Chapter 9

Frequentist Statistics

The goal of statistical analysis is to *extract information* from data by computing **statistics**, which are deterministic functions of the data. In Chapter 8 we describe several statistics from a deterministic and geometric point of view, without making any assumptions about the data-generation process. This makes it very challenging to evaluate the accuracy of the acquired information.

In this chapter we model the data-acquisition process probabilistically. This allows to analyze statistical techniques and derive theoretical guarantees on their performance. The data are interpreted as **realizations** of random variables, vectors or processes (depending on the dimensionality). The information that we want to extract can then be expressed in terms of the joint distribution of these quantities. We consider this distribution to be unknown but **fixed**, taking a **frequentist** perspective. The alternative framework of Bayesian statistics is described in Chapter 10.

9.1 Independent identically-distributed sampling

In this chapter we consider one-dimensional real-valued data, modeled as the realization of an iid sequence. Figure 9.1 depicts the corresponding graphical model. This is a very popular assumption, which holds for controlled experiments, such as randomized trials to test drugs, and can often be a good approximation in other settings. However, in practice it is crucial to evaluate to what extent the independence assumptions of a model actually hold.

The following example shows that measuring a quantity by sampling a subset of individuals randomly from a large population produces data satisfying the iid assumption, as long as we sample with replacement (if the population is large, sampling without replacement will have a negligible effect).

Example 9.1.1 (Sampling from a population)**.** Assume that we are studying a population of m individuals. We are interested in a certain quantity associated to each person, e.g. their cholesterol level, their salary or who they are voting for in an election. There are k possible values for the quantity $\{z_1, z_2, \ldots, z_k\}$, where k can be equal to m or much smaller. We denote by m_j the number of people for whom the quantity is equal to z_j, $1 \leq j \leq k$. In the case of an election with two candidates, k would equal two and m_1 and m_2 would represent the people voting for each of the candidates.

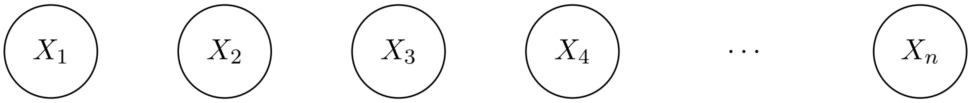

Figure 9.1: Directed graphical model corresponding to an independent sequence. If the sequence is also identically distributed, then X_1, X_2, …, X_n all have the same distribution.

Let us assume that we select n individuals independently at random with replacement, which means that one individual could be chosen more than once, and record the value of the quantity of interest. Under these assumptions the measurements can be modeled as a random sequence of independent variables $\widetilde{X}$. Since the probability of choosing any individual is the same every time we make a selection, the first-order pmf of the sequence is

$$\begin{aligned} p_{\widetilde{X}(i)}\left(z_j\right) &= P\left(\text{The } i\text{th measurement equals } z_j\right) && (9.1)\\ &= \frac{\text{People such that the quantity equals } z_j}{\text{Total number of people}} && (9.2)\\ &= \frac{m_j}{m}, \qquad 1 \leq j \leq k, && (9.3) \end{aligned}$$

for $1 \leq i \leq n$ by the law of total probability. We conclude that the data can be modeled as a realization of an iid sequence. △

9.2 Mean square error

We define an **estimator** as a deterministic function of the available data $x_1, x_2, \ldots, x_n$ which provides an approximation to a quantity associated to the distribution that generates the data

$$y := h\left(x_1, x_2, \ldots, x_n\right). \tag{9.4}$$

For example, as we will see, if we want to estimate the expectation of the underlying distribution, a reasonable estimator is the average of the data. Since we are taking a frequentist viewpoint, the quantity of interest is modeled as deterministic (in contrast to the Bayesian viewpoint which would model it as a random variable). For a fixed data set, the estimator is a deterministic function of the data. However, if we model the data as realizations of a sequence of random variables, then the estimator is also a realization of the random variable

$$Y := h\left(X_1, X_2, \ldots, X_n\right). \tag{9.5}$$

This allows to evaluate the estimator probabilistically (usually under some assumptions on the underlying distribution). For instance, we can measure the error incurred by the estimator by computing the mean square of the difference between the estimator and the true quantity of interest.

Definition 9.2.1 (Mean square error)**.** *The mean square error (MSE) of an estimator Y that approximates a deterministic quantity $\gamma \in \mathbb{R}$ is*

$$\operatorname{MSE}(Y) := \operatorname{E}\left(\left(Y-\gamma\right)^2\right). \tag{9.6}$$

The MSE can be decomposed into a **bias** term and a **variance** term. The bias term is the difference between the quantity of interest and the expected value of the estimator. The variance term corresponds to the variation of the estimator around its expected value.

Lemma 9.2.2 (Bias-variance decomposition). *The MSE of an estimator Y that approximates $\gamma \in \mathbb{R}$ satisfies*

$$MSE(Y) = \underbrace{\mathrm{E}\left((Y - \mathrm{E}(Y))^2\right)}_{\text{variance}} + \underbrace{(\mathrm{E}(Y) - \gamma)^2}_{\text{bias}}. \tag{9.7}$$

Proof. The lemma is a direct consequence of linearity of expectation. □

If the bias is zero, then the estimator equals the quantity of interest on average.

Definition 9.2.3 (Unbiased estimator). *An estimator Y that approximates $\gamma \in \mathbb{R}$ is unbiased if its bias is equal to zero, i.e. if and only if*

$$\mathrm{E}(Y) = \gamma. \tag{9.8}$$

An estimator may be unbiased but still incur in a large mean square error due to its variance.

The following lemmas establish that the sample mean and variance are unbiased estimators of the true mean and variance of an iid sequence of random variables.

Lemma 9.2.4 (The sample mean is unbiased). *The sample mean is an unbiased estimator of the mean of an iid sequence of random variables.*

Proof. We consider the sample mean of an iid sequence $\widetilde{X}$ with mean μ,

$$\widetilde{Y}(n) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}(i). \tag{9.9}$$

By linearity of expectation

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(\widetilde{X}(i)\right) \tag{9.10}$$

$$= \mu. \tag{9.11}$$

□

Lemma 9.2.5 (The sample variance is unbiased). *The sample variance is an unbiased estimator of the variance of an iid sequence of random variables.*

The proof of this result is in Section 9.7.1.

9.3 Consistency

If we are estimating a scalar quantity, the estimate should improve as we gather more data. Ideally the estimate should converge to the true value in the limit when the number of data $n \to \infty$. Estimators that achieve this are said to be consistent.

Definition 9.3.1 (Consistency)**.** *An estimator $\widetilde{Y}(n) := h\left(\widetilde{X}(1), \widetilde{X}(2), \ldots, \widetilde{X}(n)\right)$ that approximates $\gamma \in \mathbb{R}$ is consistent if it converges to γ as $n \to \infty$ in mean square, with probability one or in probability.*

The following theorem shows that the mean is consistent.

Theorem 9.3.2 (The sample mean is consistent)**.** *The sample mean is a consistent estimator of the mean of an iid sequence of random variables as long as the variance of the sequence is bounded.*

Proof. We consider the sample mean of an iid sequence $\widetilde{X}$ with mean μ,

$$\widetilde{Y}(n) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}(i). \tag{9.12}$$

The estimator is equal to the moving average of the data. As a result it converges to μ in mean square (and with probability one) by the law of large numbers (Theorem 6.2.2), as long as the variance σ^2 of each of the entries in the iid sequence is bounded. □

Example 9.3.3 (Estimating the average height)**.** In this example we illustrate the consistency of the sample mean. Imagine that we want to estimate the mean height in a population. To be concrete we consider a population of $m := 25000$ people. Figure 9.2 shows a histogram of their heights.[1] As explained in Example 9.1.1 if we sample n individuals from this population with replacement, then their heights form an iid sequence $\widetilde{X}$. The mean of this sequence is

$$\begin{aligned} \mathrm{E}\left(\widetilde{X}(i)\right) &:= \sum_{j=1}^{m} \mathrm{P}\left(\text{Person } j \text{ is chosen}\right) \cdot \text{height of person } j && (9.13)\\ &= \frac{1}{m}\sum_{j=1}^{m} h_j && (9.14)\\ &= \operatorname{av}\left(h_1, \ldots, h_m\right) && (9.15)\end{aligned}$$

for $1 \leq i \leq n$, where $h_1, \ldots, h_m$ are the heights of the people. In addition, the variance is bounded because the heights are finite. By Theorem 9.3.2 the sample mean of the n data should converge to the mean of the iid sequence and hence to the average height over the whole population. Figure 9.3 illustrates this numerically.

△

If the mean of the underlying distribution is not well defined, or its variance is unbounded, then the sample mean is not necessarily a consistent estimator. This is related to the fact that

[1] The data are available here: wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights.

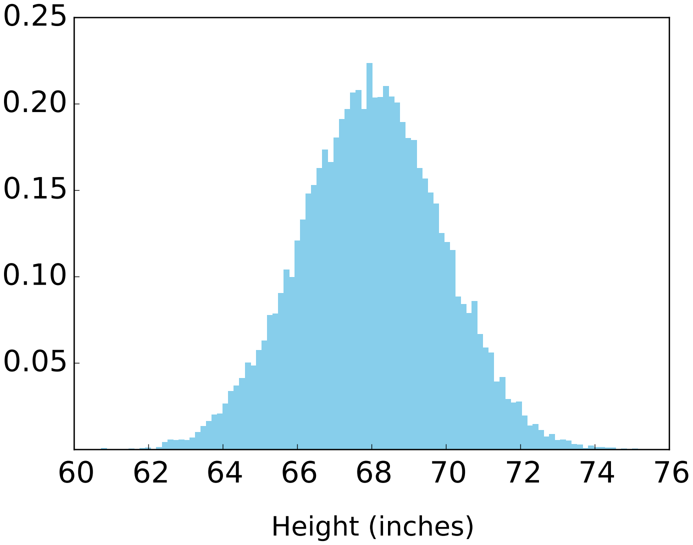

Figure 9.2: Histogram of the heights of a group of 25 000 people.

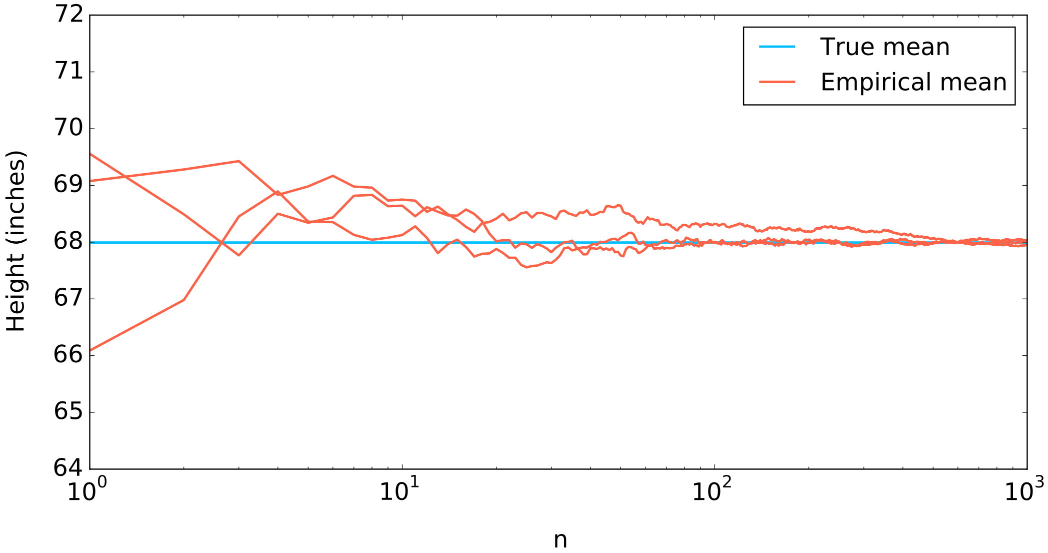

Figure 9.3: Different realizations of the sample mean when individuals from the population in Figure 9.2 are sampled with replacement.

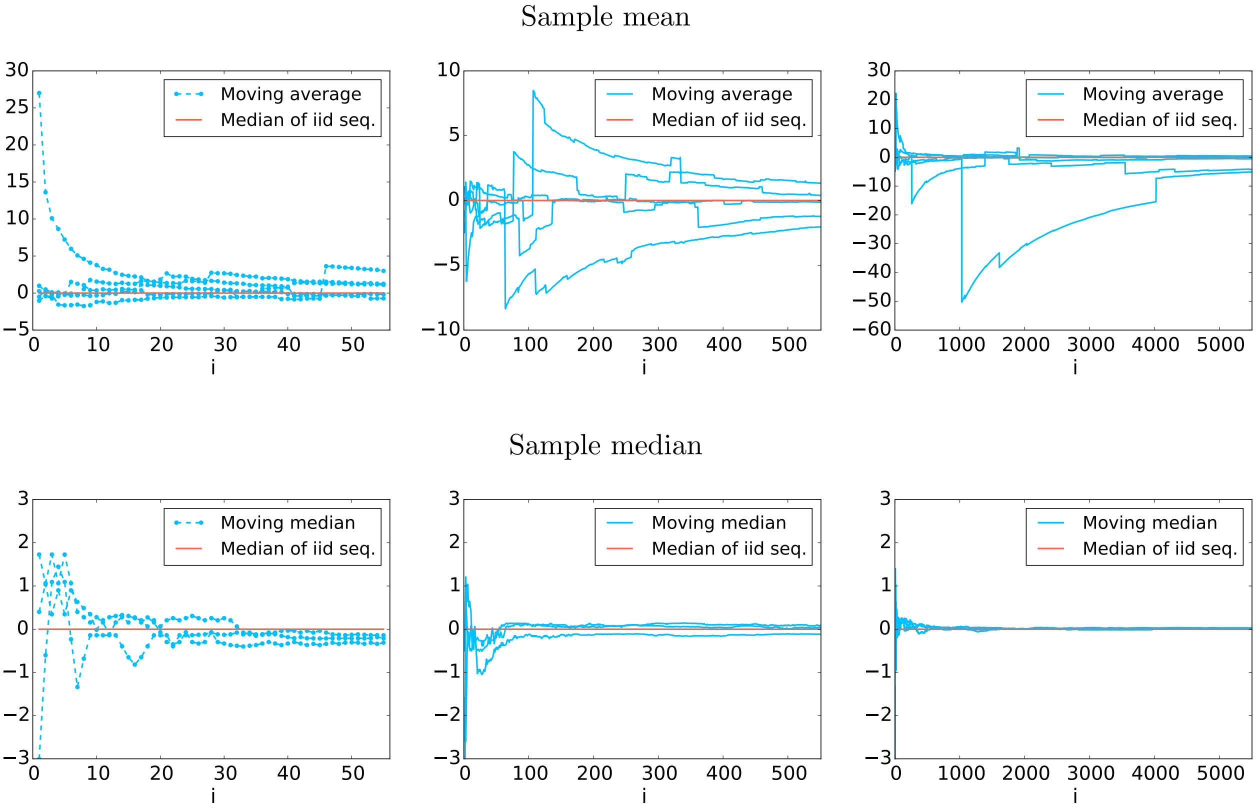

Figure 9.4: Realization of the moving average of an iid Cauchy sequence (top) compared to the moving median (bottom).

the sample mean can be severely affected by the presence of extreme values, as we discussed in Section 8.2. The sample median, in contrast, tends to be more robust in such situations, as discussed in Section 8.3. The following theorem establishes that the sample median is consistent under the iid assumption, even if the mean is not well defined or the variance is unbounded. The proof is in Section 9.7.2.

Theorem 9.3.4 (Sample median as an estimator of the median)**.** *The sample median is a consistent estimator of the median of an iid sequence of random variables.*

Figure 9.4 compares the moving average and the moving median of an iid sequence of Cauchy random variables for three different realizations. The moving average is unstable and does not converge no matter how many data are available, which is not surprising because the mean is not well defined. In contrast, the moving median does eventually converge to the true median as predicted by Theorem 9.3.4.

The sample variance and covariance are consistent estimators of the variance and covariance respectively, under certain assumptions on the higher moments of the underlying distributions. This provides an intuitive interpretation for principal component analysis (see Section 8.5.2) under the assumption that the data are realizations of an iid sequence of random vectors: the principal components approximate the eigenvectors of the true covariance matrix (see Section 4.3.3), and hence the directions of maximum variance of the multidimensional distribution. Figure 9.5

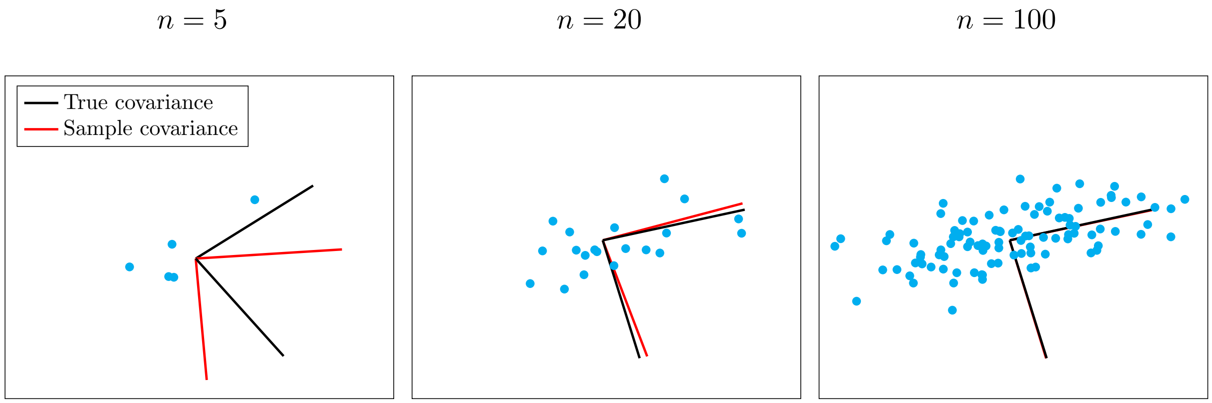

Figure 9.5: Principal directions of n samples from a bivariate Gaussian distribution (red) compared to the eigenvectors of the covariance matrix of the distribution (black).

illustrates this with a numerical example, where the principal components indeed converge to the eigenvectors as the number of data increases.

9.4 Confidence intervals

Consistency implies that an estimator will be perfect if we acquire infinite data, but this is of course impossible in practice. It is therefore important to quantify the accuracy of an estimator for a fixed number of data. Confidence intervals allow to do this from a frequentist point of view. A confidence interval can be interpreted as a *soft estimate* of the deterministic quantity of interest, which guarantees that the true value will belong to the interval with a certain probability.

Definition 9.4.1 (Confidence interval). *A $1-\alpha$ confidence interval $\mathcal{I}$ for $\gamma \in \mathbb{R}$ satisfies*

$$\mathrm{P}\left(\gamma \in \mathcal{I}\right) \geq 1-\alpha, \tag{9.16}$$

where $0 < \alpha < 1$.

Confidence intervals are usually of the form $[Y - c, Y + c]$ where Y is an estimator of the quantity of interest and c is a constant that depends on the number of data. The following theorem derives a confidence interval for the mean of an iid sequence. The confidence interval is centered at the sample mean.

Theorem 9.4.2 (Confidence interval for the mean of an iid sequence). *Let $\widetilde{X}$ be an iid sequence with mean μ and variance $\sigma^2 \leq b^2$ for some $b > 0$. For any $0 < \alpha < 1$*

$$\mathcal{I}_n := \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{b}{\sqrt{\alpha\, n}}\right], \qquad Y_n := \operatorname{av}\left(\widetilde{X}\left(1\right), \widetilde{X}\left(2\right), \ldots, \widetilde{X}\left(n\right)\right), \tag{9.17}$$

is a $1-\alpha$ confidence interval for μ.

Proof. Recall that the variance of Y_n equals $\operatorname{Var}\left(\bar{X}_n\right) = \sigma^2/n$ (see equation (6.21) in the proof of Theorem 6.2.2). We have

$$
\begin{aligned}
P\left(\mu \in \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]\right) &= 1 - P\left(|Y_n - \mu| > \frac{b}{\sqrt{\alpha\, n}}\right) && (9.18)\\
&\geq 1 - \frac{\alpha\, n \operatorname{Var}\left(Y_n\right)}{b^2} \quad \text{by Chebyshev's inequality} && (9.19)\\
&= 1 - \frac{\alpha\, \sigma^2}{b^2} && (9.20)\\
&\geq 1 - \alpha. && (9.21)
\end{aligned}
$$

□

The width of the interval provided in the theorem decreases with n for fixed α, which makes sense as incorporating more data reduces the variance of the estimator and hence our uncertainty about it.

Example 9.4.3 (Bears in Yosemite)**.** A scientist is trying to estimate the average weight of the black bears in Yosemite National Park. She manages to capture 300 bears. We assume that the bears are sampled uniformly at random with replacement (a bear can be weighed more than once). Under this assumptions, in Example 9.1.1 we show that the data can be modeled as iid samples and in Example 9.3.3 we show the sample mean is a consistent estimator of the mean of the whole population.

The average weight of the 300 captured bears is $Y := 200$ lbs. To derive a confidence interval from this information we need a bound on the variance. The maximum weight recorded for a black bear ever is 880 lbs. Let μ and σ^2 be the (unknown) mean and variance of the weights of the whole population. If X is the weight of a bear chosen uniformly at random from the whole population then X has mean μ and variance σ^2, so

$$
\begin{aligned}
\sigma^2 &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) && (9.22)\\
&\leq \mathrm{E}\left(X^2\right) && (9.23)\\
&\leq 880^2 \quad \text{because } X \leq 880. && (9.24)
\end{aligned}
$$

As a result, 880 is an upper bound for the standard deviation. Applying Theorem 9.4.2,

$$
\left[Y - \frac{b}{\sqrt{\alpha\, n}}, Y + \frac{b}{\sqrt{\alpha\, n}}\right] = [-27.2, 427.2] \tag{9.25}
$$

is a 95% confidence interval for the average weight of the whole population. The interval is not very precise because n is not very large. △

As illustrated by this example, confidence intervals derived from Chebyshev's inequality tend to be very conservative. An alternative is to leverage the central limit theorem (CLT). The CLT characterizes the distribution of the sample mean asymptotically, so confidence intervals derived from it are not guaranteed to be precise. However, the CLT often provides a very accurate approximation to the distribution of the sample mean for finite n, as we show through some numerical examples in Chapter 6. In order to obtain confidence intervals for the mean of an iid sequence from the CLT as stated in Theorem 6.3.1 we would need to know the true variance of the sequence, which is unrealistic in practice. However, the following result states that we can substitute the true variance with the sample variance. The proof is beyond the scope of these notes.

Theorem 9.4.4 (Central limit theorem with sample standard deviation)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that its variance and fourth moment $\mathrm{E}(\widetilde{X}(i)^4)$ are bounded. The sequence*

$$\frac{\sqrt{n}\left(\operatorname{av}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)-\mu\right)}{\operatorname{std}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)} \tag{9.26}$$

converges in distribution to a standard Gaussian random variable.

Recall that the cdf of a standard Gaussian does not have a closed-form expression. To simplify notation we express the confidence interval in terms of the Q function.

Definition 9.4.5 (Q function)**.** *$Q(x)$ is the probability that a standard Gaussian random variable is greater than x for positive x,*

$$Q(x) := \int_{u=x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du, \quad x > 0. \tag{9.27}$$

By symmetry, if U is a standard Gaussian random variable and $y < 0$

$$\mathrm{P}(U < y) = Q(-y). \tag{9.28}$$

Corollary 9.4.6 (Approximate confidence interval for the mean)**.** *Let $\widetilde{X}$ be an iid sequence that satisfies the conditions of Theorem 9.4.4. For any $0 < \alpha < 1$*

$$\mathcal{I}_n := \left[Y_n - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y_n + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right], \tag{9.29}$$

$$Y_n := \operatorname{av}\left(\widetilde{X}(1), \widetilde{X}(2), \ldots, \widetilde{X}(n)\right), \tag{9.30}$$

$$S_n := \operatorname{std}\left(\widetilde{X}(1), \widetilde{X}(2), \ldots, \widetilde{X}(n)\right), \tag{9.31}$$

is an approximate $1-\alpha$ confidence interval for μ, i.e.

$$P(\mu \in \mathcal{I}_n) \approx 1 - \alpha. \tag{9.32}$$

Proof. By the central limit theorem, when $n \to \infty$ $\bar{X}_n$ is distributed as a Gaussian random variable with mean μ and variance σ^2. As a result

$$\begin{aligned}
P(\mu \in \mathcal{I}_n) &= 1 - P\left(Y_n > \mu + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(Y_n < \mu - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) && (9.33)\\
&= 1 - P\left(\frac{\sqrt{n}(Y_n - \mu)}{S_n} > Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(\frac{\sqrt{n}(Y_n - \mu)}{S_n} < -Q^{-1}\left(\frac{\alpha}{2}\right)\right) && (9.34)\\
&\approx 1 - 2Q\left(Q^{-1}\left(\frac{\alpha}{2}\right)\right) \quad \text{by Theorem 9.4.4} && (9.35)\\
&= 1 - \alpha. && (9.36)
\end{aligned}$$

□

It is important to stress that the result only provides an accurate confidence interval if n is large enough for the sample variance to converge to the true variance and for the CLT to take effect.

Example 9.4.7 (Bears in Yosemite (continued))**.** The sample standard deviation of the bears captured by the scientist equals 100 lbs. We apply Corollary 9.4.6 to derive an approximate confidence interval that is tighter than the one obtained applying Chebyshev's inequality. Given that $Q(1.95) \approx 0.025$,

$$\left[Y - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right) \right] \approx [188.8, 211.3] \tag{9.37}$$

is an approximate 95% confidence interval for the mean weight of the population of bears.

△

Interpreting confidence intervals is somewhat tricky. After computing the confidence interval in Example 9.4.7 one is tempted to state:

The probability that the average weight is between 188.8 and 211.3 lbs is 0.95.

However we are modeling the average weight as a deterministic quantity, so there are no random quantities in this statement! The correct interpretation is that if we repeat the process of sampling the population and compute the confidence interval many times, then the true value will lie in the interval 95% of the time. This is illustrated in the following example and Figure 9.6.

Example 9.4.8 (Estimating the average height (continued))**.** Figure 9.6 shows several 95% confidence intervals for the average of the height population in Example 9.3.3. To compute each interval we select n individuals and then apply Corollary 9.4.6. The width of the intervals decreases as n grows, but because they are all 95% confidence intervals they all contain the true average with probability 0.95. Indeed this is the case for 113 out of 120 (94%) of the intervals that are plotted.

△

9.5 Nonparametric model estimation

In this section we consider the problem of estimating a distribution from multiple iid samples. This requires approximating the cdf, pmf or pdf of the distribution. If we assume that the distribution belongs to a predefined family, then the problem reduces to estimating the parameters that characterize that particular family, as we explain in detail in Section 9.6. Here we do not make such an assumption. Estimating a distribution directly is very challenging; clearly many (infinite!) different distributions could have generated the data. However with enough samples it is often possible to obtain models that produce an accurate approximation, as long as the iid assumption holds.

9.5.1 Empirical cdf

Under the assumption that a data set corresponds to iid samples from a certain distribution, a reasonable estimate for the cdf of the distribution at a given point x is the fraction of samples that are smaller than x. This results in a piecewise constant estimator known as the empirical cdf.

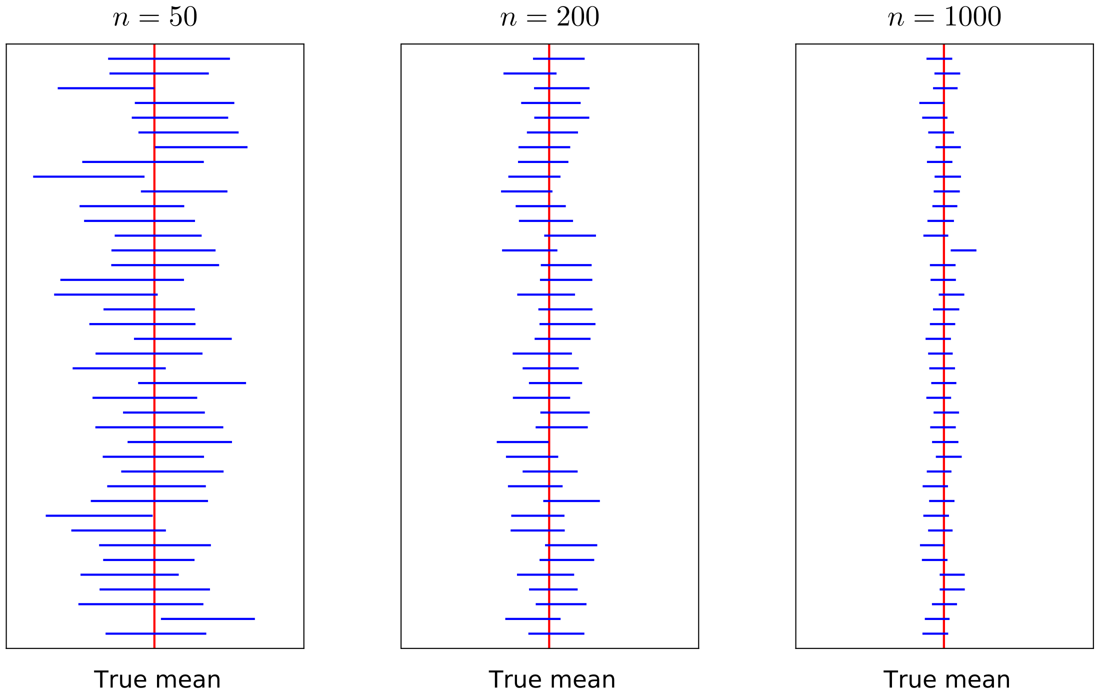

Figure 9.6: 95% confidence intervals for the average of the height population in Example 9.3.3.

Definition 9.5.1 (Empirical cdf)**.** *The empirical cdf corresponding to data $x_1, \ldots, x_n$ is*

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \le x}, \tag{9.38}$$

where $x \in \mathbb{R}$.

The empirical cdf is an unbiased and consistent estimator of the true cdf. This is established rigorously in Theorem 9.5.2 below and illustrated empirically in Figure 9.7. The cdf of the height data from 25,000 people is compared to three realizations of the empirical cdf computed from different numbers of iid samples. As the number of available samples grows, the approximation becomes very accurate.

Theorem 9.5.2. *Let $\widetilde{X}$ be an iid sequence with marginal cdf F_X. For any fixed $x \in \mathbb{R}$ $\widehat{F}_n(x)$ is an unbiased and consistent estimator of $F_X(x)$. In fact, $\widehat{F}_n(x)$ converges in mean square to $F_X(x)$.*

Proof. First, we verify

$$\begin{aligned}
\mathrm{E}\left(\widehat{F}_n(x)\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\widetilde{X}(i) \le x}\right) && (9.39)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}\left(\widetilde{X}(i) \le x\right) \quad \text{by linearity of expectation} && (9.40)\\
&= F_X(x), && (9.41)
\end{aligned}$$

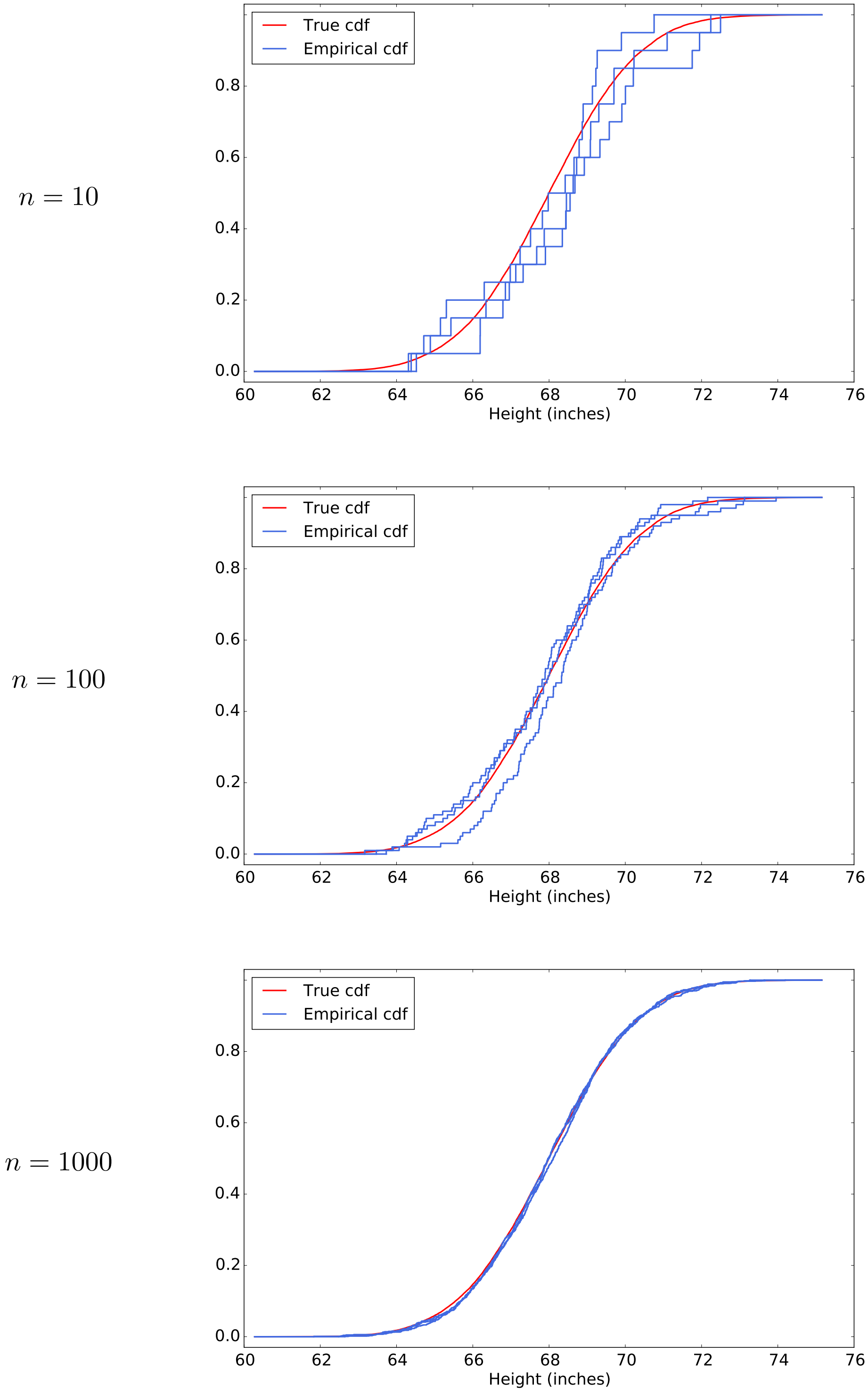

Figure 9.7: Cdf of the height data in Figure 2.13 along with three realizations of the empirical cdf computed with n iid samples for $n = 10, 100, 1000$.

so the estimator is unbiased. We now estimate its mean square

$$\mathrm{E}\left(\widehat{F}_n^2(x)\right) = \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1_{\widetilde{X}(i)\le x} 1_{\widetilde{X}(j)\le x}\right) \tag{9.42}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{P}\left(\widetilde{X}(i)\le x\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \mathrm{P}\left(\widetilde{X}(i)\le x, \widetilde{X}(j)\le x\right) \tag{9.43}$$

$$= \frac{F_X(x)}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} F_{\widetilde{X}(i)}(x)\, F_{\widetilde{X}(j)}(x) \quad \text{by independence} \tag{9.44}$$

$$= \frac{F_X(x)}{n} + \frac{n-1}{n} F_X^2(x). \tag{9.45}$$

The variance is consequently equal to

$$\mathrm{Var}\left(\widehat{F}_n(x)\right) = \mathrm{E}\left(\widehat{F}_n(x)^2\right) - \mathrm{E}^2\left(\widehat{F}_n(x)\right) \tag{9.46}$$

$$= \frac{F_X(x)(1 - F_X(x))}{n}. \tag{9.47}$$

We conclude that

$$\lim_{n\to\infty} \mathrm{E}\left(\left(F_X(x) - \widehat{F}_n(x)\right)^2\right) = \lim_{n\to\infty} \mathrm{Var}\left(\widehat{F}_n(x)\right) = 0. \tag{9.48}$$

□

9.5.2 Density estimation

Estimating the pdf of a continuous quantity is much more challenging that estimating the cdf. If we have sufficient data, the fraction of samples that are smaller than a certain x provide a good estimate for the cdf at that point. However, no matter how much data we have, there is negligible probability that we will see any samples exactly at x: a pointwise empirical density estimator would equal zero almost everywhere (except at the available samples).

Our only hope to produce an accurate estimator is if the pdf that we aim to estimate is smooth. In that case, we can estimate its value at a point x from observed samples that are situated at neighboring locations. If there are many samples close to x then this suggests that the estimate at x should be large, whereas if all the samples are far away, then it should be small. **Kernel density estimation** achieves this by averaging the samples.

Definition 9.5.3 (Kernel density estimator)**.** *The kernel density estimate with bandwidth h of the distribution of $x_1, \ldots, x_n$ at $x \in \mathbb{R}$ is*

$$\widehat{f}_{h,n}(x) := \frac{1}{n\,h}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right), \tag{9.49}$$

where k is a kernel function centered at the origin that satisfies

$$k(x) \ge 0 \quad \textit{for all } x \in \mathbb{R}, \tag{9.50}$$

$$\int_{\mathbb{R}} k(x)\; dx = 1. \tag{9.51}$$

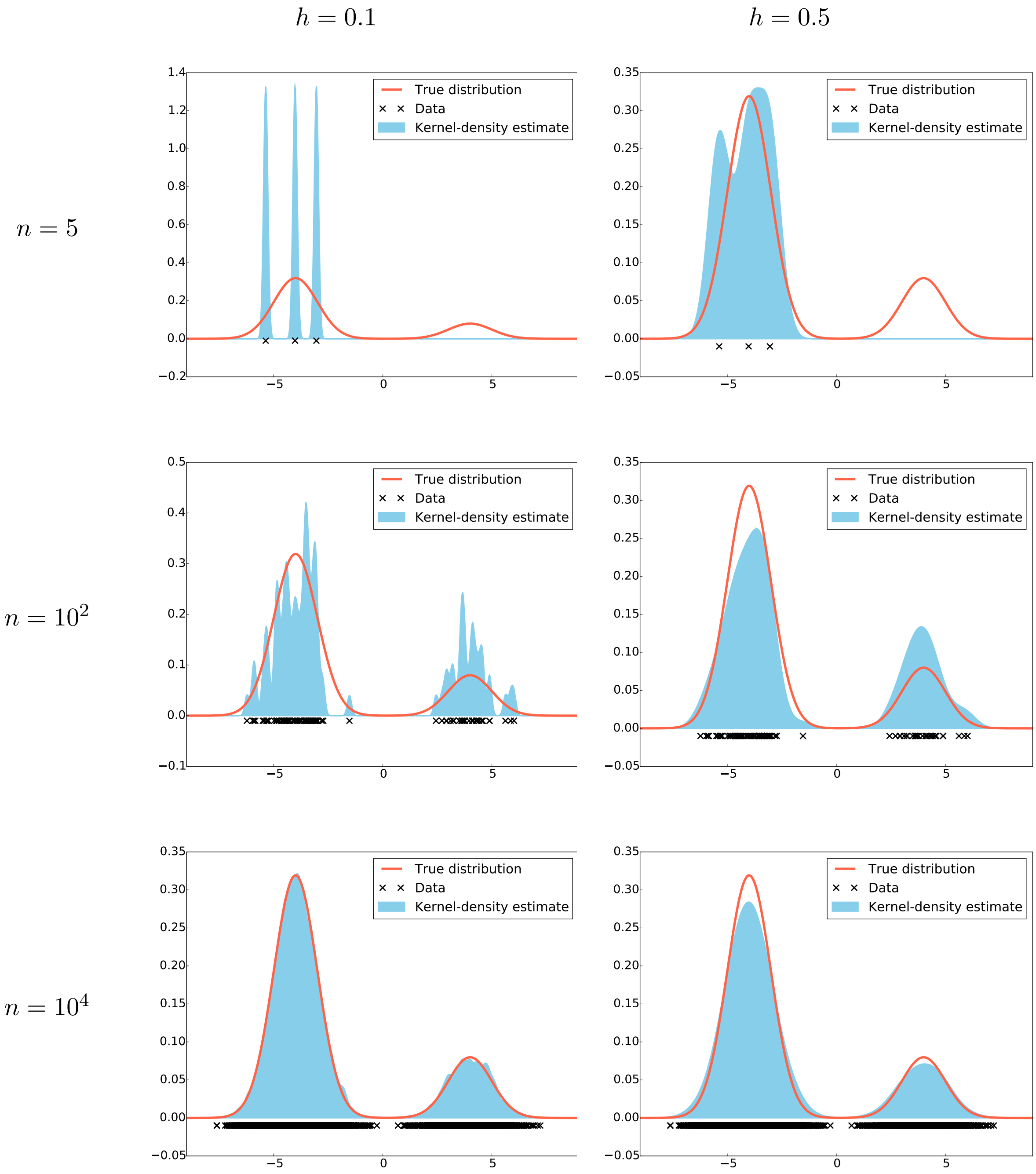

Figure 9.8: Kernel density estimation for the Gaussian mixture described in Example 9.6.5 for different number of iid samples and different values of the kernel bandwidth h.

The effect of the kernel is to weight each sample according to their distance to the point at which we are estimating the pdf x. Choosing a rectangular kernel yields an empirical density estimate that is piecewise constant and roughly looks like a histogram (the corresponding weights are constant or equal to zero). A popular alternative is the Gaussian kernel $k(x) = \exp\left(-x^2\right)/\sqrt{\pi}$, which produces a smooth density estimate. The kernel should decay so that $k\left((x - x_i)/h\right)$ is large when the sample x_i is close to x and small when it is far. This decay is governed by the bandwidth h, which is chosen before hand based on our expectations about the smoothness of the pdf and on the amount of available data. If the bandwidth is very small, individual samples have a large influence on the density estimate. This allows to reproduce irregular shapes more easily, but also yields spurious fluctuations that are not present in the true curve, especially if we don't have a lot of samples. Increasing the bandwidth smooths out such fluctuations and yields more stable estimates when the number of data is small. However, it may also over-smooth the estimate. As a rule of thumb, we should decrease the bandwidth of the kernel as the number of data increases.

Figures 9.8 and 9.9 illustrate the effect of varying the bandwidth h at different sampling rates. In Figure 9.8 Gaussian kernel density estimation is applied to estimate the Gaussian mixture described in Example 9.6.5. Figure 9.9 shows an example where the same technique is used on real data: the aim is to estimate the density of the weight of a sea-snail population.[2] The whole population consists of 4,177 individuals. The kernel density estimate is computed from 200 iid samples for different values of the kernel bandwidth.

9.6 Parametric model estimation

In the previous section, we describe how to estimate a distribution by directly estimating the cdf or pdf generating the data. In this section, we discuss an alternative route based on the assumption that the type of distribution generating the data is known beforehand. If this is the case, the problem boils down to fitting the parameters characterizing the distribution to the data. Recall that from a frequentist viewpoint, the true distribution is fixed, so the corresponding parameters are modeled as deterministic quantities (in contrast, in a Bayesian framework they are modeled as random variables).

9.6.1 The method of moments

The method of moments adjusts the parameters of a distribution so that the moments of the distribution coincide with the sample moments of the data (i.e. its mean, mean square or variance, etc.). If the distribution only depends on one parameter, then we use the sample mean as a surrogate for the true mean and compute the corresponding value of the parameter. For an exponential with parameter λ and mean μ we have

$$\mu = \frac{1}{\lambda}. \tag{9.52}$$

Assuming that we have access to n iid samples $x_1, \ldots, x_n$ from the exponential distribution, the method-of-moments estimate of λ equals

$$\lambda_{\text{MM}} := \frac{1}{\operatorname{av}\left(x_1, \ldots, x_n\right)}. \tag{9.53}$$

[2] The data are available at archive.ics.uci.edu/ml/datasets/Abalone

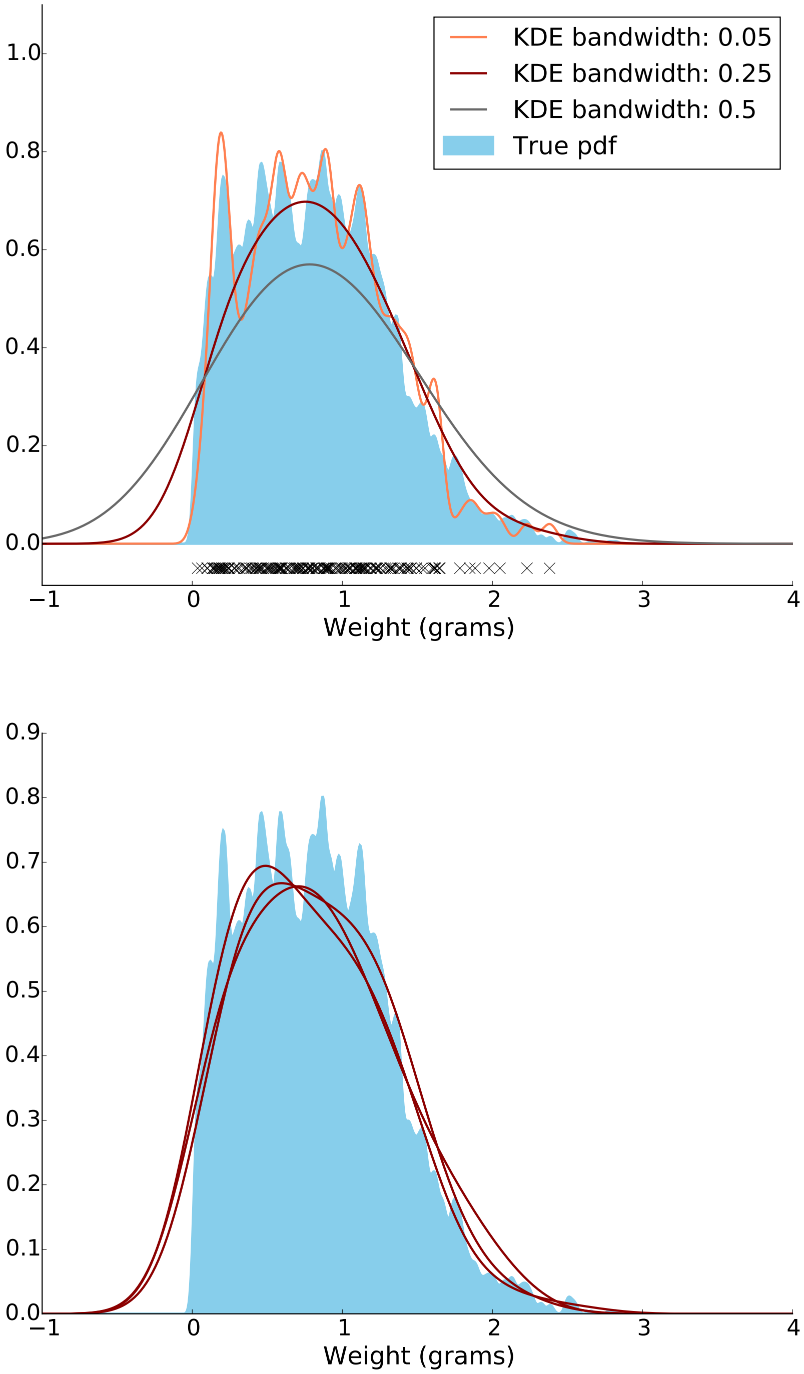

Figure 9.9: Kernel density estimate for the weight of a population of abalone, a species of sea snail. In the plot above the density is estimated from 200 iid samples using a Gaussian kernel with three different bandwidths. Black crosses representing the individual samples are shown underneath. In the plot below we see the result of repeating the procedure three times using a fixed bandwidth equal to 0.25.

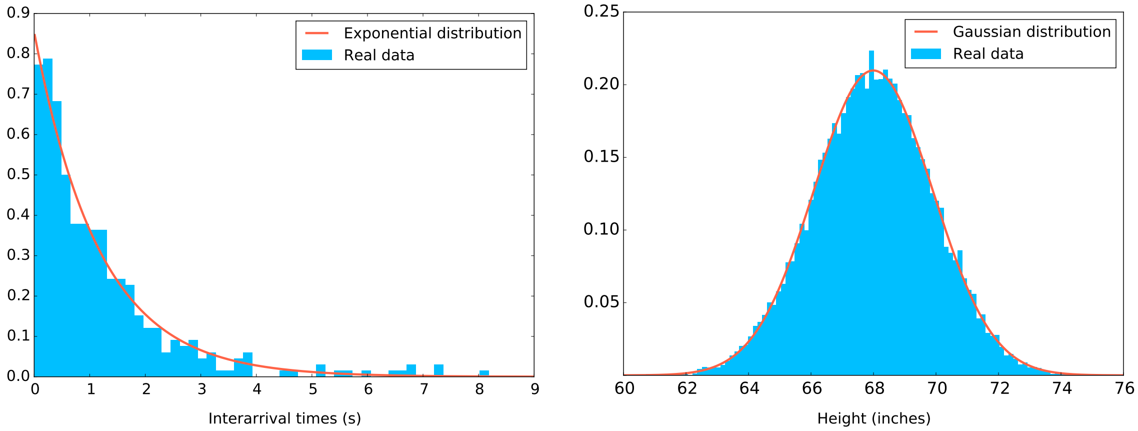

Figure 9.10: Exponential distribution fitted to data consisting of inter-arrival times of calls at a call center in Israel (left). Gaussian distribution fitted to height data (right).

The graph on the right of Figure 9.10 shows the result of fitting an exponential to the call-center data in Figure 2.11. Similarly, to fit a Gaussian using the method of moments we set the mean equal to its sample mean and the variance equal to the sample variance, as illustrated by the graph on the right of Figure 9.10 using the data from Figure 2.13.

9.6.2 Maximum likelihood

The most popular method for learning parametric models is maximum-likelihood fitting. The **likelihood** function is the joint pmf or pdf of the data, interpreted as a *function of the unknown parameters*. In more detail, let us denote the data by $x_1, \ldots, x_n$ and assume that they are realizations of a set of discrete random variables $X_1, \ldots, X_n$ which have a joint pmf that depends on a vector of parameters $\vec{\theta}$. To emphasize that the joint pmf depends on $\vec{\theta}$ we denote it by $p_{\vec{\theta}} := p_{X_1,\ldots,X_n}$. This pmf evaluated at the observed data

$$p_{\vec{\theta}}\left(x_1, \ldots, x_n\right) \tag{9.54}$$

is the likelihood function, when we interpret it as a function of $\vec{\theta}$. For continuous random variables, we use the joint pdf of the data instead.

Definition 9.6.1 (Likelihood function)**.** *Given a realization $x_1, \ldots, x_n$ of a set of discrete random variables $X_1, \ldots, X_n$ with joint pmf $p_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$ is a vector of parameters, the likelihood function is*

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := p_{\vec{\theta}}\left(x_1, \ldots, x_n\right). \tag{9.55}$$

If the random variables are continuous with pdf $f_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$, the likelihood function is

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := f_{\vec{\theta}}\left(x_1, \ldots, x_n\right). \tag{9.56}$$

*The **log-likelihood function** is equal to the logarithm of the likelihood function* $\log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right)$.

When the data are modeled as iid samples, the likelihood factors into a product of the marginal pmf or pdf, so the log likelihood can be decomposed into a sum.

In the case of discrete distributions, for a fixed $\vec{\theta}$ the likelihood is the probability that $X_1, \ldots, X_n$ equal the observed data. If we don't know $\vec{\theta}$, it makes sense to choose a value for $\vec{\theta}$ such that this probability is as high as possible, i.e. to maximize the likelihood. For continuous distributions we apply the same principle to the joint pdf of the data.

Definition 9.6.2 (Maximum-likelihood estimator)**.** *The* ***maximum likelihood*** *(ML) estimator for the vector of parameters $\vec{\theta} \in \mathbb{R}^m$ is*

$$\vec{\theta}_{\text{ML}}\left(x_1, \ldots, x_n\right) := \arg\max_{\vec{\theta}} \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) \tag{9.57}$$

$$= \arg\max_{\vec{\theta}} \log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right). \tag{9.58}$$

The maximum of the likelihood function and that of the log-likelihood function are at the same location because the logarithm is monotone.

Under certain conditions, one can show that the maximum-likelihood estimator is consistent: it converges in probability to the true parameter as the number of data increases. One can also show that its distribution converges to that of a Gaussian random variable (or vector), just like the distribution of the sample mean. These results are beyond the scope of the course. Bear in mind, however, that they only hold if the data are indeed generated by the type of distribution that we are considering.

We now show how to derive the maximum-likelihood for a Bernoulli and a Gaussian distribution. The resulting estimators for the parameters are the same as the method-of-moments estimators (except for a slight difference in the estimate of the Gaussian variance parameter).

Example 9.6.3 (ML estimator of the parameter of a Bernoulli distribution)**.** We model a set of data $x_1, \ldots, x_n$ as iid samples from a Bernoulli distribution with parameter θ (in this case there is only one parameter). The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}(\theta) = p_\theta\left(x_1, \ldots, x_n\right) \tag{9.59}$$

$$= \prod_{i=1} \left(1_{x_i=1}\theta + 1_{x_i=0}\left(1-\theta\right)\right) \tag{9.60}$$

$$= \theta^{n_1}\left(1-\theta\right)^{n_0} \tag{9.61}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\theta) = n_1 \log\theta + n_0 \log\left(1-\theta\right), \tag{9.62}$$

where n_1 are the number of samples equal to one and n_0 the number of samples equal to zero. The ML estimator of the parameter θ is

$$\theta_{\text{ML}} = \arg\max_{\theta} \log \mathcal{L}_{x_1,\ldots,x_n}(\theta) \tag{9.63}$$

$$= \arg\max_{\theta} n_1 \log\theta + n_0 \log\left(1-\theta\right). \tag{9.64}$$

We compute the derivative and second derivative of the log-likelihood function,

$$\frac{\mathrm{d} \log \mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta} = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}, \tag{9.65}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta^2} = -\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} < 0. \tag{9.66}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\theta_{\mathrm{ML}} = \frac{n_1}{n_0 + n_1}. \tag{9.67}$$

The estimate is equal to the fraction of samples that are equal to one.

△

Example 9.6.4 (ML estimator of the parameters of a Gaussian distribution)**.** Let $x_1, x_2, \ldots$ be data that we wish to model as iid samples from a Gaussian distribution with mean μ and standard deviation σ. The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.68}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.69}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = -\frac{n \log(2\pi)}{2} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.70}$$

The ML estimator of the parameters μ and σ is

$$\{\mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}}\} = \arg\max_{\{\mu,\sigma\}} \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) \tag{9.71}$$

$$= \arg\max_{\{\mu,\sigma\}} -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.72}$$

We compute the partial derivatives of the log-likelihood function,

$$\frac{\partial \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial \mu} = -\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}, \tag{9.73}$$

$$\frac{\partial \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3}. \tag{9.74}$$

The function we are trying to maximize is strictly concave in $\{\mu, \sigma\}$. To prove this, we would have to show that the Hessian of the function is positive definite. We omit the calculations that show that this is the case. Setting the partial derivatives to zero we obtain

$$\mu_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{9.75}$$

$$\sigma_{\mathrm{ML}}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})^2. \tag{9.76}$$

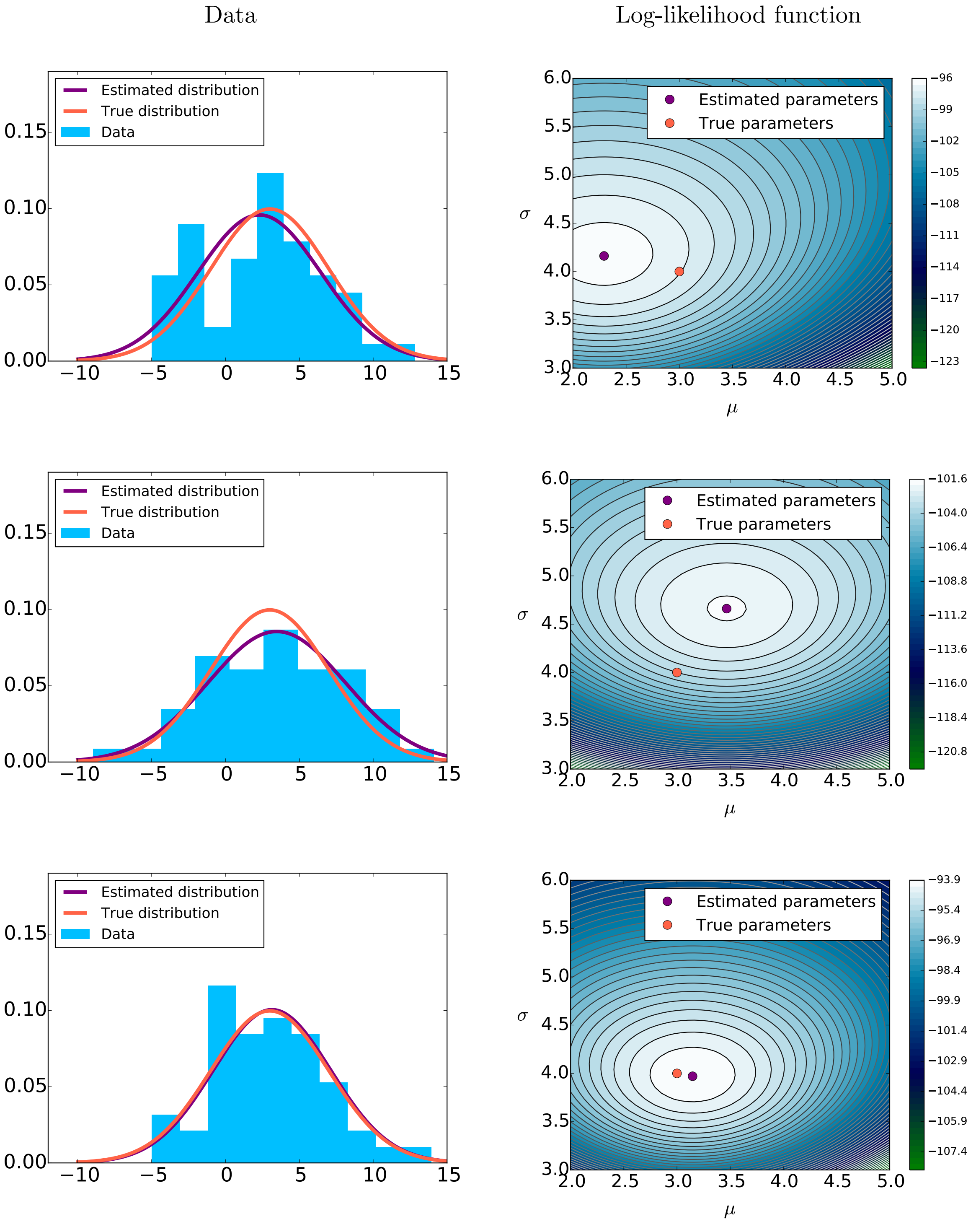

Figure 9.11: The left column shows histograms of 50 iid samples from a Gaussian distribution, together with the pdf of the original distribution, as well as the maximum-likelihood estimate. The right column shows the log-likelihood function corresponding to the data and the location of its maximum and of the point corresponding to the true parameters.

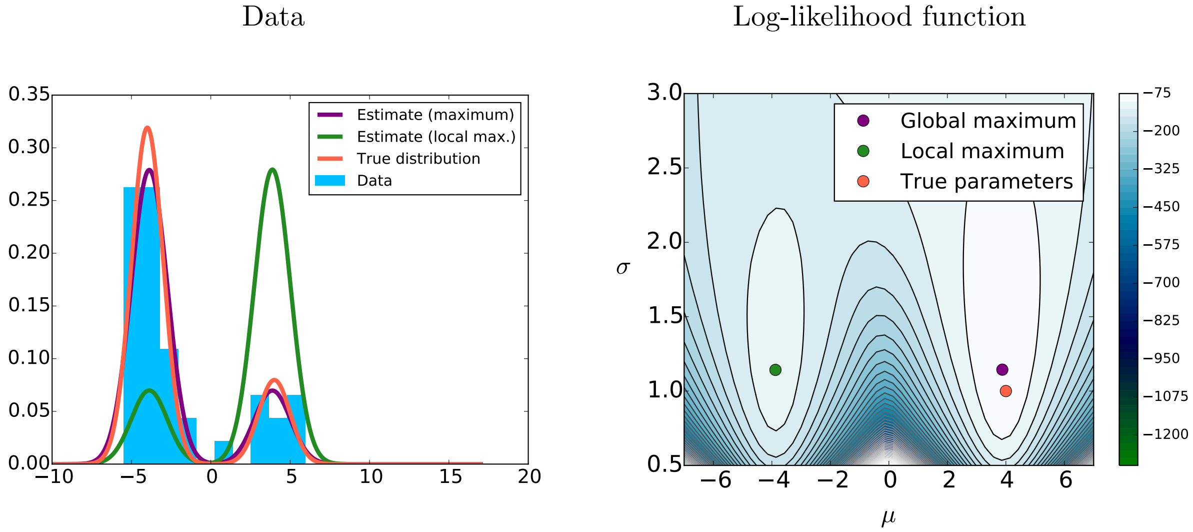

Figure 9.12: The left image shows a histogram of 40 iid samples from the Gaussian mixture defined in Example 9.6.5, together with the pdf of the original distribution. The right image shows the log-likelihood function corresponding to the data, which has a local maximum apart from the global maximum. The density estimates corresponding to the two maxima are shown on the left.

The estimator for the mean is just the sample mean. The estimator for the variance is a rescaled sample variance.

$\triangle$

Figure 9.11 displays the log-likelihood function corresponding to 50 iid samples from a Gaussian distribution with $\mu := 3$ and $\sigma := 4$. It also shows the approximation to the true pdf obtained by maximum likelihood. In Examples 9.6.3 and 9.6.4 the log-likelihood function is strictly concave. This means that the function has a unique maximum that can be located by setting the gradient to zero. When this yields nonlinear equations that cannot be solved directly, we can leverage optimization methods such as gradient ascent that will converge to the maximum. However, the log-likelihood function is not always concave. As illustrated by the following example, in such cases it can have multiple local maxima, which may make it intractable to compute the maximum-likelihood estimator.

Example 9.6.5 (Log-likelihood function of a Gaussian mixture)**.** Let X be a Gaussian mixture defined as

$$X := \begin{cases} G_1 & \text{with probability } \frac{1}{5}, \\ G_2 & \text{with probability } \frac{4}{5}, \end{cases} \tag{9.77}$$

where G_1 is a Gaussian random variable with mean $-\mu$ and variance σ^2, whereas G_2 is also Gaussian with mean μ and variance σ^2. We have parameterized the mixture with just two parameters so that we can visualize the log-likelihood in two dimensions. Let $x_1, x_2, \ldots$ be data modeled as iid samples from X. The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.78}$$

$$= \prod_{i=1}^{n} \frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.79}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = \sum_{i=1}^{n} \log\left(\frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right). \tag{9.80}$$

Figure 9.12 shows the log-likelihood function for 40 iid samples of the distribution when $\mu := 4$ and $\sigma := 1$. The function has a local maximum away from the global maximum. This means that if we use a local ascent method to find the ML estimator, we might not find the global maximum, but remain stuck at the local maximum instead. The estimate corresponding to the local maximum (shown on the left) has the same variance as the global maximum but μ is close to -4 instead of 4. Although the estimate doesn't fit the data very well, it is locally optimal, small shifts of μ and σ yield worse fits (in terms of the likelihood).

△

To finish this section, we describe a machine-learning algorithm for supervised learning based on parametric fitting using ML estimation.

Example 9.6.6 (Quadratic discriminant analysis)**.** Quadratic discriminant analysis is an algorithm for supervised learning. The input to the algorithm are two sets of training data, consisting of d-dimensional vectors $\vec{a}_1,\ldots,\vec{a}_n$ and $\vec{b}_1,\ldots,\vec{b}_n$ which belong to two different classes (the method can easily be extended to deal with more classes). The goal is to classify new instances based on the structure of the data.

To perform quadratic discriminant analysis we first fit a d-dimensional Gaussian distribution to the data of each class using the ML estimator for the mean and covariance matrix, which correspond to the sample mean and covariance matrix of the training data (up to a slight rescaling of the sample covariance). In more detail, $\vec{a}_1,\ldots,\vec{a}_n$ are used to estimate a mean $\vec{\mu}_a$ and covariance matrix Σ_a, whereas $\vec{b}_1,\ldots,\vec{b}_n$ are used to estimate $\vec{\mu}_b$ and Σ_b,

$$\{\vec{\mu}_a, \Sigma_a\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{a}_1,\ldots,\vec{a}_n}(\vec{\mu},\Sigma), \tag{9.81}$$

$$\{\vec{\mu}_b, \Sigma_b\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{b}_1,\ldots,\vec{b}_n}(\vec{\mu},\Sigma). \tag{9.82}$$

Then for each new example $\vec{x}$, the value of the density function at the example for both classes is evaluated. If

$$f_{\vec{\mu}_a,\Sigma_a}(\vec{x}) > f_{\vec{\mu}_b,\Sigma_b}(\vec{x}) \tag{9.83}$$

then $\vec{x}$ is declared to belong to the first class, otherwise it is declared to belong to the second class. Figure 9.13 shows the results of applying the method to data simulated using two Gaussian distributions.

△

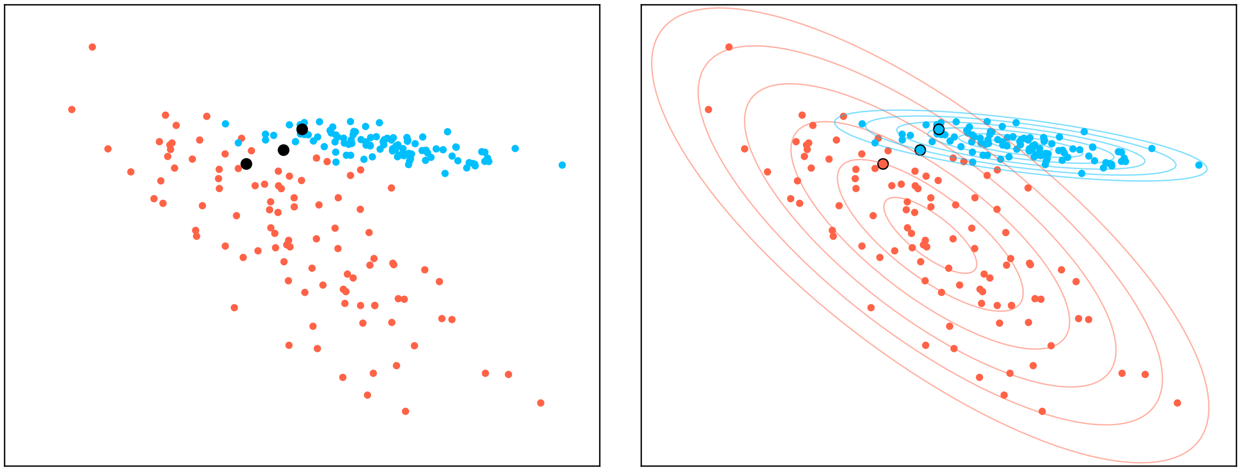

Figure 9.13: Quadratic-discriminant analysis applied to data from two different classes (left). The data corresponding to the two different classes are colored orange and blue. Three new examples are colored in black. Two bivariate Gaussians are fit to the data. Their contour lines are shown in the respective color of each class on the right. These distributions are used to classify the new examples, which are colored according to their estimated class.

9.7 Proofs

9.7.1 Proof of Lemma 9.2.5

We consider the sample variance of an iid sequence $\widetilde{X}$ with mean μ and variance σ^2,

$$\widetilde{Y}(n) := \frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{1}\widetilde{X}(j)\right) \tag{9.84}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{n}\widetilde{X}(j)\right)^2 \tag{9.85}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i)^2 + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\widetilde{X}(j)\,\widetilde{X}(k) - \frac{2}{n}\sum_{j=1}^{n}\widetilde{X}(i)\,\widetilde{X}(j)\right) \tag{9.86}$$

To simplify notation, we denote the mean square $\mathrm{E}\left(\widetilde{X}(i)^2\right) = \mu^2 + \sigma^2$ by ξ. We have

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left(\widetilde{X}(j)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathrm{E}\left(\widetilde{X}(j)\,\widetilde{X}(k)\right) \tag{9.87}$$

$$-\frac{2}{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) - \frac{2}{n}\sum_{\substack{j=1\\j\neq i}}^{n}\mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) \tag{9.88}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\xi + \frac{n\,\xi}{n^2} + \frac{n(n-1)\mu^2}{n^2} - \frac{2\,\xi}{n} - \frac{2(n-1)\mu^2}{n} \tag{9.89}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\left(\xi - \mu^2\right) \tag{9.90}$$

$$= \sigma^2. \tag{9.91}$$

9.7.2 Proof of Theorem 9.3.4

We denote the sample median by $\widetilde{Y}(n)$. Our aim is to show that for any $\epsilon > 0$

$$\lim_{n\to\infty}\mathrm{P}\left(\left|\widetilde{Y}(n) - \gamma\right| \geq \epsilon\right) = 0. \tag{9.92}$$

We will prove that

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \geq \gamma + \epsilon\right) = 0. \tag{9.93}$$

The same argument allows to establish

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \leq \gamma - \epsilon\right) = 0. \tag{9.94}$$

If we order the set $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$, then $\widetilde{Y}(n)$ equals the $(n+1)/2$th element if n is odd and the average of the $n/2$th and the $(n/2+1)$th element if n is even. The event $\widetilde{Y}(n) \geq \gamma + \epsilon$ therefore implies that at least $(n+1)/2$ of the elements are larger than $\gamma + \epsilon$.

For each individual $\widetilde{X}(i)$, the probability that $\widetilde{X}(i) > \gamma + \epsilon$ is

$$p := 1 - F_{\widetilde{X}(i)}(\gamma + \epsilon) = 1/2 - \epsilon' \tag{9.95}$$

where we assume that $\epsilon' > 0$. If this is not the case then the cdf of the iid sequence is *flat* at γ and the median is not well defined. The number of random variables in the set $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$ which are larger than $\gamma + \epsilon$ is distributed as a binomial random variable B_n with parameters n

and p. As a result, we have

$$\begin{aligned}
\mathrm{P}\left(\widetilde{Y}\left(n\right) \geq \gamma+\epsilon\right) &\leq \mathrm{P}\left(\frac{n+1}{2} \text{ or more samples are greater or equal to } \gamma+\epsilon\right) && (9.96)\\
&= \mathrm{P}\left(B_n \geq \frac{n+1}{2}\right) && (9.97)\\
&= \mathrm{P}\left(B_n - np \geq \frac{n+1}{2} - np\right) && (9.98)\\
&\leq \mathrm{P}\left(\left|B_n - np\right| \geq n\epsilon' + \frac{1}{2}\right) && (9.99)\\
&\leq \frac{\mathrm{Var}\left(B_n\right)}{\left(n\epsilon' + \frac{1}{2}\right)^2} \quad \text{by Chebyshev's inequality} && (9.100)\\
&= \frac{np\left(1-p\right)}{n^2\left(\epsilon' + \frac{1}{2n}\right)^2} && (9.101)\\
&= \frac{p\left(1-p\right)}{n\left(\epsilon' + \frac{1}{2n}\right)^2}, && (9.102)
\end{aligned}$$

which converges to zero as $n \to \infty$. This establishes (9.93).