18

Bayesian Estimation

In this chapter we will review the subject of Bayesian estimation, with a particular focus on matrix estimation. The general situation one encounters is one where the observed matrix is a noisy version of the "true" matrix one wants to estimate. For example, in the case of additive noise, one observes a matrix **E** which is the true matrix **C** plus a random matrix **X** that plays the role of noise, to wit,

$$\mathbf{E} = \mathbf{C} + \mathbf{X}. \tag{18.1}$$

In the case of multiplicative noise, the observed matrix **E** has the form

$$\mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}\mathbf{C}^{\frac{1}{2}}. \tag{18.2}$$

When **W** is a white Wishart matrix, this is the problem of sample covariance matrix encountered in Chapter 17.

In general, the true matrix **C** is unknown to us. We would like to know the probability of **C** given that we have observed **E**, i.e. compute $P(\mathbf{C}|\mathbf{E})$. This is the general subject of Bayesian estimation, which we introduce and discuss in this chapter.

18.1 Bayesian Estimation

Before doing Bayesian theory on random matrices (see Section 18.3), we first review Bayesian estimation and see it at work on simpler examples.

18.1.1 General Framework

Imagine we have an observable variable y that we would like to infer from the observation of a related variable x. The variables x and y can be scalars, vectors, matrices, higher dimensional objects . . . We postulate that we know the random process that generates y given x, i.e. y could be a noisy version of x or more generally y could be drawn from a known distribution with x as a parameter. The generation process of y is encoded in a probability distribution $P(y|x)$, which is called the *sampling distribution* or the *likelihood function*.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

Given our knowledge of $P(y|x)$, we would like to write the inference probability $P(x|y)$, also called the *posterior distribution*. To do so, we can use Bayes' rule:

$$P(x|y) = \frac{P(y|x)P_0(x)}{P(y)}. \tag{18.3}$$

To obtain the desired probability, Bayes' rule tells us that we need to know the *prior distribution* $P_0(x)$. In theory $P_0(x)$ is the distribution from which x is drawn and it is in some cases knowable. In many practical applications, however, x is actually not random but simply unknown and $P_0(x)$ encodes our ignorance of x. It should represent our best (probabilistic) guess of x before we observe the data y. The determination (or arbitrariness) of the prior $P_0(x)$ is considered to be one of the weak points of the Bayesian approach. Often $P_0(x)$ is just taken to be constant, i.e. no prior knowledge at all on x. However, note that $P_0(x) =$ constant is not invariant upon changes of variables, for if $x' = f(x)$ is a non-linear transformation of x, then $P_0(x')$ is no longer constant! In Section 18.1.3, we will see how the arbitrariness in the choice of $P_0(x)$ can be used to simplify modeling.

The other distribution appearing in Bayes' rule $P(y)$ is actually just a normalization factor. Indeed, y is assumed to be known, therefore $P(y)$ is just a fixed number that can be computed by normalizing the posterior distribution. One therefore often simplifies Bayes' rule as

$$P(x|y) = \frac{1}{Z}P(y|x)P_0(x), \qquad Z := \int \mathrm{d}x\, P(y|x)P_0(x), \tag{18.4}$$

where $P(y|x)$ represents the measurement (or noise) process and $P_0(x)$ the (often arbitrary) prior distribution.

From the posterior distribution $P(x|y)$ we can build an estimator of x. The optimal estimator depends on the problem at hand, namely, which quantity are we trying to optimize. The most common Bayesian estimators are

1 MMSE: The posterior mean $\mathbb{E}[x]_y$. It minimizes a quadratic loss function and is hence called the Minimum Mean Square Error estimator.

2 MAVE: The posterior median or Minimum Absolute Value Error estimator.

3 MAP: The Maximum *A Posteriori* estimator, defined as $\hat{x} = \operatorname{argmax}_x P(x|y)$.

18.1.2 A Simple Estimation Problem

Consider the simplest one-dimensional estimation problem:

$$y = x + \varepsilon, \tag{18.5}$$

where x is some signal to be estimated, ε is an independent noise, and y is the observation. Then $P(y|x)$ is simply $P_\varepsilon(.)$ evaluated at $y - x$:

$$P(y|x) = P_\varepsilon(y - x). \tag{18.6}$$

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

Suppose further that ε is a centered Gaussian noise with variance $\sigma_{\rm n}^2$, where the subscript n means “noise”. Then we have

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma_{\rm n}^2}} \exp\left(-\frac{(y-x)^2}{2\sigma_{\rm n}^2}\right). \tag{18.7}$$

Then we get that

$$P(x|y) \propto P_0(x) \exp\left(\frac{2xy - x^2}{2\sigma_{\rm n}^2}\right), \tag{18.8}$$

where $P_0(x)$ is the prior distribution of x and we have dropped x-independent factors. Depending on the choice of $P_0(x)$ we will get different posterior distributions and hence different estimators of x.

Gaussian Prior

Suppose first $P_0(x)$ is a Gaussian with variance $\sigma_{\rm s}^2$ (for signal) centered at x_0. Then

$$\begin{aligned} P(x|y) &\propto \exp\left(-\frac{(x-x_0)^2}{2\sigma_{\rm s}^2} + \frac{2xy - x^2}{2\sigma_{\rm n}^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\hat{x})^2}{2\sigma^2}\right), \end{aligned} \tag{18.9}$$

with

$$\hat{x} := x_0 + r(y - x_0) = (1-r)x_0 + ry; \qquad \sigma^2 := r\sigma_{\rm n}^2, \tag{18.10}$$

where the signal-to-noise ratio r is $r = \sigma_{\rm s}^2/(\sigma_{\rm s}^2 + \sigma_{\rm n}^2)$. The posterior distribution is thus a Gaussian centered around $\hat{x}$ and of variance σ^2.

For a Gaussian distribution the mean, median and maximum probability values are all equal to $\hat{x}$, which is therefore the optimal estimator in all three standard procedures, MMSE, MAVE and MAP. This estimator is called the *linear shrinkage estimator* as it is linear in the observed variable y. The linear coefficient of y is the signal-to-noise ratio r, a number smaller than one that *shrinks* the observed value towards the *a priori* mean x_0.

Note that this estimator can also be obtained in a completely different framework: it is the affine estimator that minimizes the mean square error. The estimator is affine by construction and minimization only involves first and second moments; it is therefore not too surprising that we recover Eq. (18.10), see Exercise 18.1.2. As so often in optimization problems, assuming Gaussian fluctuations is equivalent to imposing an affine solution.

Another important property of the linear shrinkage estimator is that it is rather conservative: it is biased towards x_0. By assumption x fluctuates with variance $\sigma_{\rm s}^2$ and y fluctuates with variance $\sigma_{\rm s}^2 + \sigma_{\rm n}^2$. This allows us to compute the variance of the estimator $\hat{x}(y)$ as

$$\mathbb{V}[\hat{x}(y)] = r^2(\sigma_{\rm s}^2 + \sigma_{\rm n}^2) = \frac{\sigma_{\rm s}^4}{\sigma_{\rm s}^2 + \sigma_{\rm n}^2} \leq \sigma_{\rm s}^2. \tag{18.11}$$

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

So the variance of the estimator[1] is not only smaller than that of the observed variable y it is also smaller than the fluctuations of the true variable x!

Exercise 18.1.1 Optimal affine estimator

Suppose that we observe a variable y that has some non-zero covariance with an unknown variable x that we would like to estimate. We will show that the best affine estimator of x is given by the linear shrinkage estimator (18.10). The variables x and y can be drawn from any distribution with finite variance. We write the general affine estimator

$$\hat{x} = ay + b, \tag{18.12}$$

and choose a and b to minimize the expected mean square error.

(a) Initially assume that x and y have zero mean – we will relax this assumption later. Show that

$$\mathbb{E}\left[(x - \hat{x})^2\right] = a^2\sigma_y^2 + b^2 + \sigma_x^2 - 2a\sigma_{xy}^2, \tag{18.13}$$

where σ_x^2, σ_y^2 and σ_{xy}^2 are the variances of x, y and their covariance.

(b) Show that the optimal estimator has $a = \sigma_{xy}^2/\sigma_y^2$ and $b = 0$.

(c) Compute b in the non-zero mean case by considering $x - x_0$ estimated using $y - y_0$.

(d) Compute σ_y^2 and σ_{xy}^2 when $y = x + \varepsilon$ with ε independent of x.

(e) Show that when $\mathbb{E}[\varepsilon] = 0$ we recover Eq. (18.10).

Bernoulli Prior

When $P_0(x)$ is non-Gaussian, the obtained estimators are in general non-linear. As a second example suppose that $P_0(x)$ is Bernoulli random variable with $P_0(x = 1) = P_0(x = -1) = 1/2$. Then, after a few simple manipulations one obtains

$$P(x|y) = \frac{1}{2}\left(\left(1 + \tanh\left(\frac{y}{\sigma_{\text{n}}^2}\right)\right)\delta_{x,1} + \left(1 - \tanh\left(\frac{y}{\sigma_{\text{n}}^2}\right)\right)\delta_{x,-1}\right). \tag{18.14}$$

The posterior distribution is now a discrete function that takes on only two values, namely ±1. In this case the maximum probability and the median are such that

$$\hat{x}_{\text{MAP}}(y) = \text{sign}(y). \tag{18.15}$$

[1] One should not confuse the variance of the posterior distribution $r\sigma_{\text{n}}^2$ with the variance of the estimator $r\sigma_{\text{s}}^2$. The first one measures the remaining uncertainty about x once we have observed y while the second measures the variability of $\hat{x}(y)$ when we repeat the experiment multiple times with varying x and noise ε.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

It is also easy to calculate the MMSE estimator:

$$\hat{x}_{\text{MMSE}}(y) = \mathbb{E}[x]_y = \tanh\left(\frac{y}{\sigma_{\text{n}}^2}\right). \tag{18.16}$$

It may seem odd that the MMSE estimator takes continuous values between -1 and 1 while we postulated that the true x can only be equal to ± 1. Nevertheless, in order to minimize the variance it is optimal to shoot somewhere in the middle of -1 and 1 as choosing the wrong sign costs a lot in terms of variance. The estimator $\hat{x}(y)$ is biased, i.e. $\mathbb{E}[\hat{x}_{\text{MMSE}}|x] \neq x$. It also has a variance strictly less than 1, whereas the variance of the true x is unity.

Laplace Prior

As a third example, consider a Laplace distribution

$$P_0(x) = \frac{b}{2}\mathrm{e}^{-b|x|} \tag{18.17}$$

for the prior, with variance $2b^{-2}$. In this case the posterior distribution is given by

$$P(x|y) \propto \exp\left(-b|x| + \frac{2xy - x^2}{2\sigma_{\text{n}}^2}\right). \tag{18.18}$$

The MMSE and MAVE estimators can be computed but the results are not very enlightening as they are given by an ugly combination of error functions and even inverse error functions (for MAVE). The MAP estimator is both simpler and more interesting in this case. It is given by

$$\hat{x}_{\text{MAP}}(y) = \begin{cases} 0 & \text{for } |y| < b\sigma_{\text{n}}^2, \\ y - b\sigma_{\text{n}}^2 \operatorname{sign}(y) & \text{otherwise.} \end{cases} \tag{18.19}$$

The MAP estimator is sparse in the sense that in a non-zero fraction of cases it takes the exact value of zero. Note that the true variable x itself is not sparse: it is almost surely non-zero. This example is a toy-model for the “LASSO” regularization that we will study in Section 18.2.2.

Non-Gaussian Noise

The noise in Eq. (18.5) can also be non-Gaussian. When the noise has fat tails, one can even be in the counter-intuitive situation where the estimator is not monotonic in the observed variable, i.e. the best estimate of x decreases as a function of its noisy version y. For example, if x is centered unit Gaussian and ε is a centered unit Cauchy noise, we have

$$P(x|y) \propto \frac{\mathrm{e}^{-x^2/2}}{(y-x)^2 + 1}. \tag{18.20}$$

Whereas the Cauchy noise ε and the observation y do not have a first moment, the posterior distribution of x is regularized by the Gaussian weight and all its moments

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

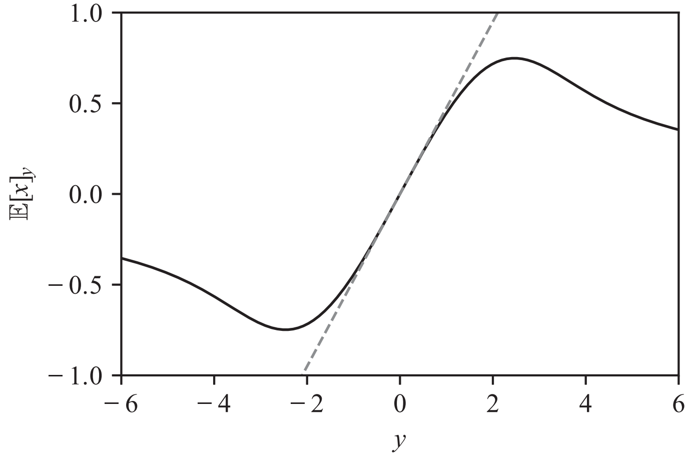

Figure 18.1 A non-monotonic optimal estimator. The MMSE estimator of a Gaussian variable corrupted by Cauchy noise (see Eq. (18.21)). For small absolute observations y, the estimator is almost linear with slope $2 - \sqrt{2/e\pi}/\text{erfc}(1/\sqrt{2}) \approx 0.475$ (dashed line).

are finite. After some tedious calculation we arrive at the conditional mean or MMSE estimator:

$$\mathbb{E}[x]_y = y + \frac{\text{Im}(\Phi)}{\text{Re}(\Phi)}, \quad \text{where} \quad \Phi = e^{iy}\text{erfc}\left(\frac{1+iy}{\sqrt{2}}\right). \tag{18.21}$$

The shape of the estimator as a function of y is not obvious from this expression but it is plotted numerically in Figure 18.1. The interpretation is the following:

- When we observe a small (order 1) value of y, we can assume that it was generated by a moderate x with moderate noise, hence we are in the regime of the linear estimator with a signal-to-noise ratio close to one-half ($\hat{x} \approx 0.475y$).
- On the other hand, when y is much larger than the standard deviation of x it becomes clear that y can only be large because the noise takes extreme values. When the noise is large our knowledge of x decreases, hence the estimator tends to zero as $|y| \to \infty$.

18.1.3 Conjugate Priors

The main weakness of Bayesian estimation is the reliance on a prior distribution for the variable we want to estimate. In many practical applications one does not have a probabilistic or statistical knowledge of $P_0(x)$. The variable x is a fixed quantity that we do not know, so how are we supposed to know about $P_0(x)$? In such cases we are left with making a reasonable practical guess. Since $P_0(x)$ is just a guess, we can at least choose a functional form for $P_0(x)$ that makes computation easy. This is the idea behind "conjugate priors".

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

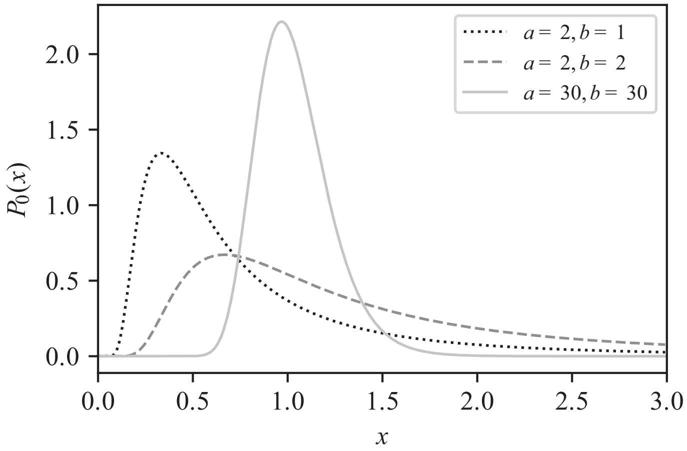

Figure 18.2 The inverse-gamma distribution Eq. (18.24). Its mean is given by $b/(a-1)$ and it becomes increasingly peaked around this mean as both a and b become large.

When we studied the one-dimensional estimation of a variable x corrupted by additive Gaussian noise (Eq. (18.5), with Gaussian ε) we found that choosing a Gaussian prior for x gave us a Gaussian posterior distribution. In many other cases, we can find a family of prior distributions that will similarly keep the posterior distribution in the same family. This concept is better explained in an example.

Imagine we are given a series of T numbers $\{y_i\}$ generated independently from a centered Gaussian distribution of variance c that is unknown to us. We use the variable c rather than σ^2 to avoid the confusion between the estimation of σ and that of $c = \sigma^2$. The joint probability of the $\mathbf{y}$'s is given by

$$P(\mathbf{y}|c) = \frac{1}{(2\pi c)^{T/2}} \exp\left(-\frac{\mathbf{y}^T\mathbf{y}}{2c}\right). \tag{18.22}$$

The posterior distribution is thus given by

$$P(c|\mathbf{y}) \propto P_0(c) c^{-T/2} \exp\left(-\frac{\mathbf{y}^T\mathbf{y}}{2c}\right). \tag{18.23}$$

Now if the prior $P_0(c)$ has the form $P_0(c) \propto c^{-a-1}\mathrm{e}^{-b/c}$, the posterior will also be of that form with modified values for a and b. Such a $P_0(c)$ will thus be our conjugate prior. This law is precisely the inverse-gamma distribution (see Fig. 18.2):

$$P_0(c) = \frac{b^a}{\Gamma(a)} c^{-a-1} \mathrm{e}^{-b/c} \qquad (c \geq 0). \tag{18.24}$$

It describes a non-negative variable, as a variance should. It is properly normalized when $a > 0$ and has mean $b/(a-1)$ whenever $a > 1$. If we choose such a law as our variance

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

prior, the posterior distribution after having observed the vector $\mathbf{y}$ is also an inverse-gamma with parameters

$$a_\mathrm{p} = a + \frac{T}{2} \quad \text{and} \quad b_\mathrm{p} = b + \frac{\mathbf{y}^T\mathbf{y}}{2}. \tag{18.25}$$

The MMSE estimator can then just be read off from the mean of an inverse-gamma distribution:

$$\mathbb{E}[c]_\mathbf{y} = \frac{b_\mathrm{p}}{a_\mathrm{p} - 1} = \frac{2b + \mathbf{y}^T\mathbf{y}}{2(a-1) + T}, \tag{18.26}$$

which can be written explicitly in the form of a linear shrinkage estimator:

$$\mathbb{E}[c]_\mathbf{y} = (1-r)c_0 + r\frac{\mathbf{y}^T\mathbf{y}}{T} \quad \text{with} \quad r = \frac{T}{2(a-1)+T}, \tag{18.27}$$

and $c_0 = b/(a-1)$ is the mean of the prior. We see that $r \to 1$ when $T \to \infty$: in this case the prior guess on c_0 disappears and one is left with the naive empirical estimator $\mathbf{y}^T\mathbf{y}/T$.

Exercise 18.1.2 Conjugate prior for the amplitude of a Laplace distribution

Suppose that we observe T variables y_i drawn from a Laplace distribution (18.17) with unknown amplitude b. We would like to estimate b using the Bayesian method with conjugate prior.

(a) Write the joint probability density of elements of the vector $\mathbf{y}$ for a given b. This is the likelihood function $P(\mathbf{y}|b)$.

(b) As a function of b, the likelihood function has the same form as a gamma distribution (4.17). Using a gamma distribution with parameters a_0 and b_0 for the prior on b show that the posterior distribution of b is also a gamma distribution. Find the posterior parameters a_p and b_p.

(c) Given that the mean of a gamma distribution is given by a/b, write the MMSE estimator in this case.

(d) Compute the estimator in the two limiting cases $T = 0$ and $T \to \infty$.

(e) Write your estimator from (c) as a shrinkage estimator interpolating between these two limits. Show that the signal-to-noise ratio r is given by $r = Tm/(Tm + 2b_0)$ where $m = \sum |y_i|/T$. Note that in this case the shrinkage estimator is non-linear in the naive estimate $\hat{b} = 1/(2m)$.

18.2 Estimating a Vector: Ridge and LASSO

A very standard problem for which Bayesian ideas are helpful is linear regression. Assume we want to estimate the parameters a_i of a multi-linear regression, where we assume that an observable y can be written as

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

$$y = \sum_{i=1}^{N} a_i x_i + \varepsilon, \tag{18.28}$$

where x_i are N observable quantities and ε is noise (not directly observable). We observe a time series of y of length T that we stack into a vector $\mathbf{y}$, whereas the different x_i are stacked into an $N \times T$ data matrix $\mathbf{H}_{it} = x_i^t$, and ε is the corresponding T-dimensional noise vector. We thus write

$$\mathbf{y} = \mathbf{H}^T \mathbf{a} + \varepsilon, \tag{18.29}$$

where $\mathbf{a}$ is an N-dimensional vector of coefficients we want to estimate. We assume the following structure for the random variables x and ε:

$$\frac{1}{T}\mathbb{E}[\varepsilon\varepsilon^T] = \sigma_{\mathrm{n}}^2 \mathbf{1}; \qquad \frac{1}{T}\mathbb{E}[\mathbf{H}\mathbf{H}^T] = \mathbf{C}, \tag{18.30}$$

where $\mathbf{C}$ can be an arbitrary covariance matrix, but we will assume it to be the identity $\mathbf{1}$ in the following, unless otherwise stated.

Classical linear regression would find the coefficient vector $\mathbf{a}$ that minimizes the error $\mathcal{E} = \|\mathbf{y} - \mathbf{H}^T\mathbf{a}\|^2$ on a given dataset. As is well known, the regression coefficients are given by

$$\mathbf{a}_{\mathrm{reg}} = \left(\mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{H}\mathbf{y}. \tag{18.31}$$

This equation can be derived easily by taking the derivatives of $\mathcal{E}$ with respect to all a_i and setting them to zero. Note that when $q := N/T < 1$, $\mathbf{H}\mathbf{H}^T$ is in general invertible, but when $q \geq 1$ (i.e. when there is not enough data), Eq. (18.31) is *a priori* ill defined.

In a Bayesian estimation framework, we want to write the posterior distribution $P(\mathbf{a}|\mathbf{y})$ and build an estimator of $\mathbf{a}$ from it. We expect that the Bayesian approach will work better than linear regression "out of sample", i.e. on a new independent sample. The reason is that the linear regression method minimizes an "in-sample" error, and is thus devised to fit best the details of the observed dataset, with no regard to overfitting considerations. These concepts will be clarified in Section 18.2.3.

Following the approach of Section 18.1, we write the posterior distribution as

$$P(\mathbf{a}|\mathbf{y}) \propto P_0(\mathbf{a}) \exp\left(-\frac{1}{2\sigma_{\mathrm{n}}^2} \left\|\mathbf{y} - \mathbf{H}^T\mathbf{a}\right\|^2\right), \tag{18.32}$$

where σ_{n}^2 is the variance of the noise ε. Now, the art is to choose an adequate prior distribution $P_0(\mathbf{a})$.

18.2.1 Ridge Regression

The likelihood function in Eq. (18.32) is a Gaussian function of $\mathbf{a}$, so choosing a Gaussian prior for $P_0(\mathbf{a})$ will give us a Gaussian posterior. To construct a Gaussian distribution for $P_0(\mathbf{a})$ we need to choose a prior mean $\mathbf{a}_0$ and a prior covariance matrix.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

Regression coefficients can be positive or negative, so the most natural prior mean is the zero vector $\mathbf{a}_0 = \mathbf{0}$. In the absence of any other information about the direction in which $\mathbf{a}$ may point, we should make a rotationally invariant prior for the covariance matrix.[2] The only rotationally invariant choice is a multiple of the identity $\sigma_{\mathrm{s}}^2\mathbf{1}$ for the prior covariance. Assuming that the coefficients a_i are IID gives the same answer. However, we do not have a good argument to set the scale of the covariance σ_{s}^2; we will come back to this point later.

The posterior distribution is then written

$$P(\mathbf{a}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma_{\mathrm{n}}^2}\left(\mathbf{a}^T\left(\mathbf{H}\mathbf{H}^T + \frac{\sigma_{\mathrm{n}}^2}{\sigma_{\mathrm{s}}^2}\mathbf{1}\right)\mathbf{a} - 2\mathbf{a}^T\mathbf{H}\mathbf{y}\right)\right). \tag{18.33}$$

As announced, the posterior is a multivariate Gaussian distribution. The MMSE, MAVE and MAP estimator are all equal to the mode of the distribution, given by[3]

$$\mathbb{E}[\mathbf{a}]_{\mathbf{y}} = \left(\frac{\mathbf{H}\mathbf{H}^T}{T} + \zeta\mathbf{1}\right)^{-1}\frac{\mathbf{H}\mathbf{y}}{T}, \qquad \zeta := \frac{\sigma_{\mathrm{n}}^2}{T\sigma_{\mathrm{s}}^2}. \tag{18.34}$$

This is called the “ridge” regression estimator, as it amounts to adding weight on the diagonal of the sample covariance matrix $(\mathbf{H}\mathbf{H}^T)/T$. This can also be seen as a shrinkage of the covariance matrix towards the identity, as we will discuss further in Section 18.3 below.

Another way to understand what ridge regression means is to notice that Eq. (18.31) involves the inverse of the covariance matrix $(\mathbf{H}\mathbf{H}^T)/T$, which can be unstable in large dimensions. This instability can lead to very large coefficients in $\mathbf{a}$. One can thus regularize the regression problem by adding a quadratic (or L^2-norm) penalty for $\mathbf{a}$ so the vector does not become too big:

$$\mathbf{a}_{\text{ridge}} = \underset{\mathbf{a}}{\operatorname{argmin}}\left[\left\|\mathbf{y} - \mathbf{H}^T\mathbf{a}\right\|^2 + T\zeta\,\|\mathbf{a}\|^2\right]. \tag{18.35}$$

Setting $\zeta = 0$ we recover the standard regression. The solution of the regularized optimization problem yields exactly Eq. (18.34); it is often called the Tikhonov regularization. Note that the resulting equation for $\mathbf{a}_{\text{ridge}}$ remains well defined even when $q \geq 1$ as long as $\zeta > 0$.

In both approaches (Bayesian and Tikhonov regularization) the result depends on the choice of the parameter $\zeta = \sigma_{\mathrm{n}}^2/(T\sigma_{\mathrm{s}}^2)$ which is hard to estimate *a priori*. The modern way of fixing ζ in practical applications is by using a validation (or cross-validation) method. The idea is to find the value of $\mathbf{a}_{\text{ridge}}$ on part of the data (the “training set”) and measure the quality of the regression on another, non-overlapping part of the data (the “validation set”). The value of ζ is then chosen as the one that gives the lowest error on the validation set.

[2] This assumption relies on our hypothesis that the covariance matrix $\mathbf{C}$ of the x’s is the identity matrix. Otherwise, the eigenvectors of $\mathbf{C}$ could be used to construct non-rotationally invariant priors.

[3] We have introduced a factor of $1/T$ in the definition of ζ so it parameterizes the shift in the *normalized* covariance matrix $(\mathbf{H}\mathbf{H}^T)/T$. It turns out to be the proper scaling in the large N limit with $q = N/T$ fixed. Note that if the elements of $\mathbf{a}$ and $\mathbf{H}$ are of order one, the variance of the elements of $\mathbf{H}^T\mathbf{a}$ is of order N; for the noise to contribute significantly in the large N limit we must have σ_{n}^2 of order N and hence ζ of order 1.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

In cross-validation, the procedure is repeated with multiple validation sets (always disjoint from the training set) and the error is then averaged over these sets.

18.2.2 LASSO

Another common estimating method for vectors is the "LASSO" method[4] which combines a Laplace prior with the MAP estimator.

In this method, the prior distribution amounts to assuming that the coefficients of **a** are IID Laplace random number with variance $2b^{-2}$. The posterior then becomes

$$P(\mathbf{a}|\mathbf{y}) \propto \exp\left(-b\sum_{i=1}^{N}|a_i| - \frac{1}{2\sigma_{\mathrm{n}}^2}\left\|\mathbf{y}-\mathbf{H}^T\mathbf{a}\right\|^2\right).$$

As in the toy model Eq. (18.18), the MMSE and MAVE estimators look rather ugly, but the MAP one is quite simple. It is given by the maximum of the argument of the above exponential:

$$\mathbf{a}_{\text{LASSO}} = \underset{\mathbf{a}}{\operatorname{argmin}}\left[2b\sigma_{\mathrm{n}}^2\sum_{i=1}^{N}|a_i| + \left\|\mathbf{y}-\mathbf{H}^T\mathbf{a}\right\|^2\right]. \tag{18.36}$$

This minimization amounts to regularizing the standard regression estimation with an absolute value penalty (also called L^1-norm penalty), instead of the quadratic penalty for the ridge regression. Interestingly, the solution to this minimization problem leads to a sparse estimator: the absolute value penalty strongly disfavors small values of $|a_i|$ and prefers to set these values to zero. Only sufficiently relevant coefficients a_i are retained – LASSO automatically selects the salient factors (this is the "SO" part in LASSO), which is very useful for interpreting the regression results intuitively.

Note that the true vector **a** is not sparse, as the probability to find a coefficient a_i to be exactly zero is itself zero for the prior Laplace distribution, which does not contain a singular $\delta(a)$ peak. The sparsity of the LASSO estimator $\mathbf{a}_{\text{LASSO}}$ is controlled by the parameter b. When $b\sigma_{\mathrm{n}}^2 \to 0$, the penalty disappears and all the coefficients of the vector **a** are non-zero (barring exceptional cases). When $b\sigma_{\mathrm{n}}^2 \to \infty$, on the other hand, all coefficients are zero. In fact, the number of non-zero coefficients is a monotonic decreasing function of $b\sigma_{\mathrm{n}}^2$. As for the parameter ζ for the ridge regression, it is hard to come up with a good prior value for b, which should be estimated again using validation or cross-validation methods (Figure 18.3). Finally we note that it is sometimes useful to combine the L^1 penalty of LASSO with the L^2 penalty of ridge, the resulting estimator is called an elastic net.

18.2.3 In-Sample and Out-of-Sample Error

Standard linear regression is built to minimize the sum of the squared-residuals on the dataset at hand. We call this error the *in-sample* error. In many cases, we are interested in

[4] LASSO stands for Least Absolute Shrinkage and Selection Operator.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

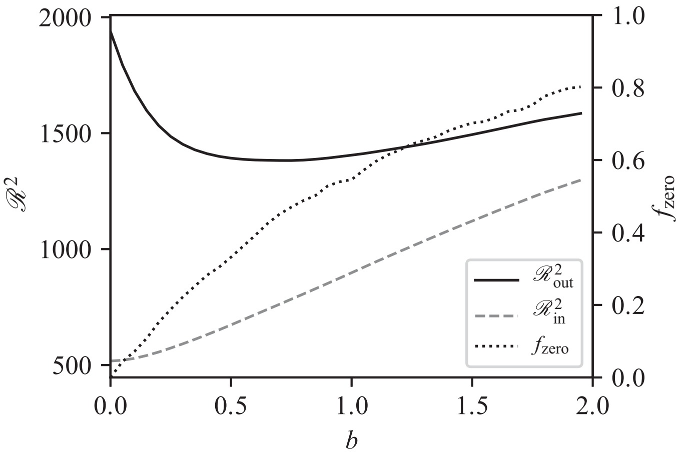

Figure 18.3 Illustration of the validation method in LASSO regularization. We built a linear model with 500 coefficients drawn from a Laplace distribution with $b = 1$ and Gaussian noise $\sigma_n^2 = 1000$. The model is estimated using $T = 1000$ unit Gaussian data and validated using a different set of the same size. The error on the training set is minimal with no regularization and gets worse as b increases (dashed line, left axis). The validation error (full line, left axis) is minimal for b about equal to 1. The dotted line (right axis) shows the fraction of the coefficient estimated to be exactly zero; this number grows from zero (no regularization) to almost 1 (strong regularization).

the predictive power of a linear model and the relevant error is the mean square error on a new independent but statistically equivalent dataset: the *out-of-sample* error. If the number of fitted variables (degree of freedom) is small with respect to the number of samples, the in-sample error is a good estimator of the out-of-sample error and the standard linear regression is also the optimal linear predictive model. The situation changes radically when the number of fitted variables becomes comparable to the number of samples, as we discuss below.

To summarize, linear regression, regularized or not, addresses two types of task:

- **In-sample estimator**: we observe some $\mathbf{H}_1$ and $\mathbf{y}_1$, and estimate $\mathbf{a}$.
- **Out-of-sample prediction**: we observe some other $\mathbf{H}_2$, non-overlapping with $\mathbf{H}_1$, and use them to predict $\mathbf{y}_2$ with the in-sample estimate of $\mathbf{a}$.

The result of the in-sample estimation is given by Eq. (18.31), which we write as

$$\mathbf{a}_{\text{reg}} = \mathbf{E}^{-1}\mathbf{b}; \qquad \mathbf{E} := \frac{1}{T}\mathbf{H}_1\mathbf{H}_1^T, \quad \mathbf{b} := \frac{1}{T}\mathbf{H}_1\mathbf{y}_1. \tag{18.37}$$

This is the best in-sample estimator. However, this is not necessarily the case for the out-of-sample prediction.

Note that both the standard regression and the ridge regression estimator (Eq. (18.34)) are of the form $\hat{\mathbf{a}} = \Xi^{-1}\mathbf{b}$ with $\Xi = \mathbf{E}$ and $\Xi = \mathbf{E} + \zeta\mathbf{1}$, respectively. We will compute

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

in the following the in-sample and out-of-sample estimation error for any estimator of that nature.

Recalling that $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma_{\text{n}}^2\mathbf{1}$ and after some calculations, one finds that the in-sample ($\mathcal{R}_{\text{in}}^2$) error is given by

$$\begin{aligned}\mathcal{R}_{\text{in}}^2(\hat{\mathbf{a}}) = \frac{1}{T}\Big\{&\sigma_{\text{n}}^2\left[T - 2\,\text{Tr}(\Xi^{-1}\mathbf{E}) + \text{Tr}(\Xi^{-1}\mathbf{E}\Xi^{-1}\mathbf{E})\right] \\ &+ \mathbf{a}^T\left(\mathbf{E} - 2\mathbf{E}\Xi^{-1}\mathbf{E} + \mathbf{E}\Xi^{-1}\mathbf{E}\Xi^{-1}\mathbf{E}\right)\mathbf{a}\Big\}.\end{aligned} \tag{18.38}$$

In the special case $\Xi = \mathbf{E}$, we have

$$\mathcal{R}_{\text{in}}^2(\mathbf{a}_{\text{reg}}) = \frac{\sigma_{\text{n}}^2}{T}(T-N) = (1-q)\sigma_{\text{n}}^2,$$

which is smaller than the true error, which is simply equal to σ_{n}^2. In fact, the error goes to zero as $q \to 1$, i.e. when the number of parameters becomes equal to the number of observations. This error reduction is called “overfitting”, or in-sample bias: if the task is to find the best model that explains past data, one can do better than the true error. Note that the above result is quite special, in the sense that it actually does not depend on either $\mathbf{E}$ or $\mathbf{a}$.

Next we calculate the expected out-of-sample ($\mathcal{R}_{\text{out}}^2$) error. We draw another matrix $\mathbf{H}_2$ of size $N \times T_2$ and consider another independent noise vector $\boldsymbol{\varepsilon}_2$ of variance σ_{n}^2 and size T_2 (where T_2 does not need to be equal to T, it can even be equal to 1). We calculate

$$\begin{aligned}\mathcal{R}_{\text{out}}^2(\hat{\mathbf{a}}) &= \frac{1}{T_2}\mathbb{E}_{\mathbf{H}_2,\boldsymbol{\varepsilon}_2}\left[\left\|\mathbf{H}_2^T\mathbf{a} + \boldsymbol{\varepsilon}_2 - \mathbf{H}_2^T\hat{\mathbf{a}}\right\|^2\right] \\ &= \frac{1}{T_2}\mathbb{E}_{\mathbf{H}_2,\boldsymbol{\varepsilon}_2}\left[\left\|\mathbf{H}_2^T\mathbf{a} + \boldsymbol{\varepsilon}_2 - \mathbf{H}_2^T\Xi^{-1}\mathbf{E}_1\mathbf{a} - \mathbf{H}_2^T\Xi^{-1}\frac{1}{T}\mathbf{H}_1\boldsymbol{\varepsilon}_1\right\|^2\right],\end{aligned} \tag{18.39}$$

where we denote $\mathbf{E}_1 := T^{-1}\mathbf{H}_1\mathbf{H}_1^T$. We now assume that $T_2^{-1}\mathbb{E}[\mathbf{H}_2\mathbf{H}_2^T] = \mathbf{C}$ with a general covariance $\mathbf{C}$.

In the standard regression case, $\Xi = \mathbf{E}$ and $\hat{\mathbf{a}} = \mathbf{a}_{\text{reg}}$ and we have

$$\mathcal{R}_{\text{out}}^2(\mathbf{a}_{\text{reg}}) = \frac{1}{T_2}\mathbb{E}_{\mathbf{H}_2,\boldsymbol{\varepsilon}_2}\left[\left\|\boldsymbol{\varepsilon}_2 - \mathbf{H}_2^T\Xi^{-1}\frac{1}{T}\mathbf{H}_1\boldsymbol{\varepsilon}_1\right\|^2\right] = \sigma_{\text{n}}^2 + \frac{\sigma_{\text{n}}^2}{T}\text{Tr}(\mathbf{E}^{-1}\mathbf{C}). \tag{18.40}$$

Now since $\mathbf{E}$ is a sample covariance matrix with true covariance $\mathbf{C}$, we have

$$\text{Tr}(\mathbf{E}^{-1}\mathbf{C}) = \text{Tr}\left(\mathbf{C}^{-\frac{1}{2}}\mathbf{W}_q^{-1}\mathbf{C}^{-\frac{1}{2}}\mathbf{C}\right) = \text{Tr}(\mathbf{W}_q^{-1}) \approx \frac{N}{1-q}, \tag{18.41}$$

where $\mathbf{W}_q$ denotes a standard Wishart matrix. Thus we find

$$\mathcal{R}_{\text{out}}^2(\mathbf{a}_{\text{reg}}) = \sigma_{\text{n}}^2 + \frac{q\sigma_{\text{n}}^2}{1-q} = \frac{\sigma_{\text{n}}^2}{1-q} = \frac{\mathcal{R}_{\text{in}}^2(\mathbf{a}_{\text{reg}})}{(1-q)^2}. \tag{18.42}$$

As an illustration see Figure 18.3 where without regularization ($b = 0$) we have indeed $\mathcal{R}^2_{\text{out}}/\mathcal{R}^2_{\text{in}} \approx (1-q)^{-2} = 4$. Thus, we see that we can make precise statements about the following intuitive inequalities:

$$\text{in-sample error} \le \text{true error} \le \text{out-of-sample error}. \tag{18.43}$$

Note that the out-of-sample error tends to ∞ as $N \to T$.

Now, let us compute the expected out-of-sample error ($\mathcal{R}^2_{\text{out}}$) for the ridge predictor $\mathbf{a}_{\text{ridge}}$, parameterized by ζ. The result reads

$$\mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{ridge}}) = \sigma_{\text{n}}^2 + \frac{\sigma_{\text{n}}^2}{T}\operatorname{Tr}(\mathbf{C}\Xi^{-1}\mathbf{E}\Xi^{-1}) + \zeta^2\operatorname{Tr}(\mathbf{C}\Xi^{-1}\mathbf{a}\mathbf{a}^T\Xi^{-1}), \tag{18.44}$$

with $\Xi = \mathbf{E} + \zeta\mathbf{1}$. Expanding to linear order for small ζ then leads to

$$\begin{aligned}\mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{ridge}}) &= \mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{reg}}) - \frac{2\sigma_{\text{n}}^2}{T}\operatorname{Tr}(\mathbf{C}\mathbf{E}^{-2})\zeta + O(\zeta^2)\\ &= \mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{reg}}) - \frac{2\sigma_{\text{n}}^2 q}{(1-q)^3}\tau(\mathbf{C}^{-1})\zeta + O(\zeta^2),\end{aligned} \tag{18.45}$$

where $\tau(.) = \operatorname{Tr}(.)/N$ and we have used the fact that $\tau(\mathbf{W}_q^{-2}) = (1-q)^{-3}$. The important point here is that the coefficient in front of ζ is *negative*, i.e. to first order, the ridge estimator has a lower out-of-sample error than the naive regression estimator:

$$\text{ridge estimation error} < \text{naive estimation error}. \tag{18.46}$$

However, the ridge estimator introduces a systematic bias since $\|\mathbf{a}_{\text{ridge}}\|^2 < \|\mathbf{a}_{\text{reg}}\|^2$ when $\zeta > 0$. This gives the third term in Eq. (18.44), which becomes large for larger ζ. So one indeed expects that there should exist an optimal value of ζ (which depends on the specific problem at hand) which minimizes the out-of-sample error. We now show how this optimal out-of-sample error can be elegantly computed in the large N limit.

The Large N Limit

In the large N limit we can recover the fact that the ridge estimator of Section 18.2.1 minimizes the out-of-sample risk without the need of a Gaussian prior on $\mathbf{a}$. We will also find an interesting relation between the Wishart Stieltjes transform and the out-of-sample risk of the ridge estimator.

In the following we will assume that the elements of the out-of-sample data $\mathbf{H}_2$ are IID with unit variance, i.e. that $\mathbf{C} = \mathbf{1}$. Then when $\Xi = \mathbf{E}_1 + \zeta\mathbf{1}$, Eq. (18.44) becomes, in the large N limit,

$$\mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{ridge}}) = \sigma_{\text{n}}^2\left(1 - q\,\mathfrak{g}_{\mathbf{W}_q}(-\zeta)\right) + \zeta\left(q\sigma_{\text{n}}^2 - \zeta|\mathbf{a}|^2\right)\mathfrak{g}'_{\mathbf{W}_q}(-\zeta), \tag{18.47}$$

where we have used that $\mathbf{E}_1$ is a Wishart matrix free from $\mathbf{a}\mathbf{a}^T$, and that

$$\tau((z\mathbf{1} - \mathbf{E}_1)^{-1}) = \mathfrak{g}_{\mathbf{W}_q}(z); \qquad \tau((z\mathbf{1} - \mathbf{E}_1)^{-2}) = -\mathfrak{g}'_{\mathbf{W}_q}(z). \tag{18.48}$$

For $\zeta = 0$, we have $\mathfrak{g}_{\mathbf{W}_q}(0) = -1/(1-q)$ and we thus recover Eq. (18.42). In the large N limit, the out-of-sample error for an estimator with $\Xi = \mathbf{E}_1 + \zeta\mathbf{1}$ depends on the

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

vector $\mathbf{a}$ only through its norm $|\mathbf{a}|^2$, regardless of the distribution of its components. The optimal value of ζ must then also only depend on $|\mathbf{a}|^2$.

Now, we know that when $\mathbf{a}$ is drawn from a Gaussian distribution, the value $\zeta_{\text{opt}} = \sigma_{\text{n}}^2/(T\sigma_{\text{s}}^2)$ is optimal. In the large N limit, $|\mathbf{a}|^2$ is self-averaging and equal to $N\sigma_{\text{s}}^2$. So $\zeta_{\text{opt}} = q\sigma_{\text{n}}^2/|\mathbf{a}|^2$. We can check directly that this value is optimal by computing the derivative of Eq. (18.47) with respect to ζ evaluated at ζ_{opt}. Indeed we have

$$\sigma_{\text{n}}^2 q \mathfrak{g}'_{\mathbf{W}_q}(-\zeta_{\text{opt}}) - \zeta_{\text{opt}}|\mathbf{a}|^2 \mathfrak{g}'_{\mathbf{W}_q}(-\zeta_{\text{opt}}) = 0. \tag{18.49}$$

For the optimal value of ζ we also have

$$\mathcal{R}^2_{\text{out}}(\mathbf{a}_{\text{ridge}}) = \sigma_{\text{n}}^2\left(1 - q\mathfrak{g}_{\mathbf{W}_q}(-\zeta_{\text{opt}})\right), \tag{18.50}$$

where $\mathfrak{g}_{\mathbf{W}_q}(z)$ is given by Eq. (4.40). Since $-\mathfrak{g}_{\mathbf{W}_q}(-z)$ is positive and monotonically decreasing for $z > 0$, we recover that the optimal ridge out-of-sample error is smaller than that of the standard regression.

18.3 Bayesian Estimation of the True Covariance Matrix

We now apply the Bayesian estimation method to covariance matrices. From empirical data, we measure the sample covariance matrix $\mathbf{E}$, and want to infer the most reliable information about the "true" underlying covariance matrix $\mathbf{C}$. Hence we write Bayes' equation for conditional probabilities for matrices:

$$P(\mathbf{C}|\mathbf{E}) \propto P(\mathbf{E}|\mathbf{C})P_0(\mathbf{C}). \tag{18.51}$$

We now recall Eq. (4.16) established in Chapter 4 for Gaussian observations:

$$P(\mathbf{E}|\mathbf{C}) \propto (\det \mathbf{C})^{-T/2} \exp\left[-\frac{T}{2}\operatorname{Tr}\left(\mathbf{C}^{-1}\mathbf{E}\right)\right]. \tag{18.52}$$

As explained in Section 18.1.3, in the absence of any meaningful prior information, it is interesting to pick a conjugate prior, which here is of the form

$$P_0(\mathbf{C}) \propto (\det \mathbf{C})^a \exp\left[-b\operatorname{Tr}\left(\mathbf{C}^{-1}\mathbf{X}\right)\right] \tag{18.53}$$

for some matrix $\mathbf{X}$, which turns out to be proportional to the prior mean of $\mathbf{C}$. Indeed, this prior is in fact the probability density of the elements of an inverse-Wishart matrix. Consider an inverse-Wishart matrix $\mathbf{C}$ of size N, T^* degree of freedom and centered at a (positive definite) matrix $\mathbf{X}$. If $T^* > N + 1$, $\mathbf{C}$ has the density (see Eq. (15.35))

$$P(\mathbf{C}) \propto (\det \mathbf{C})^{-(T^*+N+1)/2} \exp\left[-\frac{T^* - N - 1}{2}\operatorname{Tr}\left(\mathbf{C}^{-1}\mathbf{X}\right)\right]. \tag{18.54}$$

Note that here $T^* > N$ is some parameter that is unrelated to the length of the time series T. The chosen normalization is such that $\mathbb{E}_0[\mathbf{C}] = \mathbf{X}$. As $T^* \to \infty$, we have $\mathbf{C} \to \mathbf{X}$.

With this prior we thus obtain

$$P(\mathbf{C}|\mathbf{E}) \propto (\det \mathbf{C})^{-(T+T^*+N+1)/2} \exp\left[-\frac{T}{2}\operatorname{Tr}\left(\mathbf{C}^{-1}\mathbf{E}^*\right)\right], \tag{18.55}$$

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press

where we define

$$\mathbf{E}^* := \mathbf{E} + \frac{T^* - N - 1}{T}\mathbf{X}. \tag{18.56}$$

We now notice that (18.55) is, by construction, also a probability density for an inverse-Wishart with $\bar{T} = T + T^*$, with

$$\mathbb{E}[\mathbf{C}|\mathbf{E}] = \frac{T\mathbf{E}^*}{T + T^* - N - 1} = r\mathbf{E} + (1-r)\mathbf{X}, \tag{18.57}$$

with

$$r = \frac{T}{T + T^* - N - 1}. \tag{18.58}$$

Hence we recover a linear shrinkage, similar to Eq. (18.10) in the case of a scalar variable with a Gaussian prior. We will recover this shrinkage formula in the context of rotationally invariant estimators in the next chapter, see Eq. (19.49).

We end with the following remarks:

- The linear shrinkage works even for the finite N case, i.e. without the large N hypothesis.
- In general, if one has no idea of what $\mathbf{X}$ should be, one can use the identity matrix, i.e.

  $$\mathbb{E}[\mathbf{C}|\mathbf{E}] = r\mathbf{E} + (1-r)\mathbf{1}. \tag{18.59}$$

  Another simple choice is a covariance matrix $\mathbf{X}$ corresponding to a one-factor model (see Section 20.4.2):

  $$\mathbf{X}_{ij} = \sigma_s^2\left[\delta_{ij} + \rho(1-\delta_{ij})\right], \tag{18.60}$$

  where ρ is the average pairwise correlation (which can also be learned using validation).
- Note that T^* (or equivalently r) is generally unknown. It may be inferred from the data or learned using validation.
- As we will see in Chapter 20, the linear shrinkage works quite well in financial applications, showing that inverse-Wishart is not a bad prior for the true covariance matrix in that case (see Fig. 15.1).

Bibliographical Notes

- Some general references on Bayesian methods and statistical inference:
  - E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003,
  - G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 2013.

https://doi.org/10.1017/9781108768900.022 Published online by Cambridge University Press