DS-GA 1002 Lecture notes 5 October 19, 2015

Statistics: Learning models from data

Learning models from data that are assumed to be generated probabilistically from a certain unknown distribution is a crucial step in statistical analysis. The model may be of interest in itself, as a description of the data used to build it, or it may be used to perform inference, i.e. to extract conclusions about new data.

1 Parametric models

In **parametric** modeling we make the assumption that the data are sampled from a **known** family of distributions (Gaussian, Bernoulli, exponential, ...) with a small number of **unknown** parameters that must be learnt from the data. This assumption may be motivated by theoretical insights such as the Central Limit Theorem, which explains why additive disturbances are often well modeled as Gaussian, or by empirical evidence.

1.1 Frequentist parameter selection

Frequentist statistical analysis is based on treating parameters as unknown **deterministic** quantities. These parameters are fit to produce a model that is well adapted to the available data. In order to quantify the extent to which a model explains the data, we define the **likelihood** function. This function is the pmf or pdf of the distribution that generates the data, interpreted as a *function of the unknown parameters.*

Definition 1.1 (Likelihood function)**.** *Given a realization $\boldsymbol{x}$ of an iid vector of random variables $\boldsymbol{X}$ of dimension n with a distribution that depends on a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^m$, the likelihood function is defined as*

$$\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}) := \prod_{i=1}^{n} p_{X_i}(x_i, \boldsymbol{\theta}) \tag{1}$$

if $\boldsymbol{X}$ is discrete and

$$\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}) := \prod_{i=1}^{n} f_{X_i}(x_i, \boldsymbol{\theta}) \tag{2}$$

if $\boldsymbol{X}$ is continuous.

The ***log-likelihood function*** *is equal to the logarithm of the likelihood function* $\log \mathcal{L}_{\boldsymbol{x}}(\theta)$.

The likelihood function represents the probability or the probability density of the parametric distribution at the observed data, i.e. it quantifies how *likely* the data are according to the model. Therefore, higher likelihood values indicate that the model is better adapted to the samples. The **maximum likelihood** (ML) estimator is a hugely popular parameter estimator based on maximizing the likelihood (or equivalently the log-likelihood).

Definition 1.2 (Maximum-likelihood estimator)**.** *The* ***maximum likelihood*** *(ML) estimator for the vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^m$ is*

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{ML}}\left(\boldsymbol{x}\right) &:= \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{\theta}\right) && (3) \\ &= \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{\theta}\right). && (4) \end{aligned}$$

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 even 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 obviously only hold if the data are indeed generated by the type of distribution that we are considering.

Example 1.3 (ML estimator of the parameter of a Bernoulli distribution)**.** Let $x_1, x_2, \ldots$ be data that we wish to model as iid samples from a Bernoulli distribution. The likelihood function is equal to

$$\begin{aligned} \mathcal{L}_{\mathbf{x}}\left(p\right) &= \prod_{i=1}^{n} p_{X_i}\left(x_i, p\right) && (5) \\ &= \prod_{i=1} 1_{x_i=1} p + 1_{x_i=0}\left(1-p\right) && (6) \\ &= p^{n_1}\left(1-p\right)^{n_0} && (7) \end{aligned}$$

and the log-likelihood function to

$$\begin{aligned} \log \mathcal{L}_{\mathbf{x}}\left(p\right) &= \sum_{i=1}^{n} p_{X_i}\left(x_i\right) && (8) \\ &= n_1 \log p + n_0 \log\left(1-p\right), && (9) \end{aligned}$$

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 p is

$$\begin{aligned} \hat{p}_{\text{ML}} &= \arg\max_{p} \log \mathcal{L}_{\mathbf{x}}\left(p\right) && (10) \\ &= \arg\max_{p} n_1 \log p + n_0 \log\left(1-p\right). && (11) \end{aligned}$$

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

$$\frac{\mathrm{d} \log \mathcal{L}_{\mathbf{x}}(p)}{\mathrm{d} p} = \frac{n_1}{p} - \frac{n_0}{1-p}, \tag{12}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{\mathbf{x}}(p)}{\mathrm{d} p^2} = -\frac{n_1}{p^2} - \frac{n_0}{(1-p)^2} < 0. \tag{13}$$

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

$$\hat{p}_{\mathrm{ML}} = \frac{n_1}{n_0 + n_1}, \tag{14}$$

the fraction of samples that are equal to one, which is a very reasonable estimate.

Example 1.4 (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

$$\mathcal{L}_{\mathbf{x}}(\mu, \sigma) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{15}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{16}$$

and the log-likelihood function to

$$\log \mathcal{L}_{\mathbf{x}}(\mu, \sigma) = \sum_{i=1}^{n} f_{X_i}(x_i, p) \tag{17}$$

$$= -\frac{n \log(2\pi)}{2} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}. \tag{18}$$

The ML estimator of the parameters μ and σ is

$$\{\hat{\mu}_{\mathrm{ML}}, \hat{\sigma}_{\mathrm{ML}}\} = \arg\max_{\{\mu,\sigma\}} \log \mathcal{L}_{\mathbf{x}}(\mu, \sigma) \tag{19}$$

$$= \arg\max_{\{\mu,\sigma\}} -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}. \tag{20}$$

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

$$\frac{\partial \log \mathcal{L}_{\mathbf{x}}\left(\mu,\sigma\right)}{\partial \mu} = -\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}, \tag{21}$$

$$\frac{\partial \log \mathcal{L}_{\mathbf{x}}\left(\mu,\sigma\right)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{\sigma^3}. \tag{22}$$

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

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{23}$$

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \hat{\mu}_{\text{ML}}\right). \tag{24}$$

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

Figure 1 shows the result of fitting a Gaussian to the height data in Figure 1 of Lecture Notes 4 by applying the ML estimators for the mean and the variance derived in Example 1.4. For a small number of samples the estimate can be quite unstable, but for a large number of samples it provides a good description of the data.

1.2 Bayesian parameter selection

Up to now we have focused on estimating parameters that are modeled as deterministic and fixed. This is the viewpoint of frequentist statistics. **Bayesian** statistics provide an alternative perspective in which the parameters are considered *random*. This allows for greater flexibility in both building the model and in quantifying our uncertainty about it, but it also assumes that we have access to an estimate of the distribution of the parameters.

In a frequentist framework we assumed only that the likelihood function of the data was known. Bayesian inference relies on two modeling choices:

1. The **prior** distribution of the parameters encodes our uncertainty about the model before seeing the data.

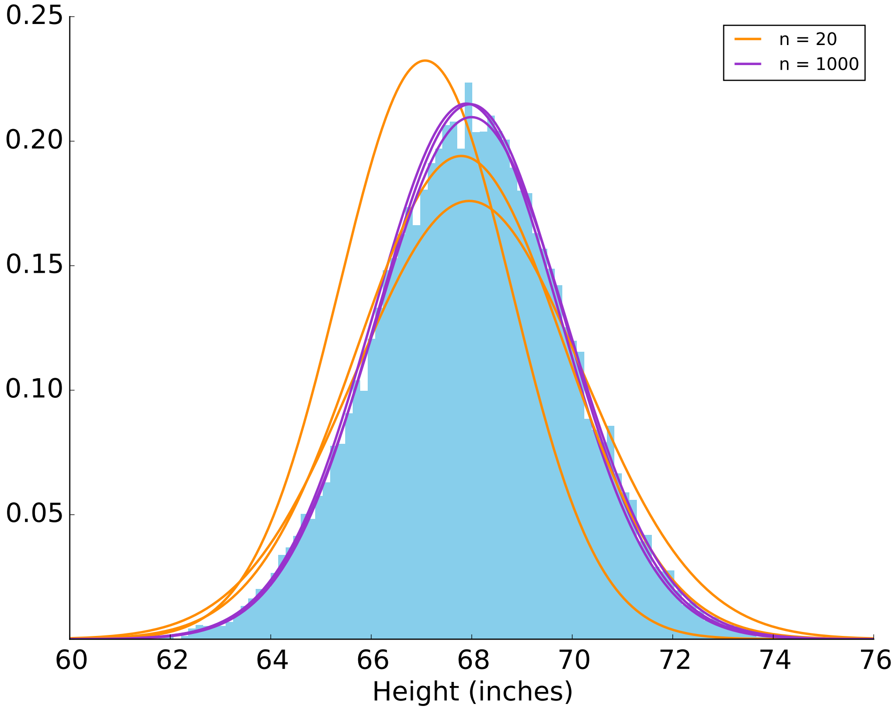

Figure 1: Result of fitting a Gaussian distribution to the data shown in Figure 1 of Lecture Notes 4 three times using 20/1000 random samples.

2. The conditional distribution of the data given the parameters specifies how the data are generated. The pmf or pdf that characterizes this distribution is equal to the **likelihood** function $\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})$ from Definition 1.1. In a Bayesian framework the likelihood is no longer just a function of the parameters; it has a probabilistic interpretation in its own right.

Once the prior and the likelihood are fixed, we apply Bayes theorem to obtain the **posterior distribution** of the parameters given the data. This distribution quantifies our uncertainty about the value of the parameters *after* processing the data.

Theorem 1.5 (Posterior distribution)**.** *The posterior distribution of the parameters given the data equals*

$$p_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\sum_{\boldsymbol{u}} p_{\boldsymbol{\Theta}}(\boldsymbol{u})\, p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{u})} \tag{25}$$

if the data and parameters are discrete,

$$f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\int_{\boldsymbol{u}} f_{\boldsymbol{\Theta}}(\boldsymbol{u})\, f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{u})\; d\boldsymbol{u}} \tag{26}$$

if the data and parameters are continuous,

$$p_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\sum_{\boldsymbol{u}} p_{\boldsymbol{\Theta}}(\boldsymbol{u})\, f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{u})} \tag{27}$$

if the data are continuous and the parameters discrete, and

$$f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\int_{\boldsymbol{u}} f_{\boldsymbol{\Theta}}(\boldsymbol{u})\, p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{u})\; d\boldsymbol{u}} \tag{28}$$

if the data are discrete and the parameters continuous.

Proof. The expressions follow from a direct application of Bayes theorem. □

The posterior distribution of the parameter given the data allows us to compute the probability that the parameter lies in a certain interval. Such intervals are called **credible intervals**, as opposed to frequentist confidence intervals. Recall that once we have computed a $1-\alpha$ confidence interval from the data, it makes no sense to state that it contains the true parameter with probability $1-\alpha$; the realization of the interval and the parameter are both deterministic. In contrast, once we have computed the posterior distribution of a parameter given the data within a Bayesian framework, it is completely correct to state that the true parameter belongs to the fixed $1-\alpha$ credible interval with probability $1-\alpha$ (if the prior and likelihood are assumed to be correct).

A question that remains is how to produce a point estimate of the parameters from their posterior distribution. A reasonable choice is the mean of the posterior distribution, which corresponds to the conditional expectation of the parameters given the data. This has a strong theoretical justification: the posterior mean minimizes the mean square error with respect to the true value of the parameters *over all possible estimators.*

Theorem 1.6 (The posterior mean minimizes the MSE)**.** *The posterior mean is the minimum mean-squared-error (MMSE) estimate of the parameter given the data. More precisely, if we represent the data and the parameters as the random vectors* $\boldsymbol{\Theta}$ *and* $\boldsymbol{X}$*,*

$$\mathrm{E}\big(\boldsymbol{\Theta}|\boldsymbol{X}\big) = \arg\min_{\hat{\boldsymbol{\theta}}(\boldsymbol{X})} \mathrm{E}\left(\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}\right) - \boldsymbol{\Theta}\right)^2\right). \tag{29}$$

Proof. Let $\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right)$ denote an arbitrary estimator. We will show that the MSE incurred by $\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right)$ is always greater or equal to the MSE incurred by $\mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right)$. We begin by computing the MSE conditioned on $\mathbf{X} = \mathbf{x}$,

$$\begin{aligned}
\mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \boldsymbol{\Theta}\right)^2 \big|\mathbf{X} = \mathbf{x}\Big) &= \mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right) + \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right) - \boldsymbol{\Theta}\right)^2 \Big|\mathbf{X} = \mathbf{x}\Big) && (30)\\
&= \left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}\right)\right)^2 + \mathrm{E}\Big(\left(\mathrm{E}\left(\boldsymbol{\Theta}\right) - \boldsymbol{\Theta}\right)^2 \Big|\mathbf{X} = \mathbf{x}\Big) && (31)\\
&\quad + 2\,\mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}\right)\right)\left(\mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}\right)\right)\Big)\\
&= \left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}\right)\right)^2 + \mathrm{E}\Big(\left(\mathrm{E}\left(\boldsymbol{\Theta}\right) - \boldsymbol{\Theta}\right)^2 \Big|\mathbf{X} = \mathbf{x}\Big). && (32)
\end{aligned}$$

By iterated expectation,

$$\begin{aligned}
\mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \Theta\right)^2\Big) &= \mathrm{E}\Big(\mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \boldsymbol{\Theta}\right)^2 \big|\mathbf{X}\Big)\Big) && (33)\\
&= \mathrm{E}\Big(\left(\hat{\boldsymbol{\theta}}\left(\mathbf{X}\right) - \mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right)\right)^2\Big) + \mathrm{E}\Big(\left(\mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right) - \boldsymbol{\Theta}\right)^2 \big|\mathbf{X}\Big) && (34)\\
&\geq \mathrm{E}\Big(\left(\mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right) - \boldsymbol{\Theta}\right)^2\Big), && (35)
\end{aligned}$$

Since the expectation of a nonnegative quantity is nonnegative. This establishes that $\mathrm{E}\left(\boldsymbol{\Theta}|\mathbf{X}\right)$ achieves the minimum MSE. □

The following example illustrates Bayesian inference applied to the problem of determining the bias of a coin by flipping it several times, or equivalently of fitting the parameter of a Bernoulli random variable from iid realizations.

Example 1.7 (Bayesian analysis of the parameter of a Bernoulli distribution)**.** Let $x_1, x_2, \ldots$ be data that we wish to model as iid samples from a Bernoulli distribution. Since we are taking a Bayesian approach we are forced to choose a prior distribution for the parameter of the Bernoulli. We will consider two different Bayesian estimators Θ_1 and Θ_2:

1. Θ_1 represents a conservative estimator in terms of prior information. We assign a uniform pdf to the parameter. Any value in the unit interval has the same probability density:

$$f_{\Theta_1}(\theta) = \begin{cases} 1 & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

2. Θ_2 is an estimator that assumes that the parameter is closer to 1 than to 0. We could use it for instance to capture the suspicion that a coin is biased towards heads. We choose a skewed pdf that increases linearly from zero to one,

$$f_{\Theta_2}(\theta) = \begin{cases} 2\,\theta & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

Recall from the likelihood estimation (7) in Example 1.3 that the conditional pmf of the data given the parameter of the Bernoulli Θ equals

$$p_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \theta^{n_1}(1-\theta)^{n_0}, \tag{38}$$

where n_1 is the number of ones in the data and n_0 the number of zeros. The posterior pdfs of the two estimators are consequently equal to

$$f_{\Theta_1|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\boldsymbol{\Theta_1}}(\theta)\, p_{\mathbf{X}|\boldsymbol{\Theta_1}}(\mathbf{x}|\theta)}{\int_u f_{\boldsymbol{\Theta_1}}(u)\, p_{\mathbf{X}|\boldsymbol{\Theta_1}}(\mathbf{x}|u)\, \mathrm{d}u} \tag{39}$$

$$= \frac{\theta^{n_1}(1-\theta)^{n_0}}{\int_u u^{n_1}(1-u)^{n_0}\, \mathrm{d}u} \tag{40}$$

$$= \frac{\theta^{n_1}(1-\theta)^{n_0}}{\beta(n_1+1, n_0+1)}, \tag{41}$$

$$f_{\Theta_2|\mathbf{X}}(\theta|\mathbf{x}) = \frac{\theta^{n_1+1}(1-\theta)^{n_0}}{\int_u u^{n_1+1}(1-u)^{n_0}\, \mathrm{d}u} \tag{42}$$

$$= \frac{\theta^{n_1+1}(1-\theta)^{n_0}}{\beta(n_1+2, n_0+1)}, \tag{43}$$

$$\tag{44}$$

where

$$\beta(x,y) := \int_u u^{x-1}(1-u)^{y-1}\, \mathrm{d}u \tag{45}$$

is a special tabulated function.

In order to obtain point estimates for the parameter we compute the posterior means:

$$\begin{aligned}
\mathrm{E}\left(\Theta_1 | \mathbf{X} = \mathbf{x}\right) &= \int_0^1 \theta f_{\Theta_1|\mathbf{X}}\left(\theta|\mathbf{x}\right) \, \mathrm{d}\theta && (46) \\
&= \frac{\int_0^1 \theta^{n_1+1} \left(1-\theta\right)^{n_0} \, \mathrm{d}\theta}{\beta\left(n_1+1, n_0+1\right)} && (47) \\
&= \frac{\beta\left(n_1+2, n_0+1\right)}{\beta\left(n_1+1, n_0+1\right)}, && (48) \\
\mathrm{E}\left(\Theta_2 | \mathbf{X} = \mathbf{x}\right) &= \int_0^1 \theta f_{\Theta_2|\mathbf{X}}\left(\theta|\mathbf{x}\right) \, \mathrm{d}\theta && (49) \\
&= \frac{\beta\left(n_1+3, n_0+1\right)}{\beta\left(n_1+2, n_0+1\right)}. && (50)
\end{aligned}$$

Figure 2 shows the plot of the posterior distribution for different values of n_1 and n_0. It also plots the posterior mean and the ML estimate. For a small number of flips, the posterior pdf of Θ_2 is skewed to the right with respect to that of Θ_1, reflecting the prior belief that the parameter is closer to 1. However for a large number of flips both posterior densities are very close and so are the posterior means and the ML estimates; the likelihood term dominates when the number of data grows.

In Figure 2 we can see that the maximum likelihood is the mode (maximum value) of the posterior distribution when the prior is uniform. This is no accident.

Lemma 1.8. *The maximum likelihood is the mode (maximum value) of the posterior distribution if the prior distribution is uniform.*

Proof. We prove the result when the model for the data and the parameters is continuous, if any or both of them are discrete the proof is identical. If the prior distribution of the parameters is uniform then

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta}} f_{\Theta|\mathbf{X}}\left(\theta|\mathbf{x}\right) &= \arg\max_{\theta} \frac{f_{\boldsymbol{\Theta}}\left(\theta\right) f_{\mathbf{X}|\boldsymbol{\Theta}}\left(\mathbf{x}|\theta\right)}{\int_u f_{\boldsymbol{\Theta}}\left(u\right) f_{\mathbf{X}|\boldsymbol{\Theta}}\left(\mathbf{x}|u\right) \, \mathrm{d}u} && (51) \\
&= \arg\max_{\theta} f_{\mathbf{X}|\boldsymbol{\Theta}}\left(\mathbf{x}|\theta\right) && (52) \\
&= \arg\max_{\theta} \mathcal{L}\left(\theta\right), && (53)
\end{aligned}$$

which is the ML estimator. □

Note that uniform priors are only well defined in situations where the parameter is restricted to a bounded set.

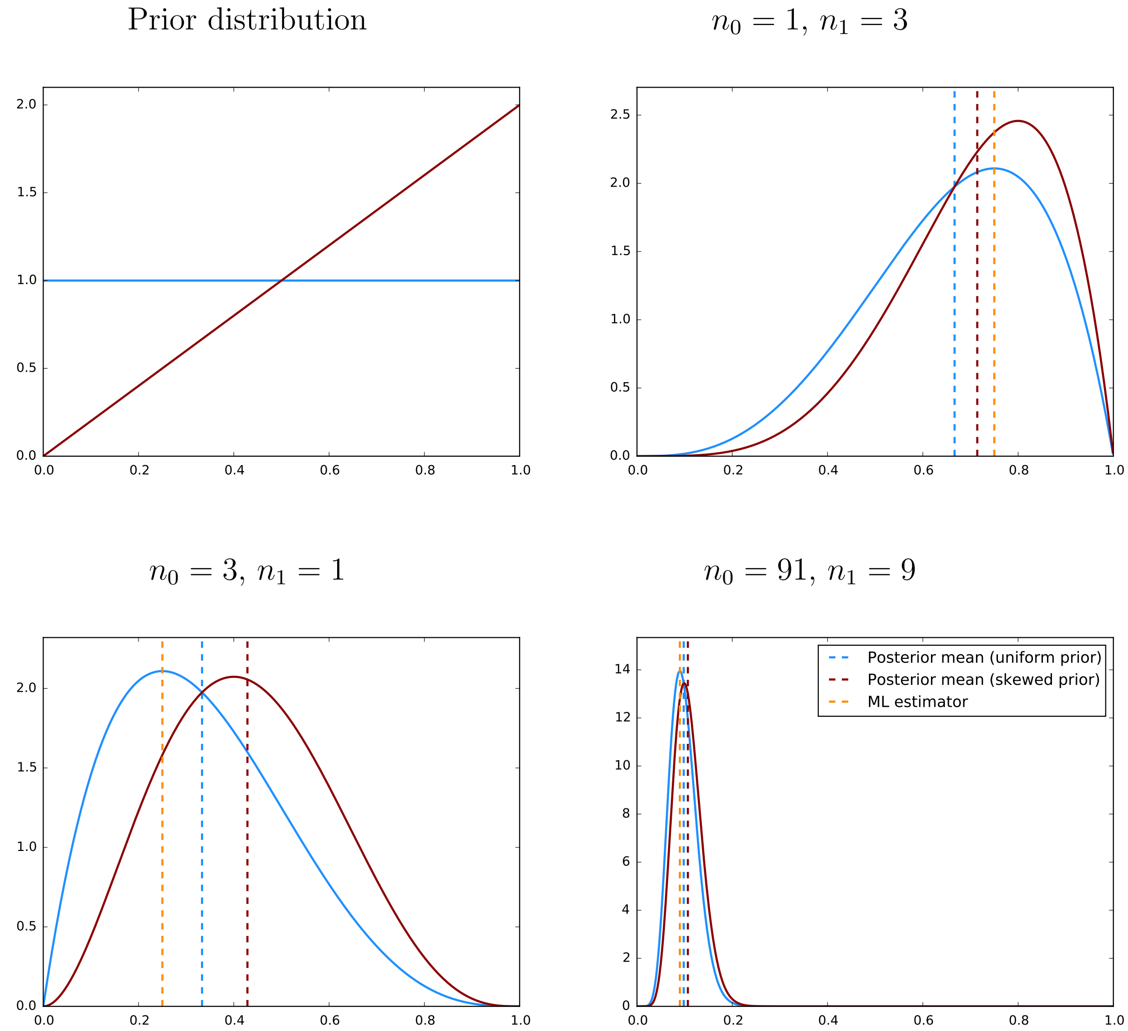

Figure 2: Posterior distributions of the parameter of a Bernoulli for two different priors and for different data realizations.

2 Nonparametric models

In situations where a parametric model is not available or does not fit the data adequately, we resort to nonparametric methods in order to characterize the unknown distribution that is supposed to generate the data. Learning a model that does not rely on a small number of parameters is challenging: we need to estimate the *whole distribution* just from the available samples. Without further assumptions this problem is ill posed; many (infinite!) different distributions could have generated the data. However, as we show below, with enough samples it is possible to obtain models that characterize the underlying distribution quite accurately.

2.1 Estimating the cdf

A way of characterizing the distribution that generates the data is to approximate its cdf. An intuitive estimate is obtained by computing the fraction of samples that are smaller than a certain value. This produces a piecewise constant estimator known as the **empirical cdf**.

Definition 2.1 (Empirical cdf)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables belonging to the same probability space. The value of the empirical cdf at any $x \in \mathbb{R}$ is*

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}. \tag{54}$$

The empirical cdf is an unbiased and consistent estimator of the true cdf. This is established rigorously in Theorem 2.2 below, but is also illustrated empirically in Figure 3. The cdf of the height data in Figure 1 of Lecture Notes 4 is compared to three realizations of the empirical cdf computed from iid samples. As the number of available samples grows, the approximation becomes very accurate.

Theorem 2.2. *Let $X_1, X_2, \ldots$ be an iid sequence with cdf F_X. For any fixed $u \in \mathbb{R}$ $\widehat{F}_{\boldsymbol{x}}(x)$ is an unbiased and consistent estimator of $F_X(x)$. In fact, $\widehat{F}_{\boldsymbol{x}}(x)$ converges in mean square to $F_X(x)$.*

Proof. First, we verify

$$\begin{aligned}
\mathrm{E}\left(\widehat{F}_n(x)\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}\right) && (55)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}\left(X_i \leq x\right) \quad \text{by linearity of expectation} && (56)\\
&= F_X(x), && (57)
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{E}\left(\widehat{F}_n^2\left(x\right)\right) &= \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1_{X_i \leq x} 1_{X_j \leq x}\right) \qquad (58)\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{P}\left(X_i \leq x\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} \mathrm{P}\left(X_i \leq x, X_j \leq x\right) \quad \text{by linearity of expectation}\\
&= \frac{F_X\left(x\right)}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} F_{X_i}\left(x\right) F_{X_j}\left(x\right) \quad \text{by independence,} \qquad (59)\\
&= \frac{F_X\left(x\right)}{n} + \frac{n-1}{n} F_X^2\left(x\right). \qquad (60)
\end{aligned}$$

The variance is consequently equal to

$$\begin{aligned}
\mathrm{Var}\left(\widehat{F}_n\left(x\right)\right) &= \mathrm{E}\left(\widehat{F}_n\left(x\right)^2\right) - \mathrm{E}^2\left(\widehat{F}_n\left(x\right)\right) \qquad (61)\\
&= \frac{F_X\left(x\right)\left(1 - F_X\left(x\right)\right)}{n}. \qquad (62)
\end{aligned}$$

We conclude that

$$\lim_{n\to\infty} \mathrm{E}\left(F_X\left(x\right) - \widehat{F}_n\left(x\right)\right) = \lim_{n\to\infty} \mathrm{Var}\left(\widehat{F}_n\left(x\right)\right) = 0. \tag{63}$$

□

2.2 Estimating the pdf

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).. How should we estimate $f\left(x\right)$ then?

Intuitively, an estimator for $f\left(x\right)$ should take into account the presence of samples at neighboring locations. If there are many samples close to x then we should estimate a higher probability density at x, whereas if all the samples are far away, then the estimate for $f\left(x\right)$ should be small. The **kernel density estimator** implements these ideas by computing a local weighted average at each point x such that the contribution of each sample depends on its distance to x.

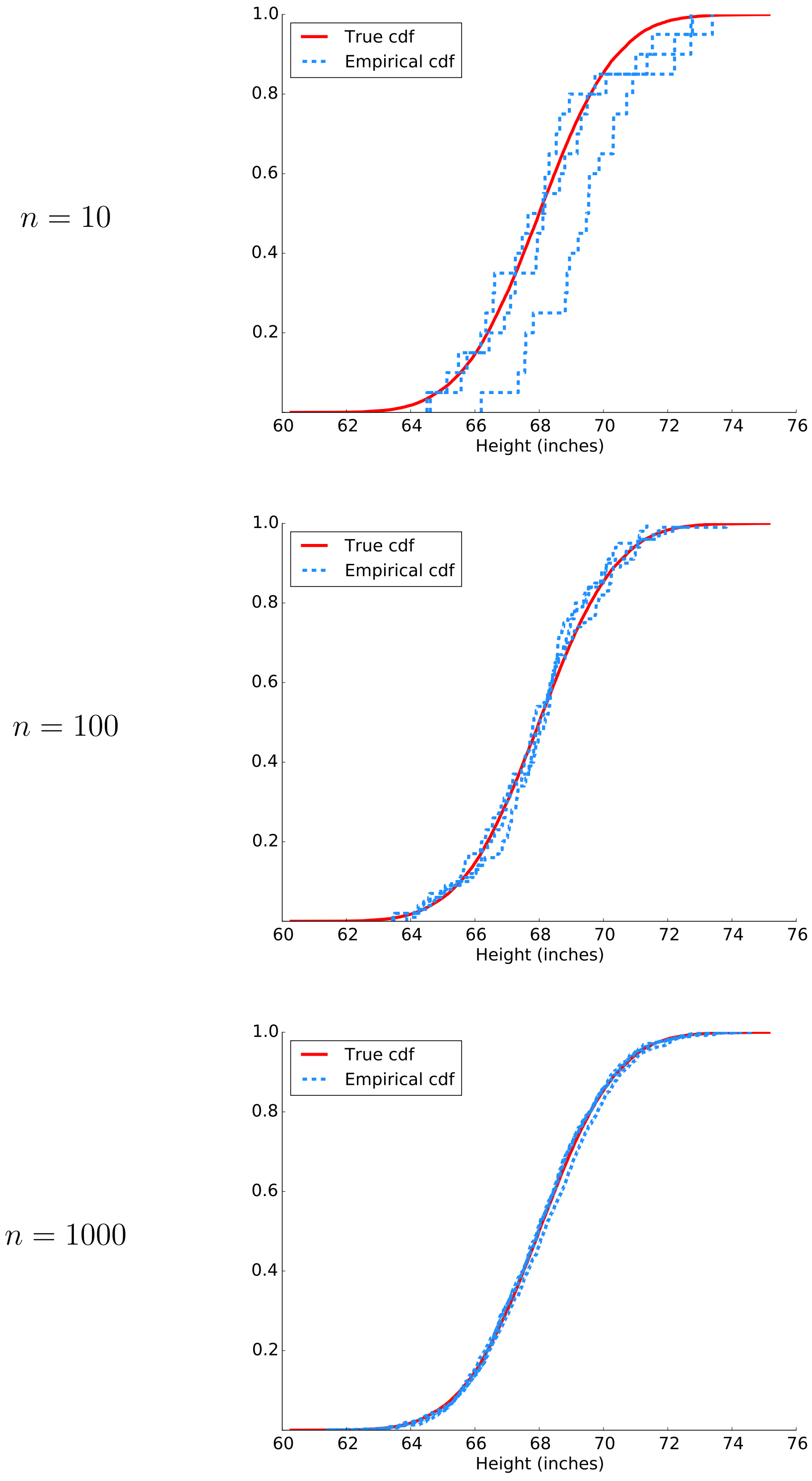

Figure 3: Cdf of the height data in Figure 1 of Lecture Notes 4 along with three realizations of the empirical cdf computed with n iid samples for $n = 10, 100, 1000$.

Definition 2.3 (Kernel density estimator). *Let $X_1, X_2, \ldots$ be a sequence of random variables belonging to the same probability space. The value of the kernel density estimator with bandwidth h at $x \in \mathbb{R}$ is*

$$\widehat{f}_{h,n}(x) := \frac{1}{n\,h} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \tag{64}$$

where k is a kernel function with a maximum at the origin which decreases away from the origin and satisfies

$$k(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}, \tag{65}$$

$$\int_{\mathbb{R}} k(x)\ dx = 1. \tag{66}$$

Choosing a rectangular kernel yields an empirical density estimate that looks like a histogram. A more popular kernel is the Gaussian kernel $k(x) = e^{-x^2}$, which produces a smooth density estimate. Figure 4 shows the result of computing a Gaussian kernel to estimate the probability density of the weight of a population of sea snails[1]. The whole population consists of 4177 individuals. Our task is to estimate this distribution from just 200 iid samples. The plots show the enormous influence that the bandwidth parameter, which determines the width of the kernel, can have on the result. 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. Increasing the bandwidth smooths out such fluctuations. However, increasing the bandwidth too much smooths out structure that may be actually present in the true pdf. A good tradeoff is difficult to achieve. In practice this parameter must be calibrated from the data.

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

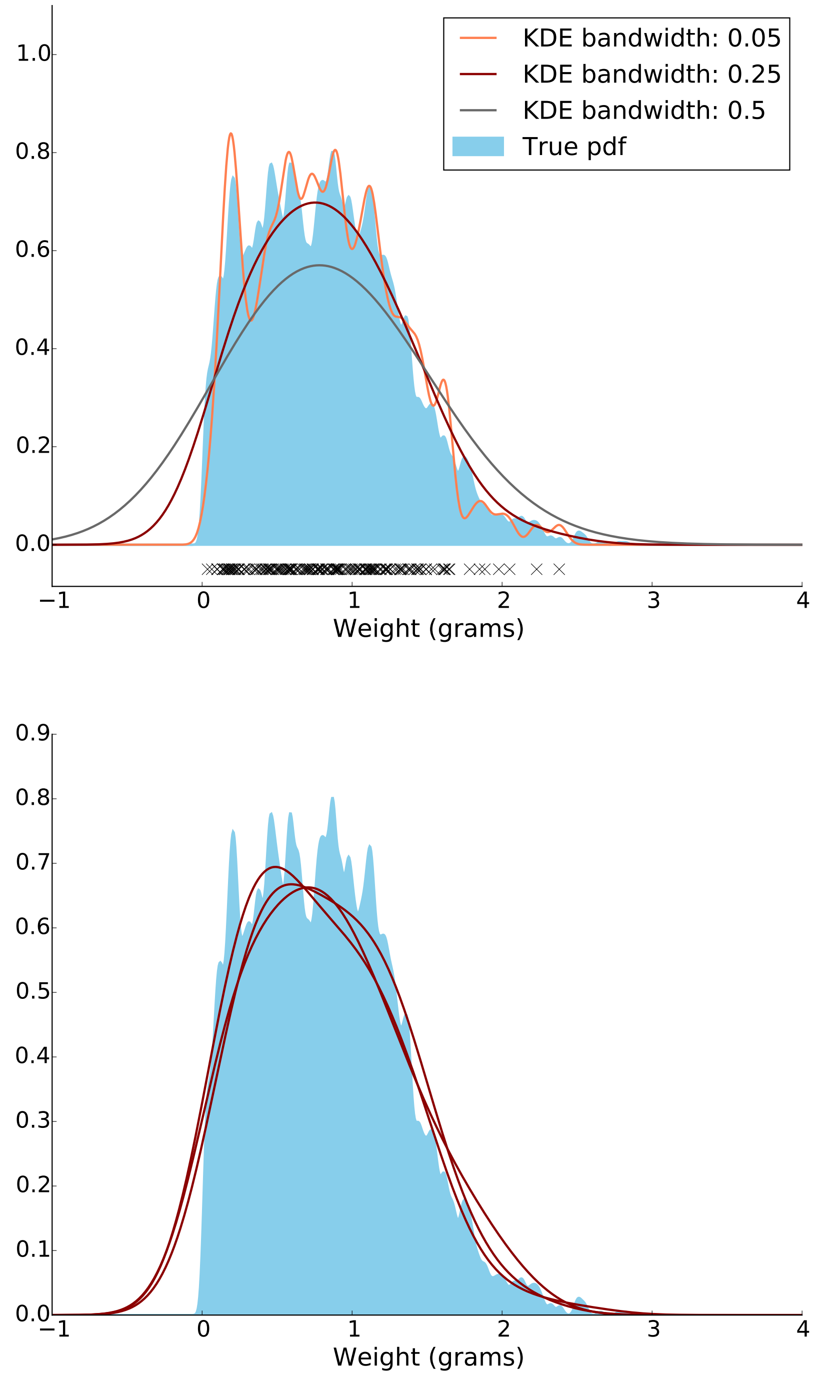

Figure 4: Kernel density estimate for the weight of a population of abalone, a species of sea snail. In the plot above the density is estimated from 200 iid samples using a Gaussian kernel with three different bandwidths. Black crosses representing the individual samples are shown underneath. In the plot below we see the result of repeating the procedure three times using a fixed bandwidth equal to 0.25.