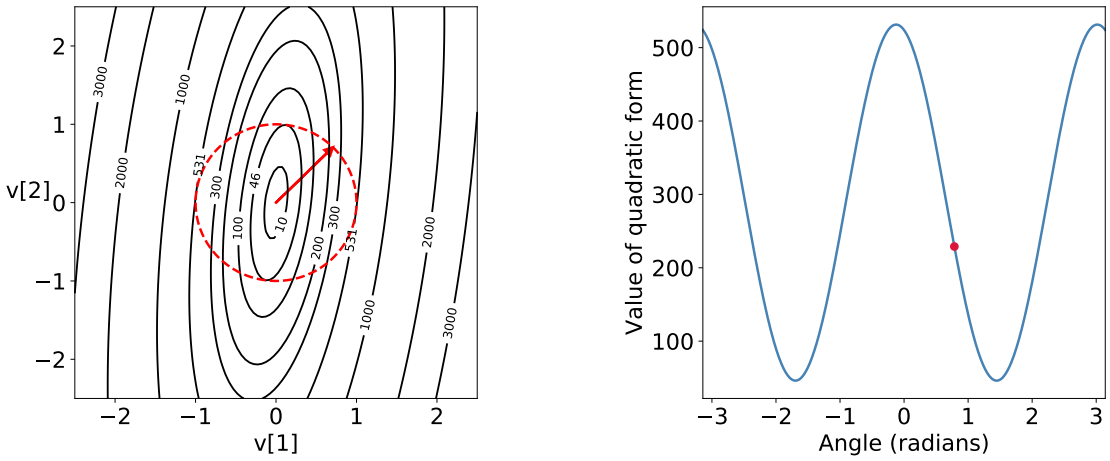

Figure 3: **Sample variance in different directions.** The left plot shows the contours of the quadratic form $v^T \Sigma_X v$, where Σ_X is the sample covariance matrix of the data in Figure 1. The unit circle, where $||v||_2 = 1$, is drawn in red. The red arrow is a unit vector collinear with the dashed red line on the left plot of Figure 2. 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.

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,*

$$\lambda_1 = \max_{||x||_2=1} x^T A x, \tag{43}$$

$$u_1 = \arg\max_{||x||_2=1} x^T A x, \tag{44}$$

$$\lambda_k = \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \leq k \leq d-1, \tag{45}$$

$$u_k = \arg\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \leq k \leq d-1, \tag{46}$$

$$\lambda_d = \min_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \tag{47}$$

$$u_d = \arg\min_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x. \tag{48}$$

In order to characterize the variance of a random vector in different directions, 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, ..., 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 2.2. *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

$$\lambda_1 = \max_{||v||_2=1} \text{Var}(v^T \tilde{x}), \tag{49}$$

$$u_1 = \arg \max_{||v||_2=1} \text{Var}(v^T \tilde{x}), \tag{50}$$

$$\lambda_k = \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d, \tag{51}$$

$$u_k = \arg \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \text{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d. \tag{52}$$

Proof. Covariance matrices are symmetric by definition. The result follows automatically from Theorem 2.1 and Lemma 1.6. □

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*,

$$\widetilde{pc}[i] := u_i^T c(\tilde{x}), \quad 1 \leq i \leq d \tag{53}$$

By Theorem 2.2 the variance of each principal component is the corresponding eigenvalue of the covariance matrix,

$$\text{Var}\left(\widetilde{pc}[i]\right) = u_i^T \Sigma_{\tilde{x}} u_i \tag{54}$$

$$= \lambda_i u_i^T u_i \tag{55}$$

$$= \lambda_i. \tag{56}$$

Interestingly, the principal components of a random vectors are uncorrelated, which means that there is no linear relationship between them.

Lemma 2.3. *The principal components of a random vector $\tilde{x}$ are uncorrelated.*

Proof. Let u_i be the eigenvector of the covariance matrix corresponding to the ith principal component. We have

$$\text{E}(\widetilde{pc}[i]\widetilde{pc}[j]) = \text{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) \tag{57}$$

$$= u_i^T \text{E}(c(\tilde{x}) c(\tilde{x})^T) u_j \tag{58}$$

$$= u_i^T \Sigma_{\tilde{x}} u_j \tag{59}$$

$$= \lambda_j u_i^T u_j \tag{60}$$

$$= 0, \tag{61}$$

by orthogonality of the eigenvectors of a symmetric matrix. □

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

Algorithm 2.4 (Principal component analysis (PCA))**.** *Given a dataset X containing n vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with d features each, where $n > d$.*

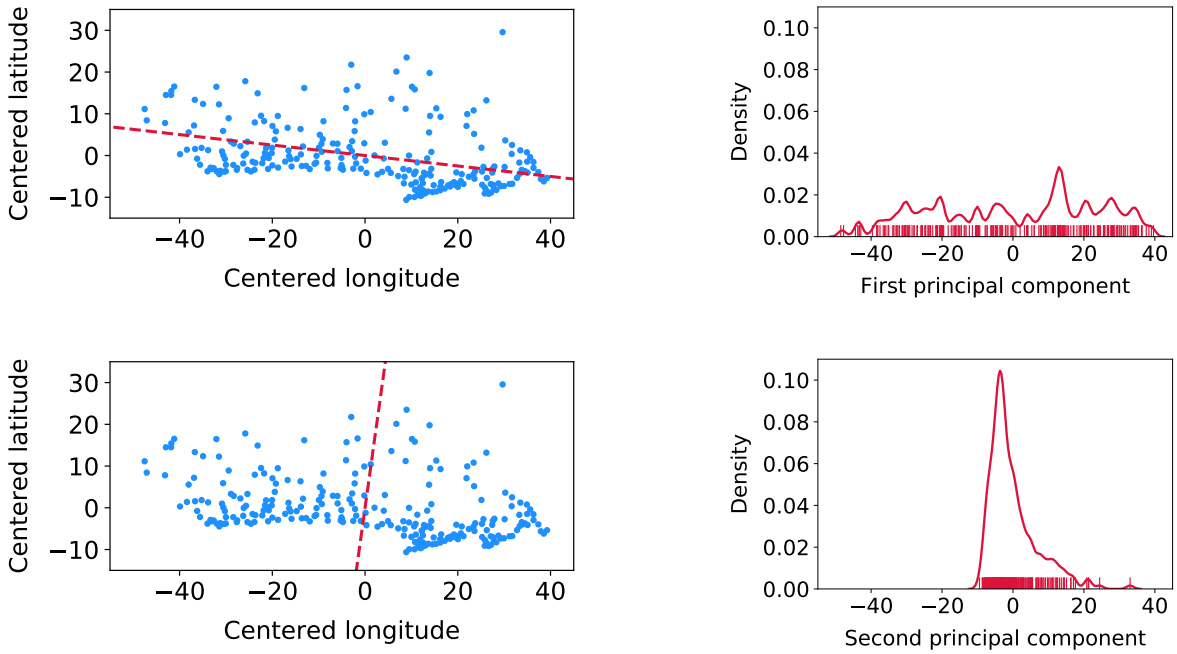

Figure 4: **Principal directions.** The scatterplots in the left column show the centered data from Figure 1, and the first (top) and second (bottom) principal directions of the data represented by lines going through the origin. The right column shows the first (top) and second (bottom) principal components 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)

1. *Compute the sample covariance matrix of the data* Σ_X.
2. *Compute the eigendecomposition of* Σ_X*, to find the principal directions* $u_1, \ldots, u_d$.
3. *Center the data and compute the principal components*

$$pc_i[j] := u_j^T c(x_i), \quad 1 \leq i \leq n, \; 1 \leq j \leq d, \tag{62}$$

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

When we perform PCA on a dataset, the resulting principal directions maximize (and minimize) the sample variance. This again follows from the spectral theorem (Theorem 2.1), in this case combined with Lemma 1.10.

Theorem 2.5. *Let* X *contain* n *vectors* $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ *with sample covariance matrix* Σ_X*, and let* $u_1, \ldots, u_d$*, and* $\lambda_1 > \ldots > \lambda_d$ *denote the eigenvectors and corresponding eigenvalues of*

Σ_X. *We have*

$$\lambda_1 = \max_{||v||_2=1} \sigma^2_{X_v}, \tag{63}$$

$$u_1 = \arg\max_{||v||_2=1} \sigma^2_{X_v}, \tag{64}$$

$$\lambda_k = \max_{||v||_2=1, v\perp u_1,\dots,u_{k-1}} \sigma^2_{X_v}, \quad 2 \leq k \leq d, \tag{65}$$

$$u_k = \arg\max_{||v||_2=1, v\perp u_1,\dots,u_{k-1}} \sigma^2_{X_v}, \quad 2 \leq k \leq d. \tag{66}$$

Proof. Sample covariance matrices are symmetric by definition. The result follows automatically from Theorem 2.1 and Lemma 1.10. □

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 4 shows the principal directions and the principal components for the data in Figure 1. Comparing the principal components to the component in the direction shown in Figure 2, we confirm that the first principal component has larger sample variance, and the second principal component has smaller sample variance.

Example 2.6 (PCA of faces)**.** The Olivetti Faces dataset[2] contains 400 64×64 images taken from 40 different subjects (10 per subject). We vectorize each image so that each pixel is interpreted as a different feature. Figure 5 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. △

3 Gaussian random vectors

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

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right), \tag{67}$$

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.

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

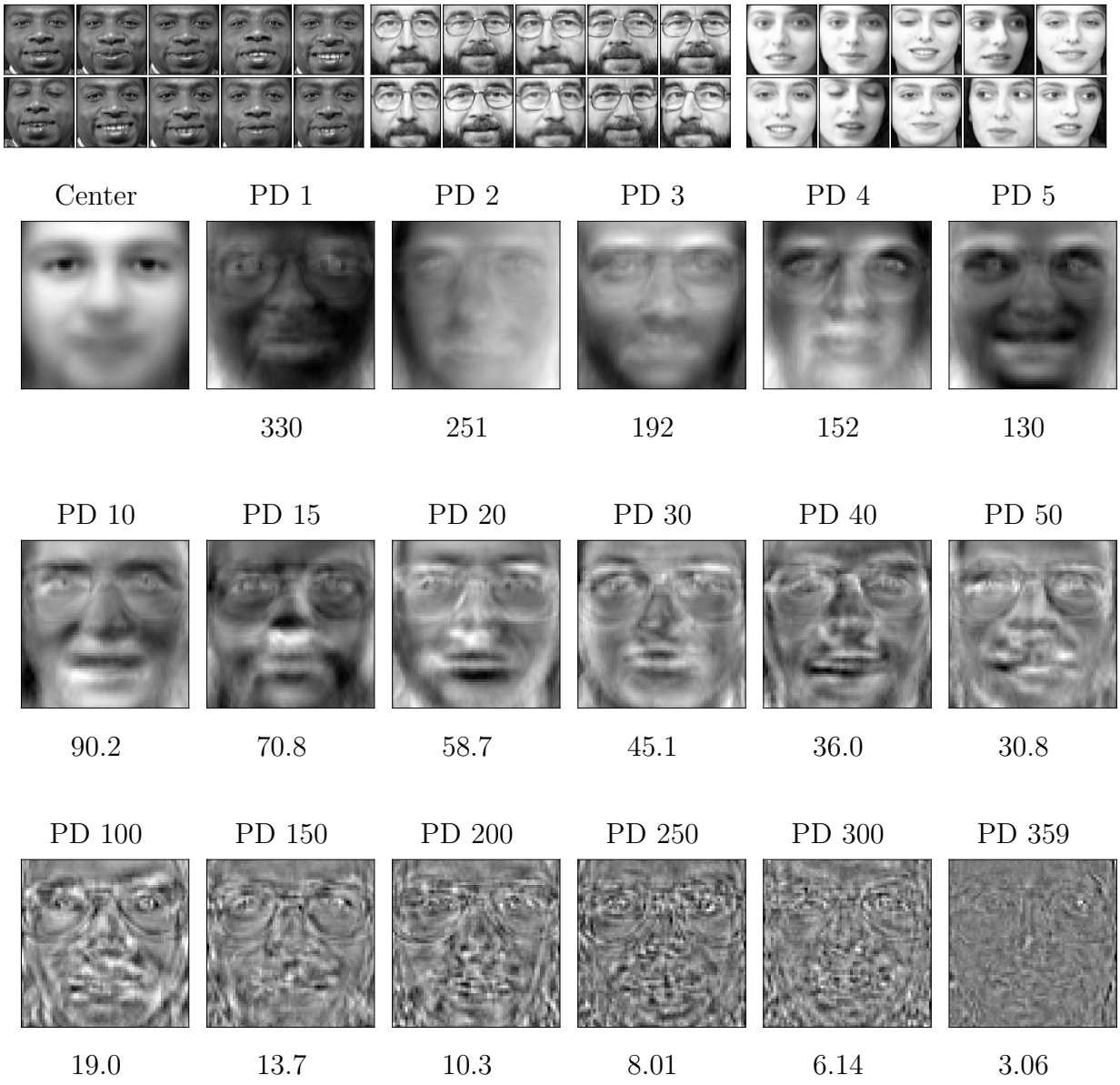

Figure 5: The top row shows the data corresponding to three different individuals in the Olivetti dataset. The sample mean and the principal directions (PD) obtained by applying PCA to the centered data are depicted below. The sample standard deviation of each principal component is listed below the corresponding principal direction.

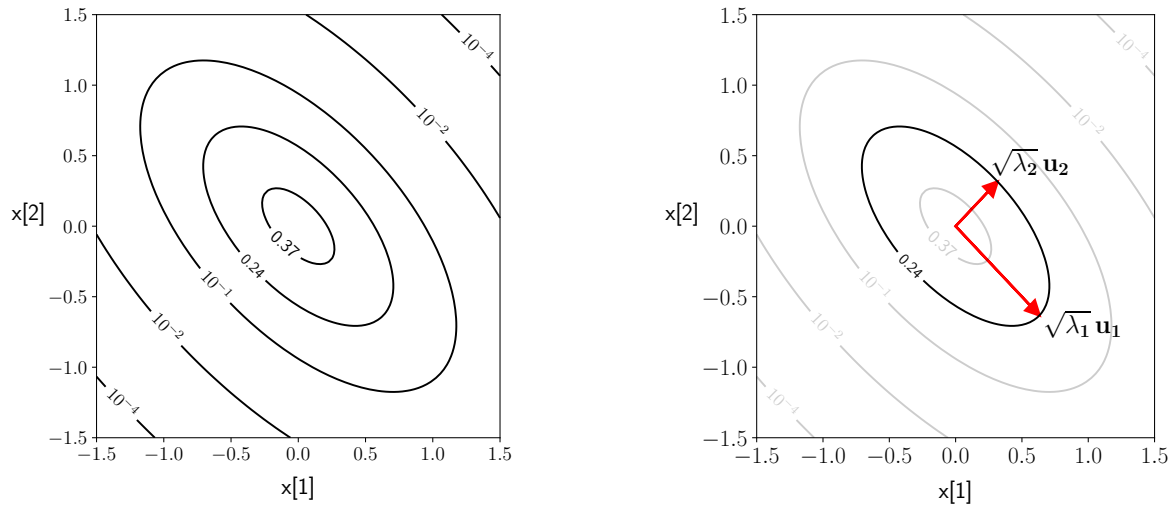

Figure 6: **Contour surfaces of a Gaussian vector.** The left image shows a contour plot of the probability density function of the two-dimensional Gaussian random vector defined in Example 3.2. The axes align with the eigenvectors of the covariance matrix, and are proportional to the square root of the eigenvalues, as shown on the right image for a specific contour.

In order to better understand the geometry of the pdf of Gaussian random vectors, we analyze their contour surfaces. The contour surfaces are sets of points where the density is constant. The spectral theorem (Theorem 2.1) 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

$$\begin{aligned} c &= x^T\Sigma^{-1}x && (68)\\ &= x^T U\Lambda^{-1}U^T x && (69)\\ &= \sum_{i=1}^{d} \frac{(u_i^T x)^2}{\lambda_i}. && (70) \end{aligned}$$

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.

Example 3.2 (Two-dimensional Gaussian)**.** We illustrate the geometry of the Gaussian probability distribution function with a two-dimensional example where μ is zero and

$$\Sigma = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}. \tag{71}$$

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

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}. \tag{72}$$

The left plot of Figure 6 shows several contours of the density. The right plot shows the axes for the contour line

$$\frac{(u_1^T x)^2}{\lambda_1} + \frac{(u_2^T x)^2}{\lambda_2} = 1, \tag{73}$$

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 3.3 (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

$$\Sigma_{\tilde{x}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}, \tag{74}$$

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

$$\Sigma_{\tilde{x}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{bmatrix}, \tag{75}$$

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

$$f_{\tilde{x}}(a) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{76}$$

$$= \frac{1}{\prod_{i=1}^d \sqrt{(2\pi)}\sigma_i} \exp\left(\sum_{i=1}^d -\frac{(x[i]-\mu[i])^2}{2\sigma_i^2}\right) \tag{77}$$

$$= \prod_{i=1}^d \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{(x[i]-\mu[i])^2}{2\sigma_i^2}\right) \tag{78}$$

$$= \prod_{i=1}^d f_{\tilde{x}[i]}(x[i]). \tag{79}$$

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 the univariate result. We omit the proof, which is very similar.

Theorem 3.4 (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 3.4 and Lemma 3.3, 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

$$\widetilde{pc} := U^T\tilde{x} \tag{80}$$

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 Figure 4 for example).

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

$$(\mu_{\text{ML}}, \Sigma_{\text{ML}}) := \arg \max_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \log \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right) \tag{81}$$

$$= \arg \min_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \frac{n}{2} \log |\Sigma| \,. \tag{82}$$

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.