Statistics and ML – 1

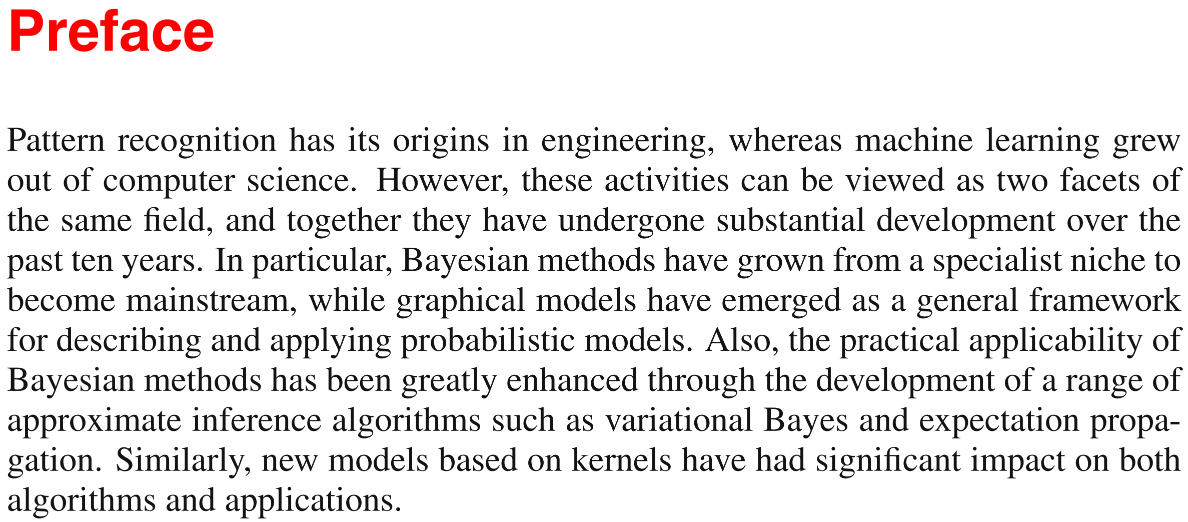
A. Naive Bayesian Classifiers
B. The Expectation-Maximization (EM) algorithm

Edoardo Milotti

Advanced Statistics for Physics

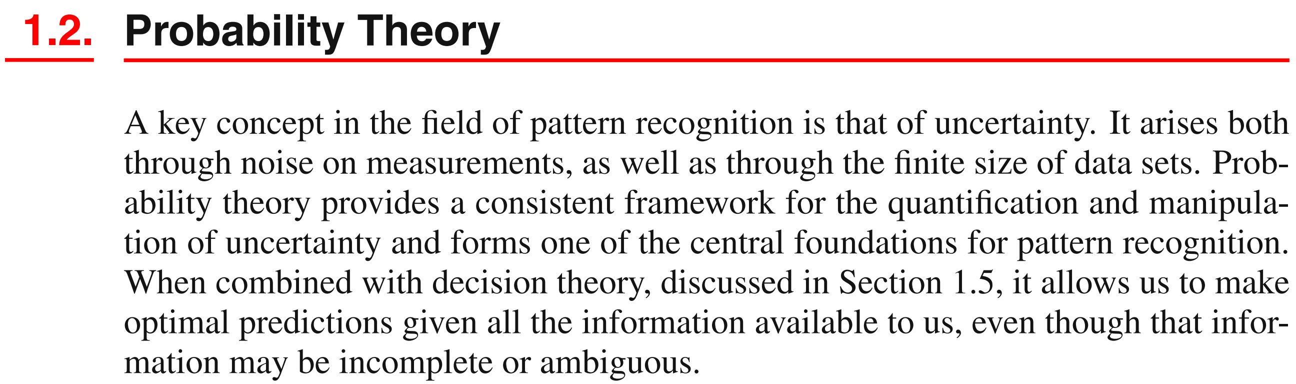
A.Y. 2023-24

Statistics and Machine Learning (ML)



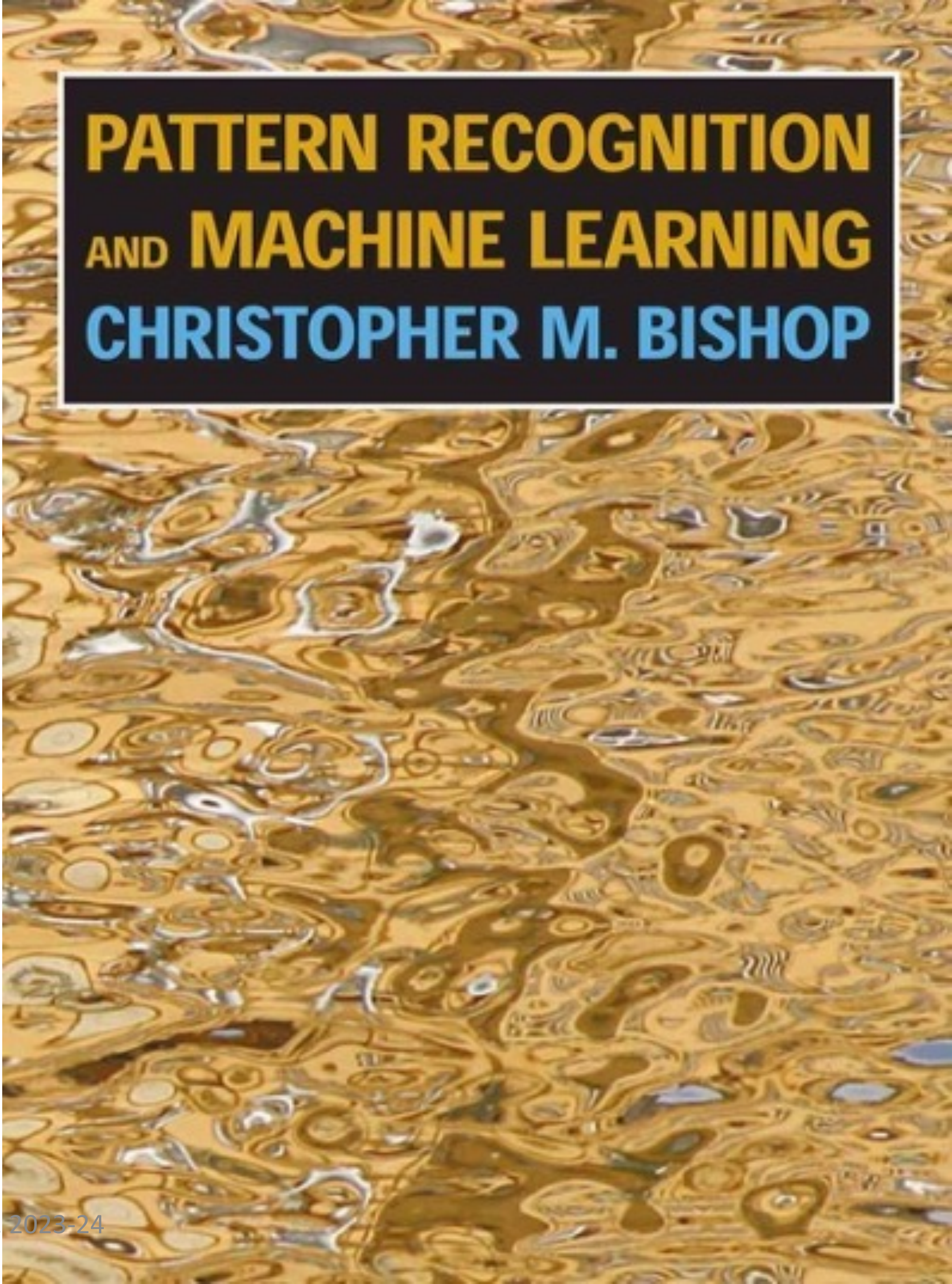
Preface

Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field, and together they have undergone substantial development over the past ten years. In particular, Bayesian methods have grown from a specialist niche to become mainstream, while graphical models have emerged as a general framework for describing and applying probabilistic models. Also, the practical applicability of Bayesian methods has been greatly enhanced through the development of a range of approximate inference algorithms such as variational Bayes and expectation propagation. Similarly, new models based on kernels have had significant impact on both algorithms and applications.



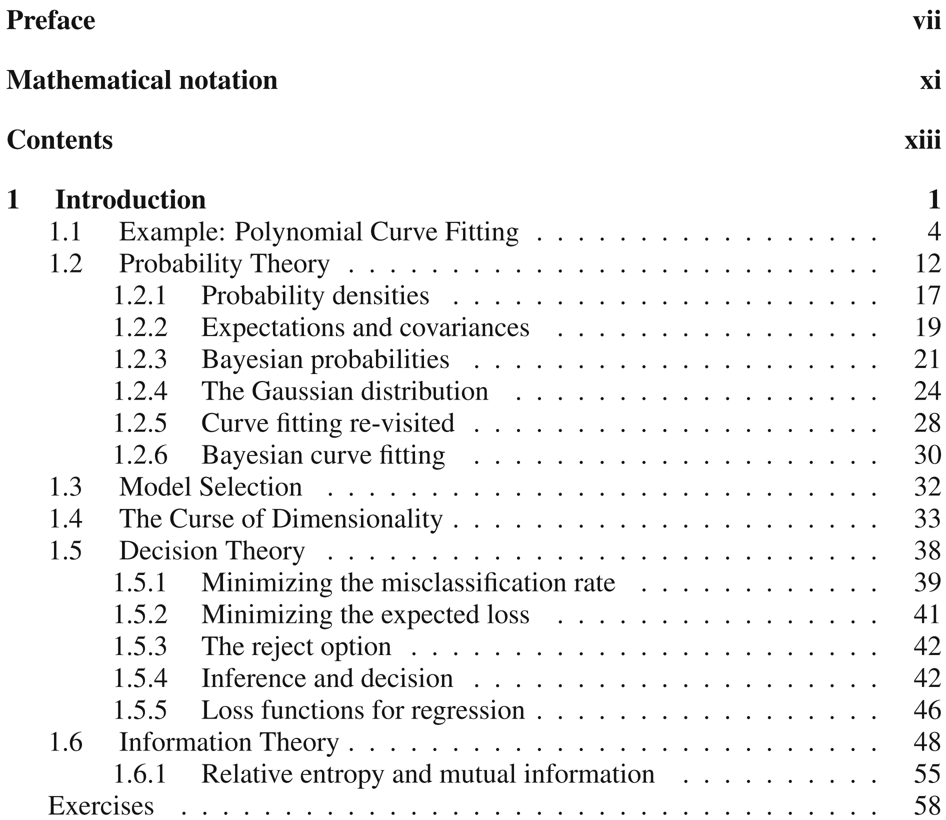
1.2. Probability Theory

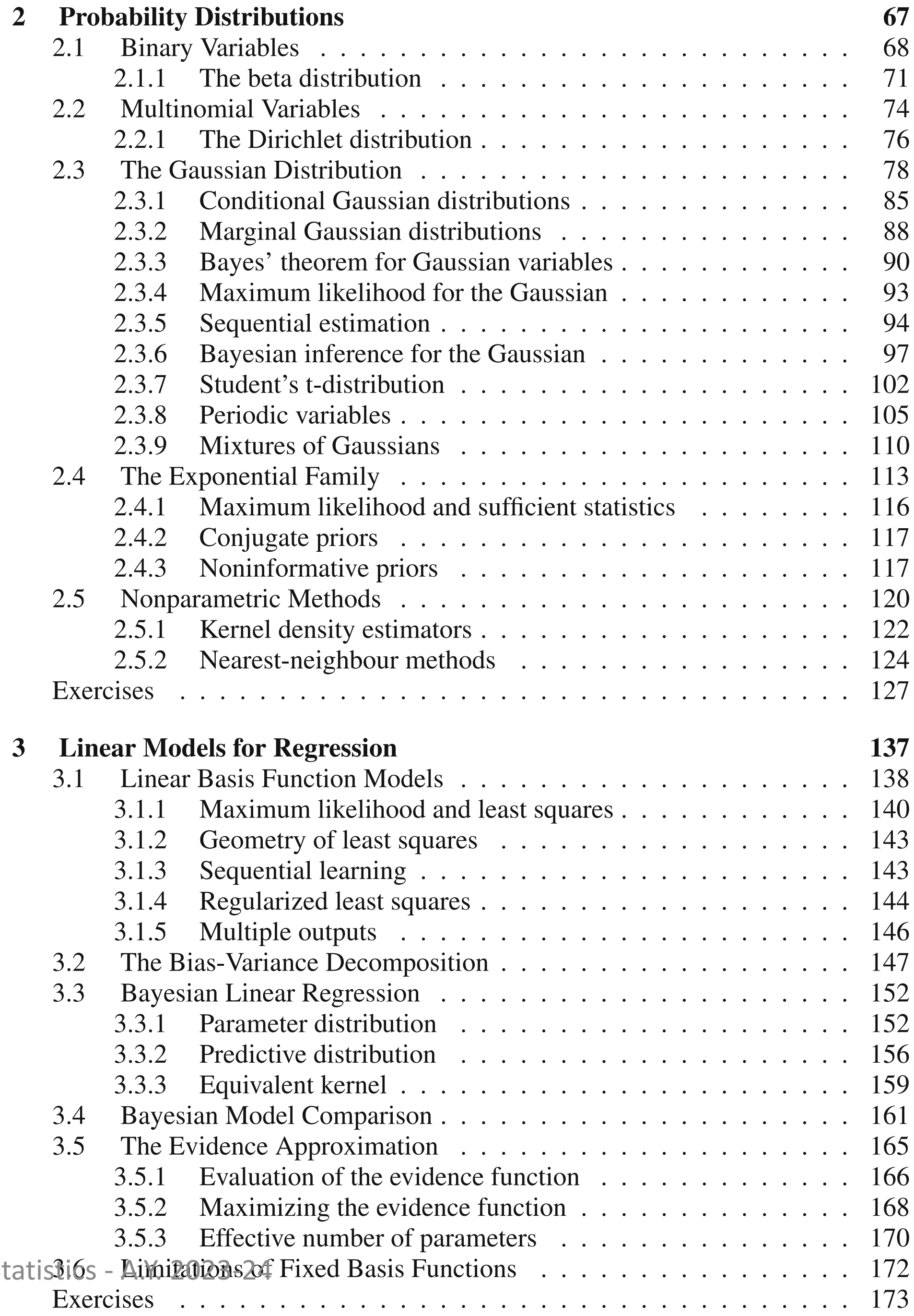
A key concept in the field of pattern recognition is that of uncertainty. It arises both through noise on measurements, as well as through the finite size of data sets. Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition. When combined with decision theory, discussed in Section 1.5, it allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

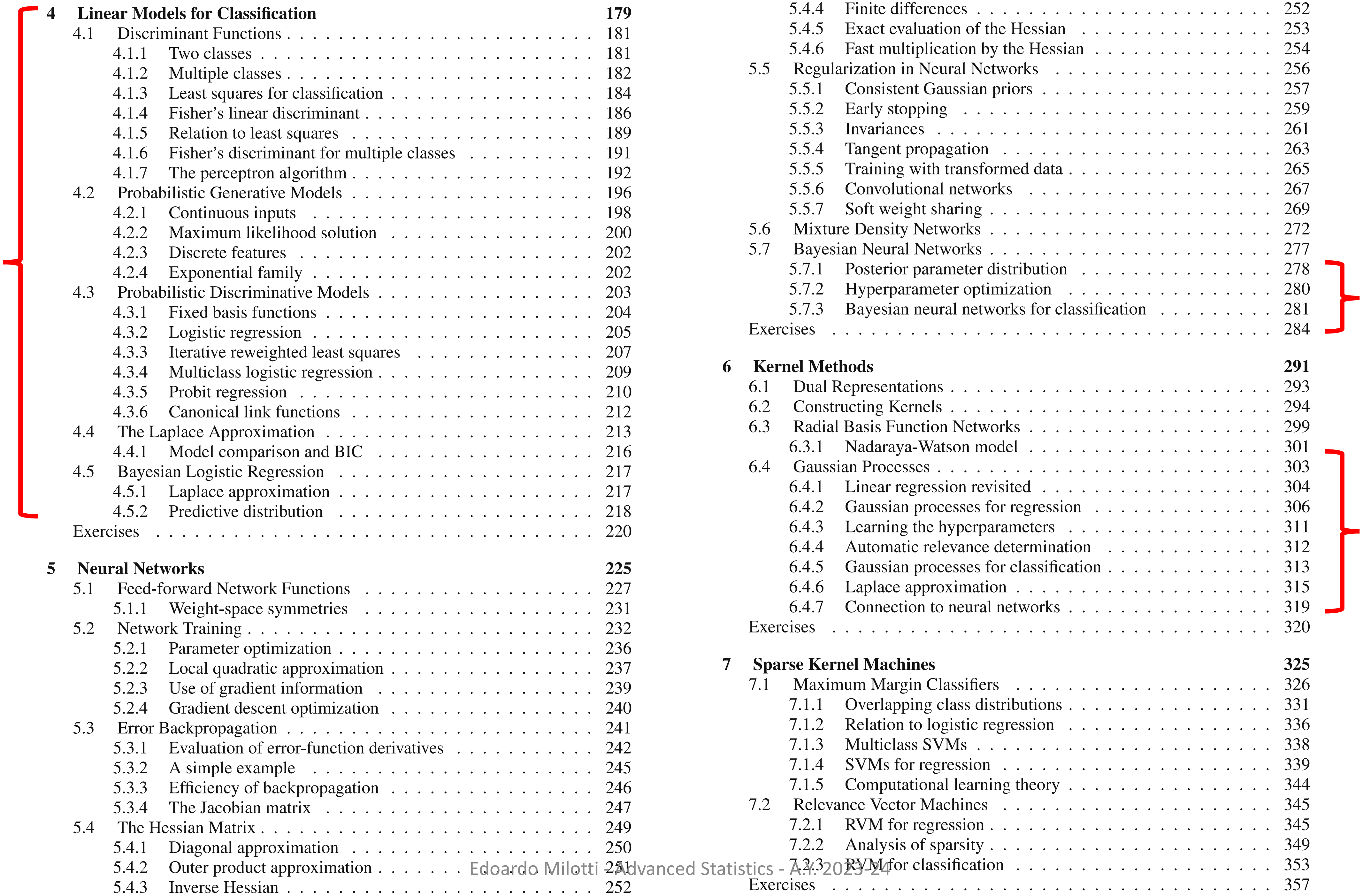


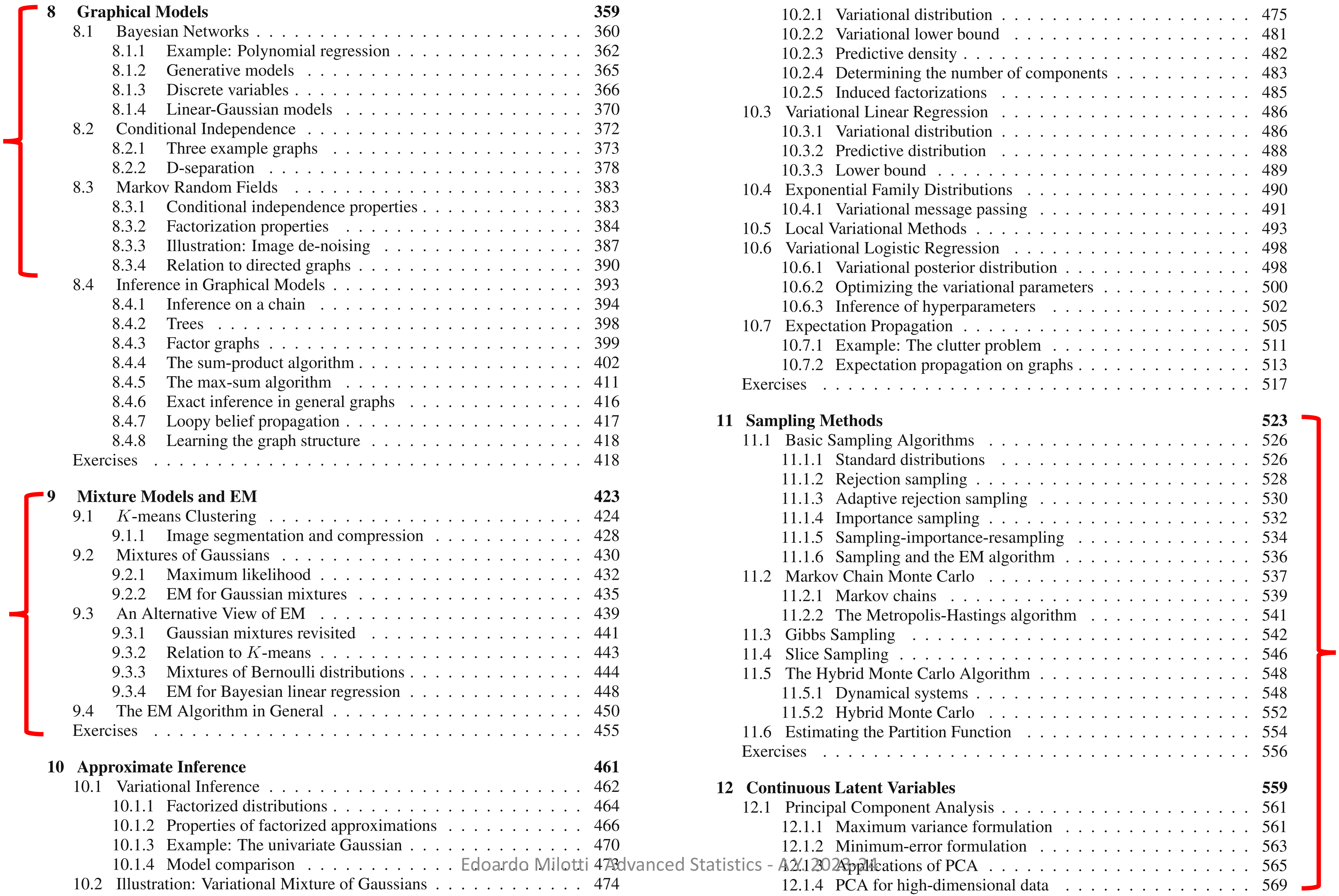
https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/

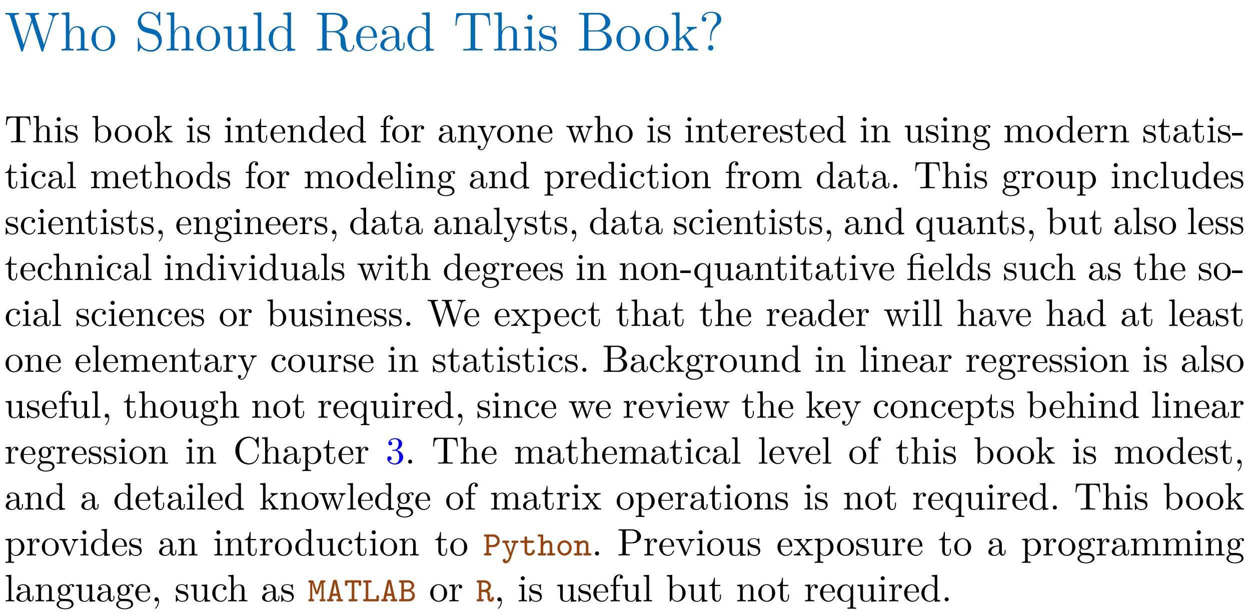
Contents





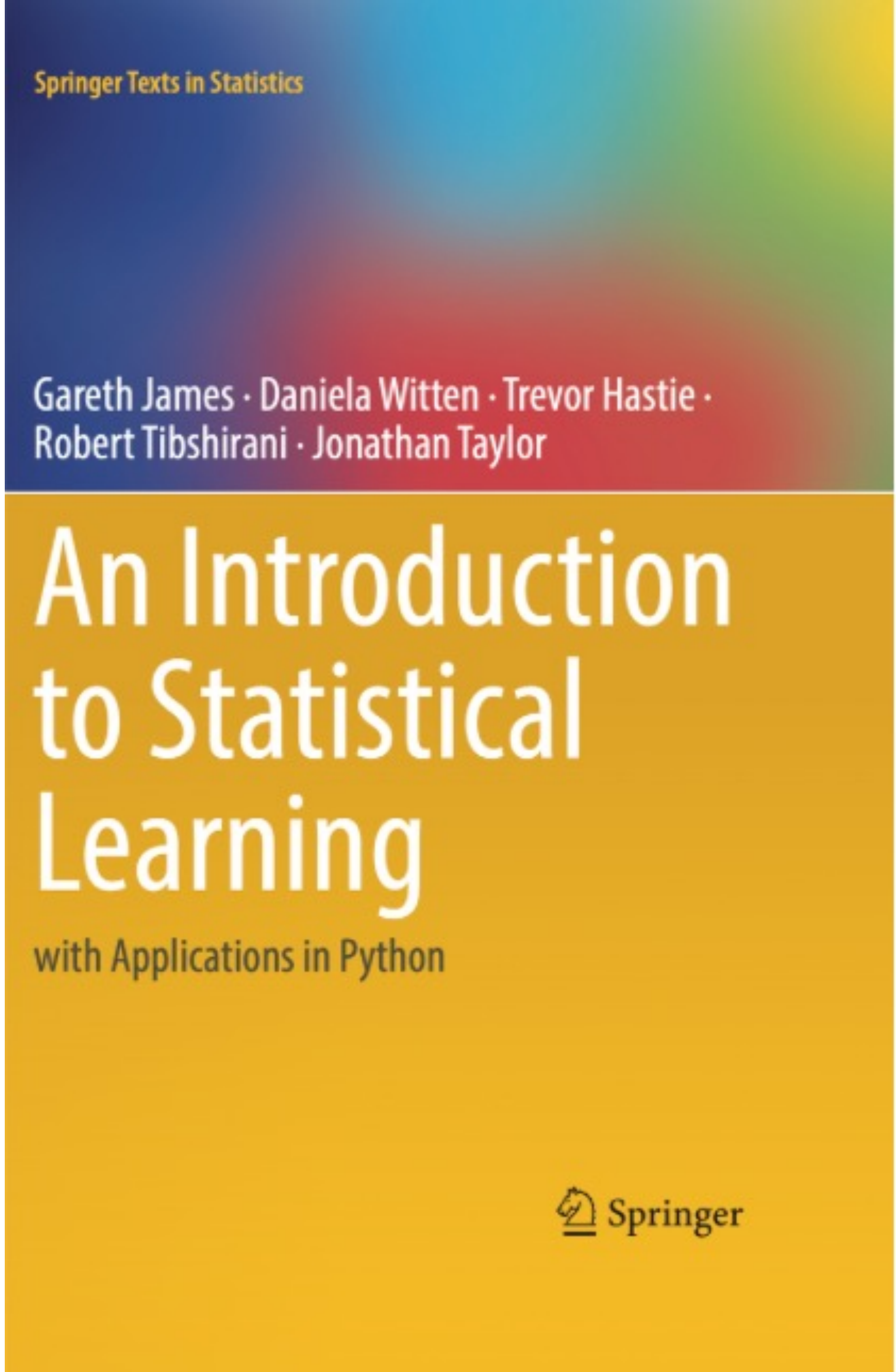






Who Should Read This Book?

This book is intended for anyone who is interested in using modern statistical methods for modeling and prediction from data. This group includes scientists, engineers, data analysts, data scientists, and quants, but also less technical individuals with degrees in non-quantitative fields such as the social sciences or business. We expect that the reader will have had at least one elementary course in statistics. Background in linear regression is also useful, though not required, since we review the key concepts behind linear regression in Chapter 3. The mathematical level of this book is modest, and a detailed knowledge of matrix operations is not required. This book provides an introduction to `Python`. Previous exposure to a programming language, such as `MATLAB` or `R`, is useful but not required.



Springer Texts in Statistics

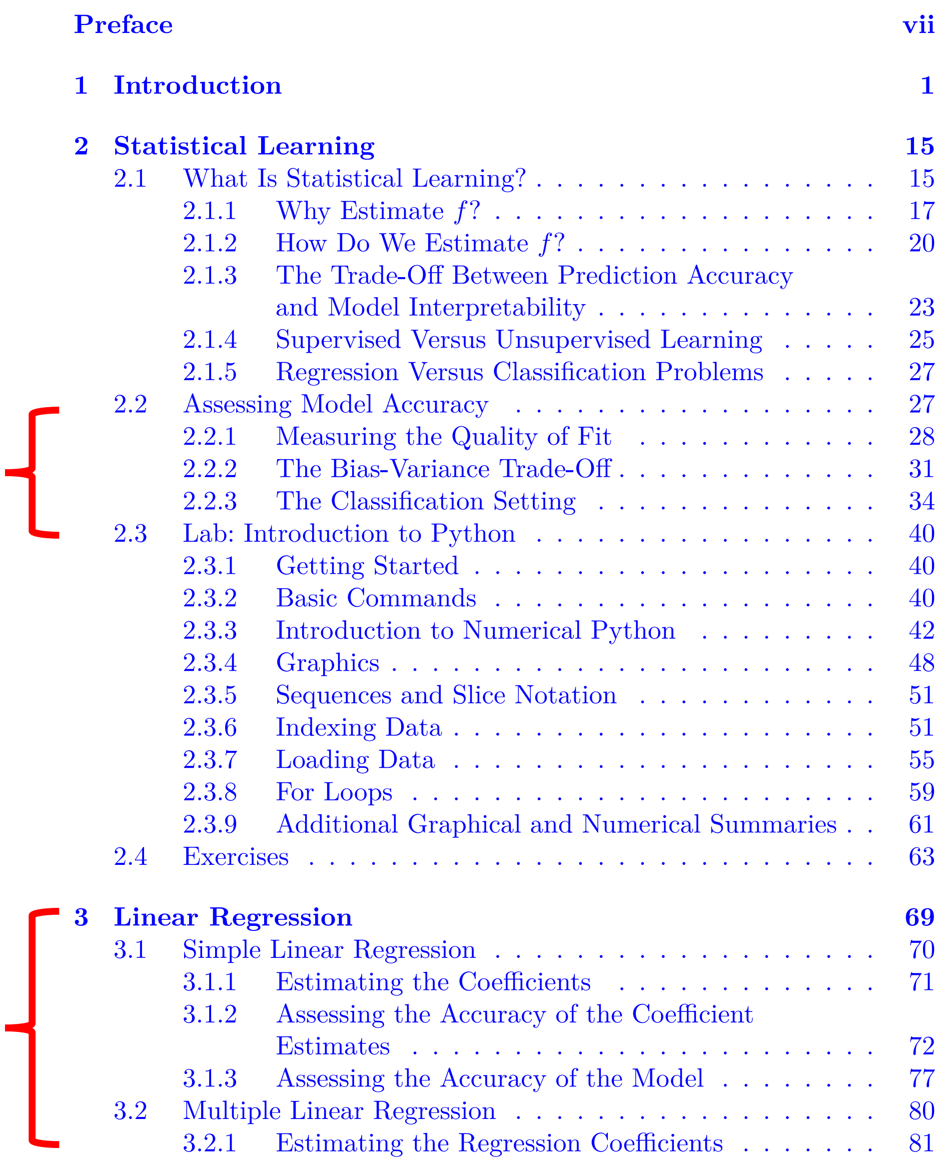
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor

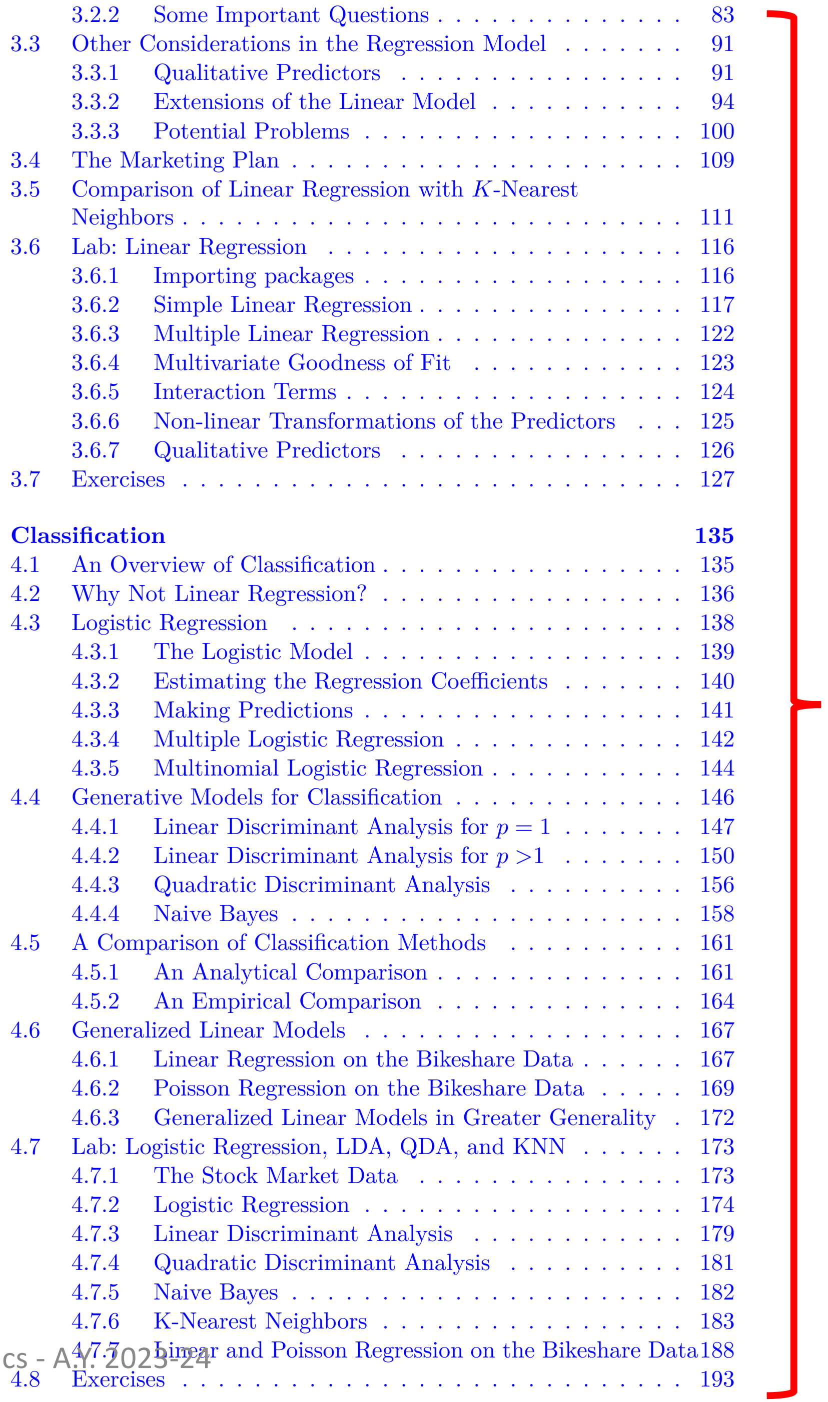
An Introduction to Statistical Learning

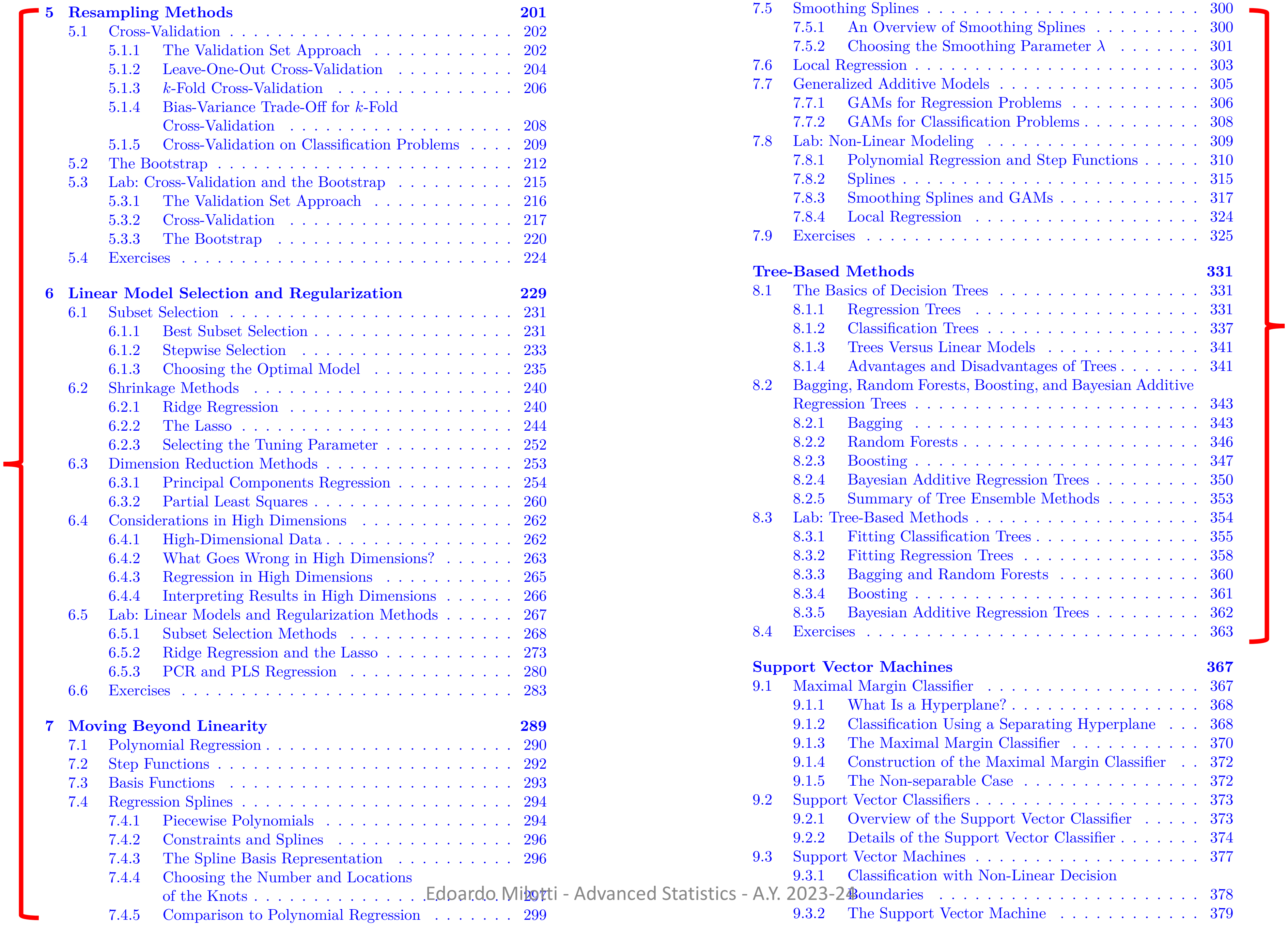
with Applications in Python

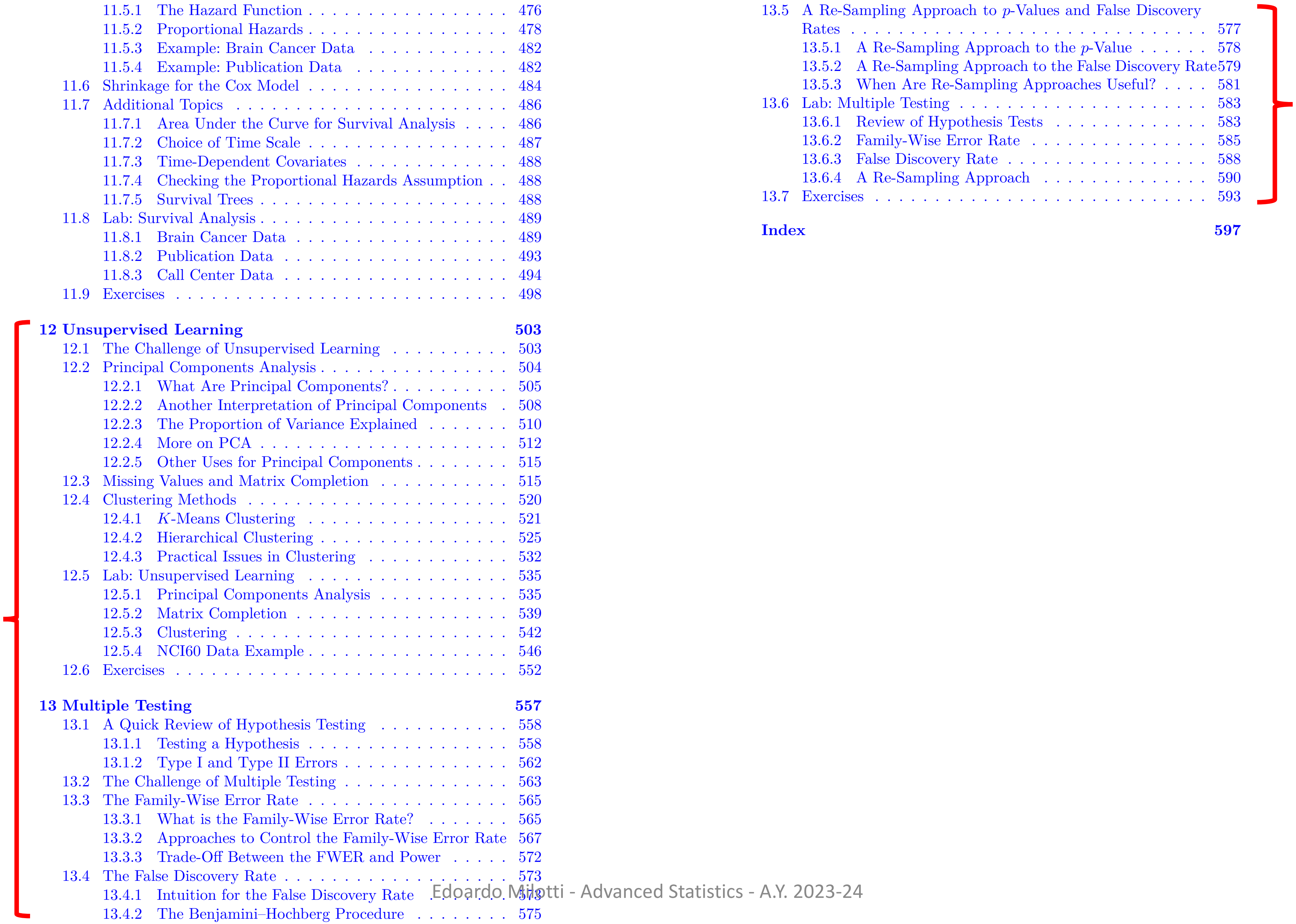
Springer

https://www.statlearning.com









A. Naive Bayesian Classifiers

1. Bayesian classification

data X, classes C

this likelihood is defined by training data

$$P(C|X) = \frac{P(X|C)}{P(X)} P(C)$$

the prior is also defined by training data

we can use the prior learning to assign a class to new data

$$C_k = \arg\max_{C_k} \frac{P(X|C_k)}{P(X)} P(C_k) = \arg\max_{C_k} P(X|C_k) P(C_k)$$

Consider a vector of *N* attributes given as Boolean variables $\mathbf{x} = \{x_i\}$ and classify the data vectors with a single Boolean variable.

The learning procedure must yield:

$P(y)$

it is easy to obtain it as an empirical distribution from a histogram of training class data: y is Boolean, the histogram has just two bins, and a hundred examples suffice to determine the empirical distribution to better than 10%.

$P(\mathbf{x}|y)$

there is a bigger problem here: the arguments have 2^{N+1} different values, and we must estimate $2(2^N-1)$ parameters ... for instance, with N = 30 there are more than 2 billion parameters!

How can we reduce the huge complexity of learning?

we assume the conditional independence of the x_n's:
naive Bayesian learning

for instance, with just two attributes

$$P\left(x_1, x_2 \middle| y\right) = P\left(x_1 \middle| x_2, y\right) P\left(x_2 \middle| y\right) = P\left(x_1 \middle| y\right) P\left(x_2 \middle| y\right)$$

conditional independence assumption

with more than 2 attributes

$$P\left(\mathbf{x} \middle| y\right) \approx \prod_{k=1}^{N} P\left(x_k \middle| y\right)$$

Therefore:

$$P\left(y_k|\mathbf{x}\right) = \frac{P\left(\mathbf{x}|y_k\right)}{P\left(\mathbf{x}\right)}P\left(y_k\right) = \frac{P\left(\mathbf{x}|y_k\right)}{\sum_j P\left(\mathbf{x}|y_j\right)P\left(y_j\right)}P\left(y_k\right)$$

$$\approx \frac{\prod_{n=1}^{N} P\left(x_n|y_k\right)}{\sum_j P\left(y_j\right)\prod_{n=1}^{N} P\left(x_n|y_j\right)}P\left(y_k\right)$$

and we assign the class according to the rule (MAP)

$$\boxed{y = \arg\max_{y_k} \frac{\prod_{n=1}^{N} P\left(x_n|y_k\right)}{\sum_j P\left(y_j\right)\prod_{n=1}^{N} P\left(x_n|y_j\right)}P\left(y_k\right)}$$

More general discrete inputs

If any of the N x input variables has J different values, and if there are K classes, then we must estimate in all $NK(J\text{-}1)$ free parameters with the Naive Bayes Classifier (this includes normalization) (compare this with the $K(J^N\text{-}1)$ parameters needed by a complete classifier)

Continuous inputs and discrete classes – the Gaussian case

$$P\left(x_n \middle| y_k\right) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{\left(x_n - \mu_{nk}\right)^2}{2\sigma_{nk}^2}\right]$$

here we must estimate $2NK$ parameters + the shape of the distribution $P(y)$ (this adds up to another K-1 parameters)

Gaussian special case with class-independent variance and Boolean classification (two classes only):

$$P\left(y=0\middle|\mathbf{x}\right)=\frac{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)+P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)}$$

$$P\left(x_n\middle|y=0\right)=\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{\left(x_n-\mu_{n0}\right)^2}{2\sigma_n^2}\right]$$

$$P\left(x_n\middle|y=1\right)=\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{\left(x_n-\mu_{n1}\right)^2}{2\sigma_n^2}\right]$$

$$
\begin{aligned}
P\left(y=0\middle|\mathbf{x}\right) &= \frac{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)+P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)} \\
&= \frac{1}{1+\dfrac{P\left(\mathbf{x}\middle|y=1\right)P\left(y=1\right)}{P\left(\mathbf{x}\middle|y=0\right)P\left(y=0\right)}} \\
&= \frac{1}{1+\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\displaystyle\prod_{n=1}^{N}\exp\left[-\dfrac{\left(x_n-\mu_{n1}\right)^2}{2\sigma_n^2}+\dfrac{\left(x_n-\mu_{n0}\right)^2}{2\sigma_n^2}\right]} \\
&= \frac{1}{1+\exp\left\{\ln\left(\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\right)+\displaystyle\sum_{n=1}^{N}\left[\dfrac{\left(\mu_{n1}-\mu_{n0}\right)x_n}{\sigma_n^2}+\dfrac{\mu_{n0}^2-\mu_{n1}^2}{2\sigma_n^2}\right]\right\}}
\end{aligned}
$$

$$w_0 = \ln\left(\frac{P(y=1)}{P(y=0)}\right) + \sum_{n=1}^{N}\left[\frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2}\right]$$

$$w_n = \frac{(\mu_{n1} - \mu_{n0})}{\sigma_n^2}$$

⇓

logistic shape

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P(y=1|\mathbf{x}) = 1 - P(y=0|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

Finally, an input vector belongs to class $y = 0$ if

$$\frac{P(y=0|\mathbf{x})}{P(y=1|\mathbf{x})} > 1$$

$$\Downarrow$$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P(y=1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$\Rightarrow \quad \exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right) < 1$$

$$\Rightarrow \quad w_0 + \sum_{n=1}^{N} w_n x_n < 0$$

B. The Expectation-Maximization (EM) algorithm

The EM algorithm is used to maximize likelihood with incomplete information (e.g., "latent variables"), and it has two main steps that are iterated until convergence:

E. expectation of the log-likelihood, averaged with respect to missing data:

parameters (with respect to which we want to maximize the expression

measured data

missing data

previous parameter estimate (constant values)

likelihood

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(n-1)}\right) = E_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\middle|\mathbf{x},\boldsymbol{\theta}^{(n-1)}\right]$$

$$= \int_Y \left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\right] p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(n-1)}\right) d\mathbf{y}$$

M. maximization of the averaged log-likelihood with respect to parameters:

$$\boldsymbol{\theta}^{(n)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(n-1)}\right)$$

Example: an experiment with an exponential model (Flury and Zoppè)

Light bulbs fail following an exponential distribution with mean failure time θ

To estimate the mean two experiments are performed

1. n light bulbs are tested, all failure times u_i are recorded
2. m light bulbs are tested, only the total number r of bulbs failed up to time t are recorded

1.

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left(-\frac{u_i}{\theta} \right) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_i u_i}{\theta} \right) = \frac{1}{\theta^n} \exp\left(-\frac{n\langle u \rangle}{\theta} \right)$$

2.

$$\mathcal{L} = \prod_{i=1}^{m} \frac{1}{\theta} \exp\left(-\frac{v_i}{\theta} \right)$$

missing data!

combined likelihood

$$\frac{1}{\theta^n}\exp\left(-\frac{n\langle u\rangle}{\theta}\right)\cdot\prod_{i=1}^{m}\frac{1}{\theta}\exp\left(-\frac{v_i}{\theta}\right)$$

log-likelihood

$$-n\ln\theta-\frac{n\langle u\rangle}{\theta}-\sum_{i=1}^{m}\left(\ln\theta+\frac{v_i}{\theta}\right)$$

expected failure time for a bulb that is still burning at time t

$$t+\theta$$

expected failure time for a bulb that is not burning at time t

$$\theta-\frac{t\exp\left(-t/\theta\right)}{1-\exp\left(-t/\theta\right)}$$

Note on mean failure time for a bulb that is not burning at time t

$$p(t') \propto \frac{1}{\theta} e^{-t'/\theta} \qquad 0 \le t' \le t$$

$$\text{normalization} = \int_0^t p(t')\,dt' = \int_0^t \frac{dt'}{\theta} e^{-t'/\theta} = 1 - e^{-t/\theta}$$

$$\text{mean failure time} = \int_0^t t' p(t')\,dt' = \frac{1}{1-e^{-t/\theta}} \int_0^t t' e^{-t'/\theta} \frac{dt'}{\theta}$$

$$= \frac{\theta}{1-e^{-t/\theta}} \left[1 - e^{-t/\theta} - (t/\theta) e^{-t/\theta} \right]$$

$$= \theta - \frac{t e^{-t/\theta}}{1 - e^{-t/\theta}}$$

average log-likelihood

$$
\begin{aligned}
Q &= E\left[-n\ln\theta - \frac{n\langle u\rangle}{\theta} + \sum_{i=1}^{m}\left(-\ln\theta - \frac{v_i}{\theta} \right) \right] \\
&= -(n+m)\ln\theta - \frac{n\langle u\rangle}{\theta} - \frac{r}{\theta}\left(\theta - \frac{t\exp(-t/\theta)}{1-\exp(-t/\theta)} \right) - \frac{(m-r)}{\theta}(\theta+t)
\end{aligned}
$$

this ends the expectation step

the max of the mean likelihood

$$Q = -(n+m)\ln\theta - \frac{1}{\theta}\left[n\langle u\rangle + r\left(\theta - \frac{t\exp\left(-t/\theta\right)}{1-\exp\left(-t/\theta\right)}\right) + (m-r)(\theta+t)\right]$$

can be found by maximizing the approximate expression

$$Q \approx -(n+m)\ln\theta - \frac{1}{\theta}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)}\right) + (m-r)\left(\theta^{(k)}+t\right)\right]$$

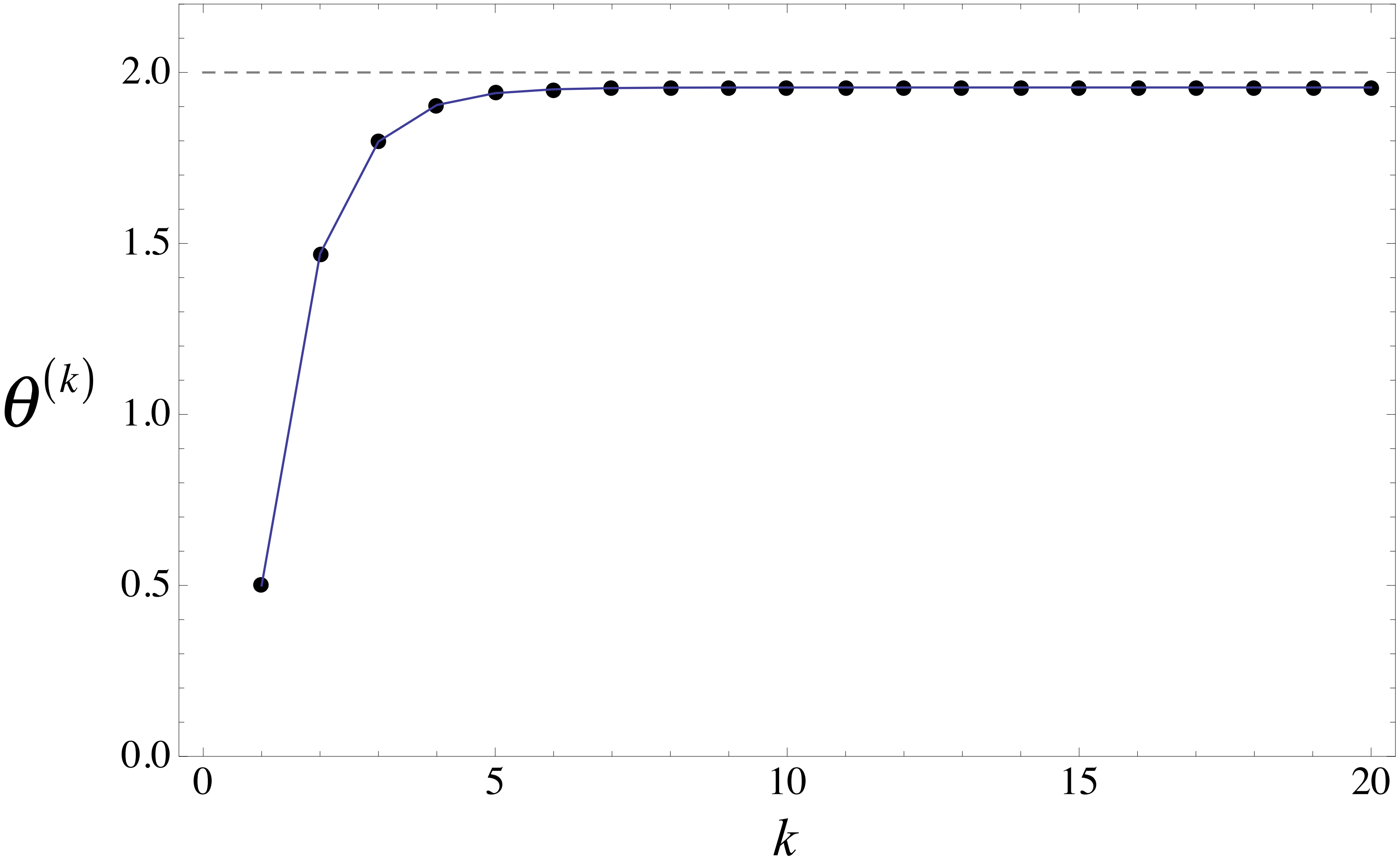
$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2}\left[n\langle u\rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)}\right) + (m-r)\left(\theta^{(k)}+t\right)\right] = 0$$

$$\frac{dQ}{d\theta} \approx -(n+m)\frac{1}{\theta} + \frac{1}{\theta^2}\left[n\langle u \rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)} \right) + (m-r)\left(\theta^{(k)}+t\right) \right] = 0$$

$$\theta^{(k+1)} = \frac{1}{n+m}\left[n\langle u \rangle + r\left(\theta^{(k)} - \frac{t\exp\left(-t/\theta^{(k)}\right)}{1-\exp\left(-t/\theta^{(k)}\right)} \right) + (m-r)\left(\theta^{(k)}+t\right) \right]$$

this formula summarizes expectation and maximization: therefore, the recipe is to iterate this until convergence ...

Example with mean failure time = 2 (a.u.), and randomly generated data ($n = 100$; $m = 100$). In this example $r = 36$.

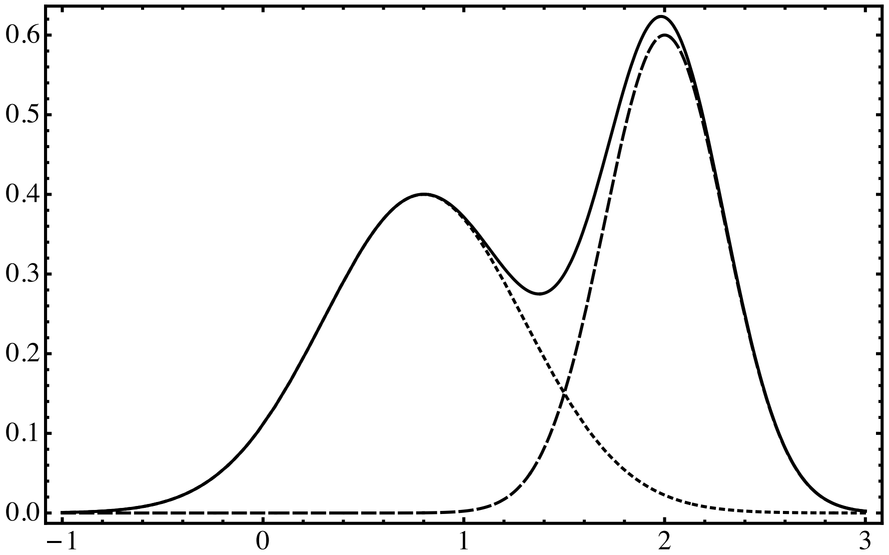


The EM method is often used to estimate the parameters of “mixture models”.

$$\boxed{p\left(x_n \middle| \boldsymbol{\theta}\right) = \sum_{i=1}^{M} \alpha_i p_i\left(x_n \middle| \boldsymbol{\theta}_i\right)}$$

$$\boldsymbol{\theta} = \left(\overbrace{\alpha_1,\ldots,\alpha_M}^{\substack{\text{"hidden parameters"} \\ \text{(also "latent parameters")}}};\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\right)$$

$$\sum_{i=1}^{M} \alpha_i = 1$$



Example: a Gaussian mixture model (M=2)

direct maximization of log likelihood

$$\log \mathcal{L}(\mathbf{x},\boldsymbol{\theta}) = \log \prod_n p(x_n|\boldsymbol{\theta}) = \sum_n \log p(x_n|\boldsymbol{\theta})$$

$$= \sum_n \log\left[\sum_{i=1}^{M} \alpha_i p_i(x_n|\boldsymbol{\theta}_i)\right]$$

difficult numerical treatment … however we can manage with a reinterpretation of the mixture model parameters …

α_k = probability of drawing the k-th component of the mixture model

⇨ new ("hidden" or "latent") variable: y = index of component (integer values only)

therefore, we **must** redefine data and parameters.

new likelihood which includes the hidden variables

$$
\begin{aligned}
\log \mathcal{L}'(\mathbf{x},\mathbf{y},\boldsymbol{\theta}) &= \log p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) \\
&= \log \prod_n p(x_n, y_n|\boldsymbol{\theta}) \\
&= \sum_n \log\left[p(x_n|y_n,\boldsymbol{\theta}) p(y_n|\boldsymbol{\theta})\right] \\
&= \sum_n \log\left[\alpha_{y_n} p_{y_n}(x_n|\boldsymbol{\theta}_{y_n})\right]
\end{aligned}
$$

($\boldsymbol{\theta}_i$ is the parameter vector restricted to the i-th component)

The structure is simpler now, there is no sum in the argument of the logarithm, however there are new hidden variables y.

Now we proceed by averaging the likelihood (Expectation step)

new parameter estimate

previous parameter estimate

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}\right)=E_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right]$$

$$=\int_{Y}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\right]p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right)d\mathbf{y}$$

$$\to\sum_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\right]p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right)$$

sum instead of integral, because the y variates are discrete

prior probabilities in the expression of the averaged log-likelihood

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}\right)=\sum_{\mathbf{y}}\left[\log p\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\theta}\right)\right]p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right)$$

and now we use Bayes' theorem:

$$p\left(y_n\middle|x_n,\boldsymbol{\theta}\right)=\frac{p\left(x_n\middle|y_n,\boldsymbol{\theta}\right)p\left(y_n\middle|\boldsymbol{\theta}\right)}{p\left(x_n\middle|\boldsymbol{\theta}\right)}=\frac{\alpha_{y_n}p_{y_n}\left(x_n\middle|\boldsymbol{\theta}_{y_n}\right)}{\sum_{k=1}^{M}\alpha_k p_k\left(x_n\middle|\boldsymbol{\theta}_k\right)}$$

$$p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}\right)=\prod_{n=1}^{N}p\left(y_n\middle|x_n,\boldsymbol{\theta}\right)=\prod_{n=1}^{N}\frac{\alpha_{y_n}p_{y_n}\left(x_n\middle|\boldsymbol{\theta}_{y_n}\right)}{\sum_{k=1}^{M}\alpha_k p_k\left(x_n\middle|\boldsymbol{\theta}_k\right)}$$

Therefore, using $\log \mathcal{L}'(\mathbf{x},\mathbf{y},\boldsymbol{\theta}) = \sum_n \log\left[\alpha_{y_n} p_{y_n}\left(x_n \middle| \boldsymbol{\theta}_{y_n}\right)\right]$

and $p\left(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}\right) = \prod_{n=1}^{N} p\left(y_n|x_n,\boldsymbol{\theta}\right)$

we find

$$
\begin{aligned}
Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}\right) &= \sum_{\mathbf{y}} \left[\log p\left(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}\right)\right] p\left(\mathbf{y}\middle|\mathbf{x},\boldsymbol{\theta}^{(i-1)}\right) \\
&= \sum_{\mathbf{y}} \sum_{k=1}^{N} \log\left[\alpha_{y_k} p_{y_k}\left(x_k\middle|\boldsymbol{\theta}_{y_k}\right)\right] \prod_{j=1}^{N} p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{k=1}^{N} \log\left[\alpha_{y_k} p_{y_k}\left(x_k\middle|\boldsymbol{\theta}_{y_k}\right)\right] \prod_{j=1}^{N} p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)
\end{aligned}
$$

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}\right)=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{k=1}^{N}\log\left[\alpha_{y_k}p_{y_k}\left(x_k\middle|\boldsymbol{\theta}_{y_k}\right)\right]\prod_{j=1}^{N}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)$$

$$=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{k=1}^{N}\boxed{\sum_{\ell=1}^{M}\delta_{\ell,y_k}}\log\left[\alpha_\ell p_\ell\left(x_k\middle|\boldsymbol{\theta}_\ell\right)\right]\prod_{j=1}^{N}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)$$

to decouple the variables, we add one sum and one Kronecker's delta...

after the decoupling, we can use the normalization of conditional probabilities

$$\sum_{y_j=1}^{M}p\left(y_j\middle|x_j,\boldsymbol{\theta}^{(i-1)}\right)=1$$

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}\right)=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{k=1}^{N}\sum_{\ell=1}^{M}\delta_{\ell,y_k}\log\left[\alpha_\ell p_\ell\left(x_k|\boldsymbol{\theta}_\ell\right)\right]\prod_{j=1}^{N}p\left(y_j|x_j,\boldsymbol{\theta}^{(i-1)}\right)$$

$$\Big\Downarrow$$

$$\begin{aligned}
Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i-1)}) &= \sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln\left[\alpha_\ell p_\ell(x_k|\boldsymbol{\theta}_\ell)\right]\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{\ell,y_k}\prod_{j=1}^{N}p\left(y_j|x_j,\boldsymbol{\theta}^{(i-1)}\right) \\
&= \sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln\left[\alpha_\ell p_\ell(x_k|\boldsymbol{\theta}_\ell)\right]\left\{\sum_{y_1=1}^{M}\cdots\sum_{y_{k-1}=1}^{M}\sum_{y_{k+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{\substack{j=1\\ j\neq k}}^{N}p\left(y_j|x_j,\boldsymbol{\theta}^{(i-1)}\right)\right\}p\left(\ell|x_k,\boldsymbol{\theta}^{(i-1)}\right) \\
&= \sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln\left[\alpha_\ell p_\ell(x_k|\boldsymbol{\theta}_\ell)\right]\left\{\prod_{\substack{j=1\\ j\neq k}}^{N}\sum_{y_j=1}^{M}p\left(y_j|x_j,\boldsymbol{\theta}^{(i-1)}\right)\right\}p\left(\ell|x_k,\boldsymbol{\theta}^{(i-1)}\right) \\
&= \sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln\left[\alpha_\ell p_\ell(x_k|\boldsymbol{\theta}_\ell)\right]p\left(\ell|x_k,\boldsymbol{\theta}^{(i-1)}\right)
\end{aligned}$$

these sums all add to 1 (normalization of conditional probabilities)

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}\right) &= \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln\left[\alpha_\ell p(\ell|x_k, \boldsymbol{\theta})\right] p_\ell(x_k, \boldsymbol{\theta}^{(i-1)}) \\
&= \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln \alpha_\ell \; p_\ell(x_k, \boldsymbol{\theta}^{(i-1)}) + \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln p(\ell|x_k, \boldsymbol{\theta}) \; p_\ell(x_k, \boldsymbol{\theta}^{(i-1)})
\end{aligned}
$$

this depends only on the α parameters

this term depends on the parameters of the component distributions

Thus, there are two terms that can be maximized separately.
Moreover, the first term must be maximized with the normalization constraint, i.e.

$$
\frac{\partial}{\partial \alpha_m}\left[\sum_{\ell=1}^{M} \sum_{k=1}^{N} \log \alpha_\ell p\left(\ell \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right) + \lambda\left(\sum_{\ell=1}^{M} \alpha_\ell - 1\right)\right] = 0
$$

$$\Big\Downarrow$$

$$
\sum_{k=1}^{N} \frac{1}{\alpha_m} p\left(m \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right) + \lambda = 0
$$

$$\sum_{k=1}^{N} \frac{1}{\alpha_m} p\left(m \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right) + \lambda = 0$$

$$\Downarrow$$

$$\sum_{k=1}^{N} p\left(m \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right) = -\lambda \alpha_m$$

$$\Downarrow$$

$$\sum_{m=1}^{M} \sum_{k=1}^{N} p\left(m \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right) = -\lambda \sum_{m=1}^{M} \alpha_m \quad \Rightarrow \quad \lambda = -N \quad \Rightarrow \quad \alpha_m = \frac{1}{N} \sum_{k=1}^{N} p\left(m \middle| x_k, \boldsymbol{\theta}^{(i-1)}\right)$$

This is as far as we can go without introducing an explicit form for the component distributions: to evaluate the other term we explicitly consider the 1D Gaussian mixture model:

$$p_\ell\left(x|\mu_\ell,\sigma_\ell\right)=\frac{1}{\sqrt{2\pi\sigma_\ell^2}}\exp\left(-\frac{\left(x-\mu_\ell\right)^2}{2\sigma_\ell^2}\right)$$

$$\sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln p_\ell(x_k,\boldsymbol{\theta})\;p(\ell|x_k,\boldsymbol{\theta}^{(i-1)})=\sum_{\ell=1}^{M}\sum_{k=1}^{N}\left[-\frac{1}{2}\ln(2\pi\sigma_\ell^2)-\frac{(x_k-\mu_\ell)^2}{2\sigma_\ell^2}\right]\;p(\ell|x_k,\mu_\ell^{(i-1)},\sigma_\ell^{(i-1)})$$

$$\frac{\partial}{\partial\mu_m}\sum_{\ell=1}^{M}\sum_{k=1}^{N}\ln p_\ell(x_k,\boldsymbol{\theta})\;p(\ell|x_k,\boldsymbol{\theta}^{(i-1)})=-2\sum_{k=1}^{N}\frac{(x_k-\mu_m)}{2\sigma_m^2}\;p(m|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)})=0$$

$$\frac{\partial}{\partial \mu_m} \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln p_\ell(x_k, \boldsymbol{\theta})\; p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) = -2 \sum_{k=1}^{N} \frac{(x_k - \mu_m)}{2\sigma_m^2}\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$

⇓

$$\mu_m = \frac{\sum_{k=1}^{N} x_k\, p\left(m \middle| x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}\right)}{\sum_{k=1}^{N} p\left(m \middle| x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}\right)}$$

moreover, if we let $c_m = 1/\sigma_m^2$

$$\begin{aligned}
\frac{\partial}{\partial c_m} \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln p_\ell(x_k, \boldsymbol{\theta})\; p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) &= \frac{\partial}{\partial c_m} \sum_{\ell=1}^{M} \sum_{k=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma_\ell^2) - \frac{(x_k - \mu_\ell)^2}{2\sigma_\ell^2} \right]\; p(\ell | x_k, \mu_\ell^{(i-1)}, \sigma_\ell^{(i-1)}) \\
&= \sum_{k=1}^{N} \left[\frac{1}{2c_m} - \frac{1}{2}(x_k - \mu_m)^2 \right]\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) \\
&= \sum_{k=1}^{N} \left[\frac{\sigma_m^2}{2} - \frac{1}{2}(x_k - \mu_m)^2 \right]\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0
\end{aligned}$$

$$\frac{\partial}{\partial c_m} \sum_{\ell=1}^{M} \sum_{k=1}^{N} \ln p_\ell(x_k, \boldsymbol{\theta})\; p(\ell | x_k, \boldsymbol{\theta}^{(i-1)}) = \sum_{k=1}^{N} \left[\frac{\sigma_m^2}{2} - \frac{1}{2}(x_k - \mu_m)^2 \right]\; p(m | x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}) = 0$$

⇓

$$\sigma_m^2 = \frac{\sum_{k=1}^{N} \left(x_k - \mu_m\right)^2 p\left(m \middle| x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}\right)}{\sum_{k=1}^{N} p\left(m \middle| x_k, \mu_m^{(i-1)}, \sigma_m^{(i-1)}\right)}$$

Finally, we find the following set of recursive formulas, that combine the E and M steps:

$$p_m\left(x\middle|\mu_m,\sigma_m\right)=\frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left(-\frac{\left(x-\mu_m\right)^2}{2\sigma_m^2}\right)$$

$$p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)=\frac{\alpha_m^{(i-1)}p_m\left(x_n\middle|\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}{\sum_{k=1}^{M}\alpha_m^{(i-1)}p_m\left(x_n\middle|\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}$$

$$\alpha_m^{(i)}=\frac{1}{N}\sum_{k=1}^{N}p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)$$

$$\mu_m^{(i)}=\frac{\sum_{k=1}^{N}x_k p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}{\sum_{k=1}^{N}p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)} \qquad \left(\sigma_m^{(i)}\right)^2=\frac{\sum_{k=1}^{N}\left(x_k-\mu_m^{(i)}\right)^2 p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}{\sum_{k=1}^{N}p\left(m\middle|x_k,\mu_m^{(i-1)},\sigma_m^{(i-1)}\right)}$$

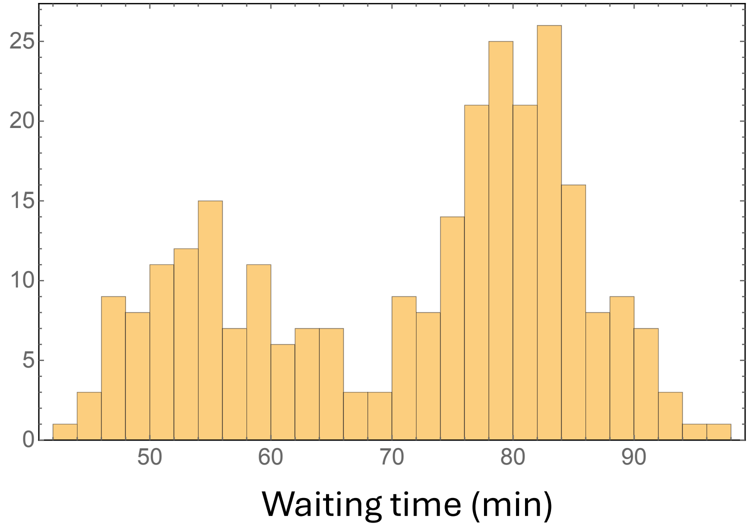
We remark that the probabilities

$$p\left(y_n \middle| x_n, \boldsymbol{\theta}\right) = \frac{\alpha_{y_n} p_{y_n}\left(x_n \middle| \boldsymbol{\theta}_{y_n}\right)}{\sum_{k=1}^{M} \alpha_k p_k\left(x_n \middle| \boldsymbol{\theta}_k\right)}$$

are an estimate of the frequencies of the y_n using the observed data x_n, and this amounts to a classification result (selection of one of the component distributions).

Straightforward example: waiting times between eruptions of the Old Faithful Geiser (Yellowstone National Park – Wyoming)

Here we analyze the waiting times assuming a 2-Gaussian mixture model for the waiting time distribution
(data taken from an R example)



In this case, the mixture model has two Gaussian components

$$p(w|\boldsymbol{\theta}) = \alpha N(w; \mu_1, \sigma_1) + (1-\alpha) N(w; \mu_2, \sigma_2)$$

where the vector of parameters is

$$\boldsymbol{\theta} = (\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$$

The resulting log likelihood with n waiting times w_i is

$$\ln \mathcal{L} = \sum_i \ln \left[\alpha N(w_i; \mu_1, \sigma_1) + (1-\alpha) N(w_i; \mu_2, \sigma_2)\right]$$

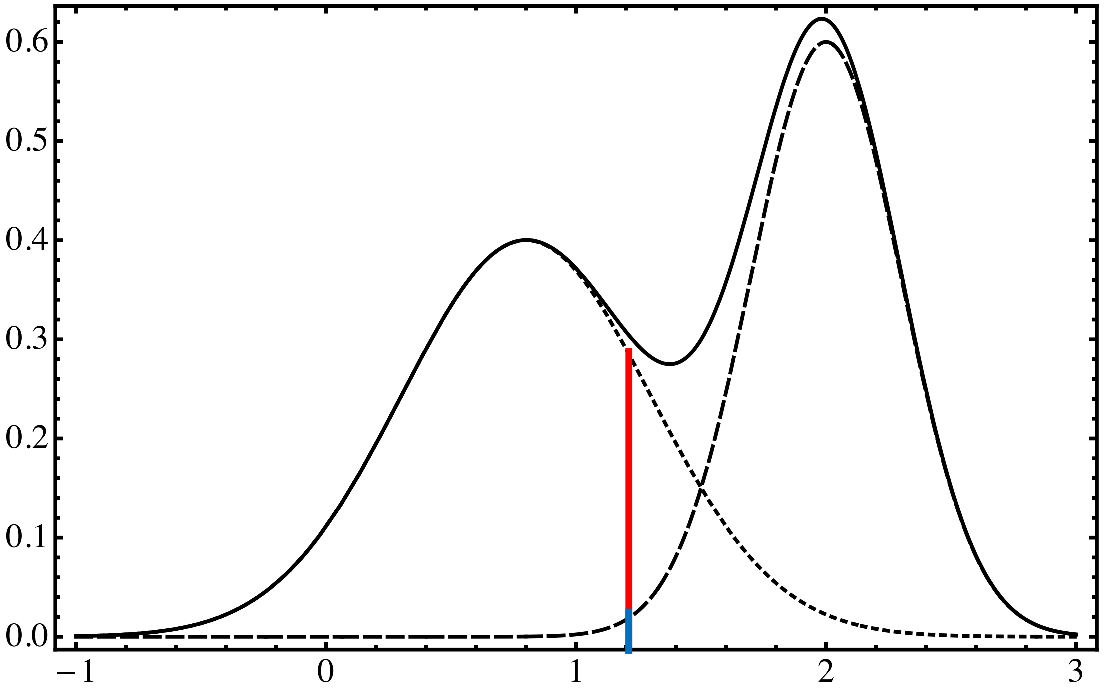
Again, we substitute the likelihood with the new one

$$\mathcal{L} = \prod_i \alpha^{y_i} (1-\alpha)^{1-y_i} [N(w_i; \mu_1, \sigma_1)]^{y_i} [N(w_i; \mu_2, \sigma_2)]^{1-y_i}$$

where the new, unobserved data y_i are indicator variables that select extraction from the first ($y_i = 1$) or the second ($y_i = 0$) Gaussian.

Then

$$\ln \mathcal{L} = \sum_i \left[y_i \ln \alpha + (1-y_i) \ln(1-\alpha) + y_i \left(-\frac{1}{2} \ln(2\pi\sigma_1) - \frac{(w_i - \mu_1)^2}{2\sigma_1^2} \right) \right.$$
$$\left. +(1-y_i) \left(-\frac{1}{2} \ln(2\pi\sigma_2) - \frac{(w_i - \mu_2)^2}{2\sigma_2^2} \right) \right]$$



The probability that a given time interval belongs to the first Gaussian is

this probability is also equal to the mean value of the indicator variable

$$
\begin{aligned}
p_i &= \frac{\alpha \times N(w_i; \mu_1, \sigma_1)}{\alpha \times N(w_i; \mu_1, \sigma_1) + (1-\alpha) \times N(w_i; \mu_2, \sigma_2)} \\
&= \frac{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2 / 2(\sigma_1^{(k)})^2] / \sqrt{2\pi(\sigma_1^{(k)})^2}}{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2 / 2(\sigma_1^{(k)})^2] / \sqrt{2\pi(\sigma_1^{(k)})^2} + (1-\alpha^{(k)}) \exp[-(w_i - \mu_2^{(k)})^2 / 2(\sigma_2^{(k)})^2] / \sqrt{2\pi(\sigma_2^{(k)})^2}}
\end{aligned}
$$

Now, averaging the log likelihood with respect to the missing data we find

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_i \left[p_i^{(k)} \ln \alpha + (1 - p_i^{(k)}) \ln(1-\alpha) + p_i^{(k)} \left(-\frac{1}{2} \ln(2\pi\sigma_1^2) - \frac{(w_i - \mu_1)^2}{2\sigma_1^2} \right) \right.$$
$$\left. + (1 - p_i^{(k)}) \left(-\frac{1}{2} \ln(2\pi\sigma_2^2) - \frac{(w_i - \mu_2)^2}{2\sigma_2^2} \right) \right]$$

(the mean value of the indicator variable is equal to the current estimate probability α)

Next, we maximize with respect to all the remaining parameters, and we find:

$$\alpha^{(k+1)} = \frac{1}{N} \sum_i p_i^{(k)}$$

$$\left(\sigma_1^{(k+1)}\right)^2 = \frac{\sum_i p_i^{(k)} (w_i - \mu_1^{(k)})^2}{\sum_i p_i^{(k)}}; \qquad \mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

$$\left(\sigma_2^{(k+1)}\right)^2 = \frac{\sum_i (1 - p_i^{(k)}) (w_i - \mu_2^{(k)})^2}{\sum_i (1 - p_i^{(k)})}; \qquad \mu_2^{(k+1)} = \frac{\sum_i (1 - p_i^{(k)}) w_i}{\sum_i (1 - p_i^{(k)})}$$

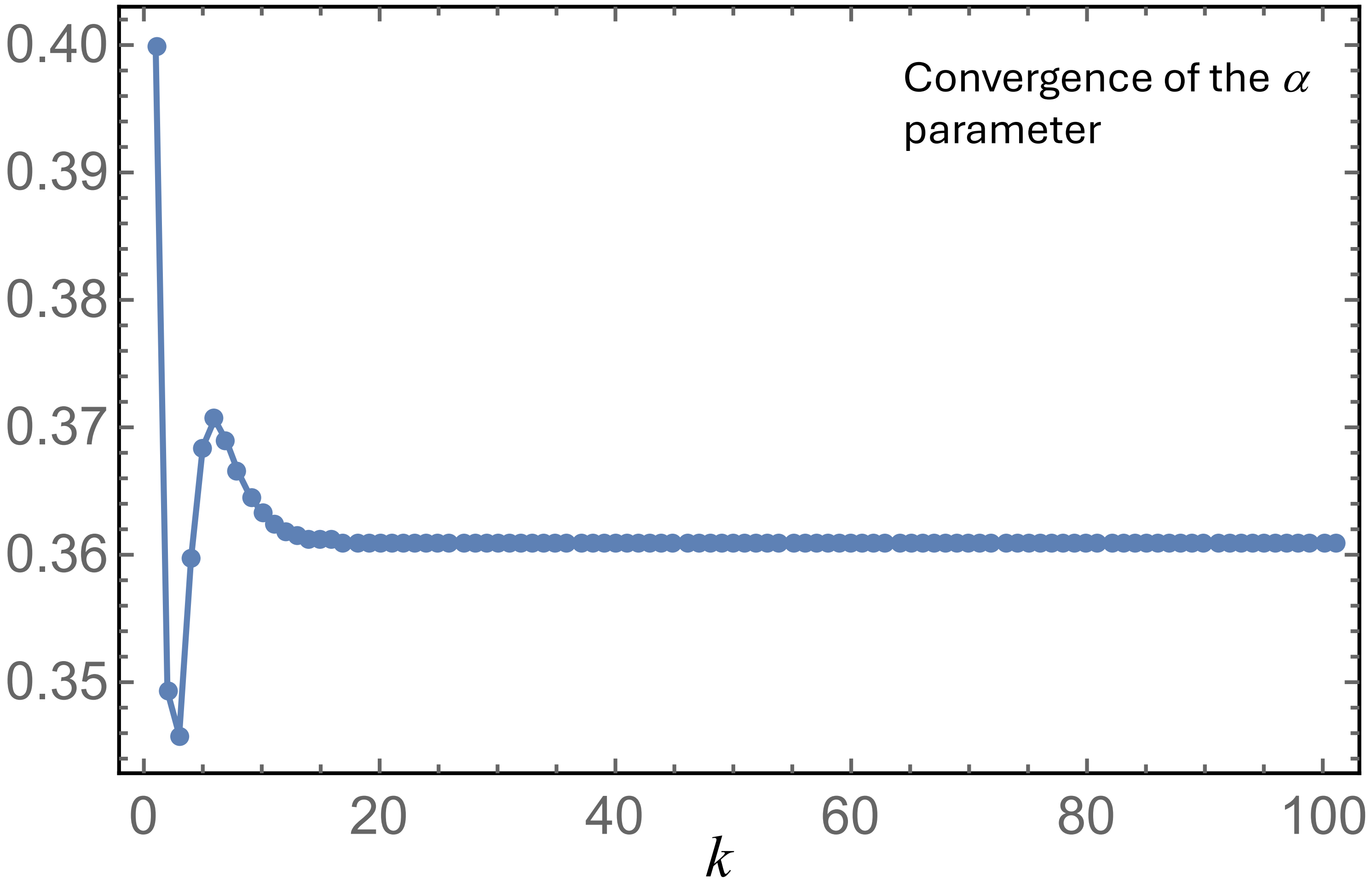
Finally, we find the following set of equations:

$$p_i^{(k)} = \frac{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2}}{\alpha^{(k)} \exp[-(w_i - \mu_1^{(k)})^2/2(\sigma_1^{(k)})^2]/\sqrt{2\pi(\sigma_1^{(k)})^2} + (1-\alpha^{(k)}) \exp[-(w_i - \mu_2^{(k)})^2/2(\sigma_2^{(k)})^2]/\sqrt{2\pi(\sigma_2^{(k)})^2}}$$

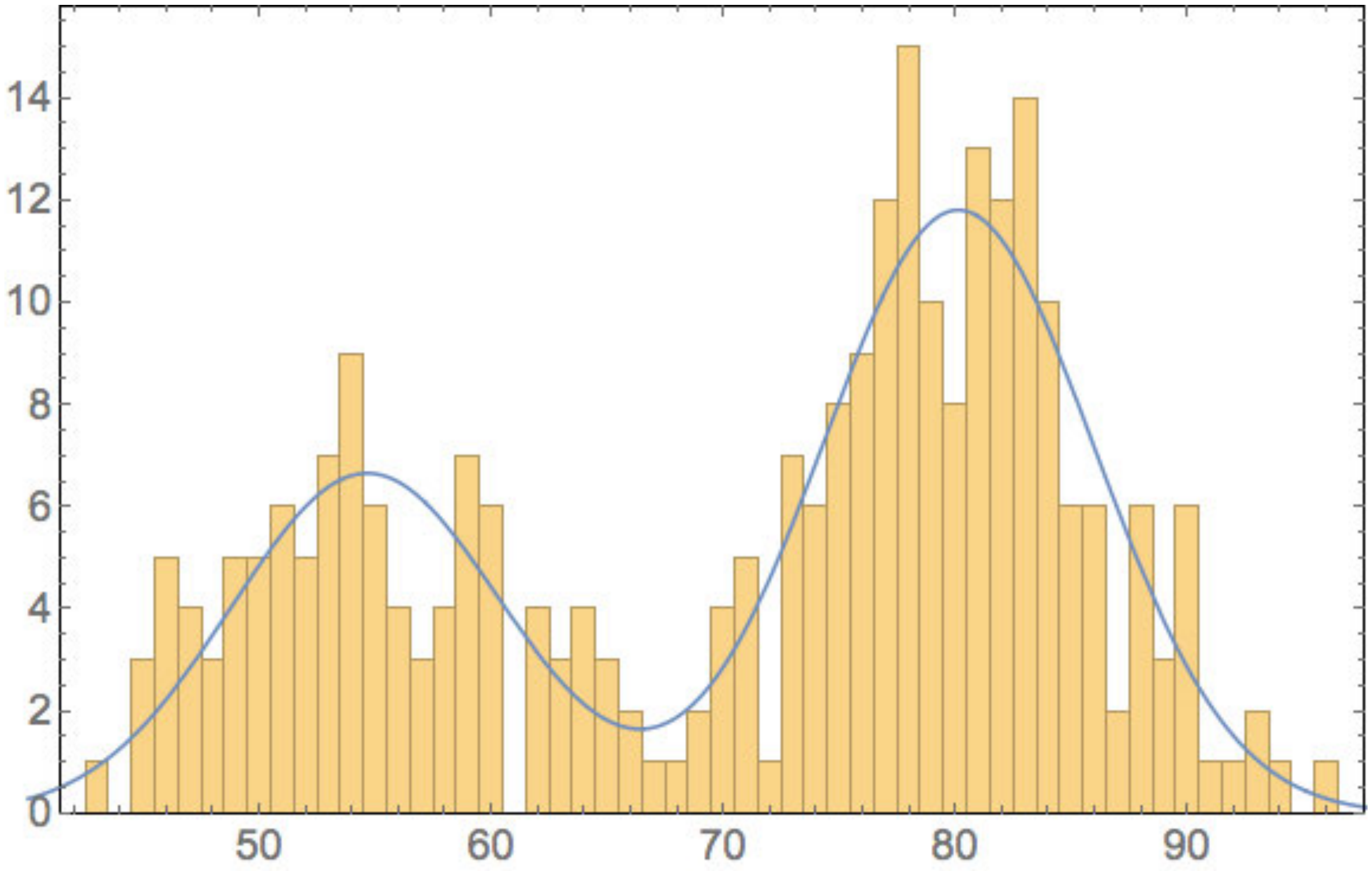
$$\alpha^{(k+1)} = \frac{1}{N}\sum_i p_i^{(k)}$$

$$\left(\sigma_1^{(k+1)}\right)^2 = \frac{\sum_i p_i^{(k)}(w_i - \mu_1^{(k)})^2}{\sum_i p_i^{(k)}}; \qquad \mu_1^{(k+1)} = \frac{\sum_i p_i^{(k)} w_i}{\sum_i p_i^{(k)}}$$

$$\left(\sigma_2^{(k+1)}\right)^2 = \frac{\sum_i (1 - p_i^{(k)})(w_i - \mu_2^{(k)})^2}{\sum_i (1 - p_i^{(k)})}; \qquad \mu_2^{(k+1)} = \frac{\sum_i (1 - p_i^{(k)}) w_i}{\sum_i (1 - p_i^{(k)})}$$


0.40
0.39
0.38
0.37
0.36
0.35
0
20
40
60
80
100
k
Convergence of the α parameter

Comparison of the original data with the mixture model obtained with the EM algorithm



Waiting time (min)