Chapter 7

Statistical estimation

7.1 Frequency approach

One approach to statistical estimation is a frequentist approach. In this approach we assign a probability to events based on their frequency of occurrence. For example, we would say that the probability of an even coin landing on its head is 50%. This means that if we repeat the experiment many times, roughly half of the times the coin will land head's up.

A different approach to probability, which we discuss later is probability based on information. For example, we say that the chance of rain tonight is 20%. Obviously, we cannot repeat the "experiment" many times. Yet, we seem to "understand" the meaning of this expression. We discuss this approach in the next section. Here we concentrate on the frequentist approach that assign probability only to measurable events.

7.1.1 Bias and Variance

A classical tool for inverse problem is the bias variance decomposition. First, let us define the bias and the variance

Bias
We say that an estimator, $\widehat{m}$ of an unknown parameter m is unbiased if

$$\mathrm{E}\widehat{m} = m$$

for any possible value of m. In other words, on the average the estimator $\widehat{m}$ provides the correct value. If $\widehat{m}$ is not unbiased (it is biased) we want to know how far of we expect to be on the average. The bias of $\widehat{m}$ is dened as

$$\mathrm{Bias} = \mathrm{E}\,\widehat{m} - m$$

Variance

We define the variance of an estimator, $\widehat{m}$ as

$$\text{Var} = \text{E}(\text{E}\widehat{m} - \widehat{m})^2$$

Obviously, unbiased estimators are desired however, in most inverse problems such estimators perform poorly. This is because such estimators may have a huge variance. To demonstrate, consider the following example

$$\begin{pmatrix} 1 & 0 \\ 0 & \zeta \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$

With ϵ Gaussian and iid with variance σ^2. The least squares solution is

$$\widehat{m}_1 = d_1 \quad \widehat{m}_2 = \zeta^{-1} d_2$$

Obviously, this estimator is unbiased since

$$\text{E}\widehat{m}_1 = \text{E}d_1 = m_1$$

and

$$\text{E}\widehat{m}_2 = \text{E}\zeta^{-1} d_2 = m_2$$

Now, assume that ζ is small. In this case the variance of $\widehat{m}_1$ is σ^2 however, the variance of $\widehat{m}_2$ is $\sigma^2 \zeta^{-1}$. This can be a huge variance and thus lead to an unstable result. Consider on the other hand the estimator

$$\widehat{m}_1 = d_1 \quad \widehat{m}_2 = 0$$

Obviously, this estimator is biased, nevertheless, the variance of this estimator is very small. Finally, consider the estimator

$$\widehat{m}_1 = d_1 \quad \widehat{m}_2 = (\zeta + \alpha)^{-2} \zeta d_2$$

It is easy to verify that for small m_2 the bias is very small while for large m_2 the bias is larger. However, this estimator has a small variance which makes it useful in practical setting.

7.1.2 The Bias-Variance decomposition

An important question is what is the error in recovering the model by an estimator $\widehat{m}$, that is we seek to estimate the Mean Square Error (MSE) defined as

$$\text{MSE} = \text{E} \left\| \widehat{m} - m \right\|^2$$

A nice trick can be used.

$$\begin{aligned} \text{MSE} = \text{E} \left\| \widehat{m} - m \right\|^2 &= \text{E} \| \widehat{m} - \text{E}\widehat{m} + \text{E}\widehat{m} - m \|^2 \\ &= \text{E} \| \widehat{m} - \text{E}\widehat{m} \|^2 + \text{E} \| \text{E}\widehat{m} - m \|^2 + 2\text{E} \left((\text{E}\widehat{m} - m)^\top (\widehat{m} - \text{E}\widehat{m}) \right) \end{aligned}$$

Now note that $\mathrm{E}\left((\mathrm{E}\widehat{m}-m)^\top(\widehat{m}-\mathrm{E}\widehat{m})\right)=0$ and therefore

$$\mathrm{MSE}=\mathrm{E}\,\|\widehat{m}-m\|^2=\mathrm{E}\|\widehat{m}-\mathrm{E}\widehat{m}\|^2+\mathrm{E}\|\mathrm{E}\widehat{m}-m\|^2=\mathrm{Var}+\|\mathrm{Bias}\|^2$$

This decomposition is a classical one known as the *Bias-Variance* decomposition. It show that controlling the total error requires the solution of both the bias and the variance. For many problems there is a tradeoff. One could reduce the bias and increase the variance and vise-versa.

Consider Tikhonov regularization with a *fixed* regularization parameter α. The estimator $\widehat{m}$ can be written as

$$\widehat{m}=(A^\top A+\alpha L^\top L)^{-1}A^\top d.$$

Define $C=(A^\top A+\alpha L^\top L)$ we compute the MSE and obtain

$$\mathrm{MSE}=\mathrm{E}\,\|\widehat{m}-m\|^2=\alpha^2\|C^{-1}L^\top Lm\|^2+\sigma^2\,\mathrm{trace}(AC^{-2}A^\top)$$

For non-fixed regularization parameter (e.g. when GCV or discrepancy principle are used) there is no such "clean" decomposition because the regularization parameter α depends on the true model. Nevertheless, one can hope that for a wide range of models, similar regularization parameters "work". In this case one can get an idea of the uncertainty of the estimator using the above formula. If a more accurate estimate of the MSE is needed then one needs to differ to Monte-Carlo simulations. This is done as follows

- $e=0, j=1$
- while 1
  - Generate data $d_j=Am+\epsilon_j$
  - solve for $\widehat{m}_j$
  - $\mathrm{MSE}_j=\frac{1}{j}\sum\|\widehat{m}-m\|^2$
  - termination criteria

The above algorithm can be used to estimate the MSE for a known model. As many Monte-Carlo methods it can be slowly to converge however, unfortunately, there are no other alternatives when the estimator $\widehat{m}$ is nonlinear with respect to the data

7.1.3 Uncertainty

The above analysis separate the error of an estimator $\widehat{m}$ into bias and variance. The variance depends on the noise alone while the bias depends on the true model. If the only information given is the data then it may be possible to assess the

variance but it is impossible to assess the bias based on the data alone. This is why uncertainty estimation is very difficult for ill-posed problems while it is relatively straight forward to do so for well posed problems.

There are two main approaches to estimate the bias. The first approach is to look at some average case. Recall that the bias is defined as

$$\text{Bias} = \text{E}\widehat{m} - m.$$

If we have a distribution of all possible models then we can (only in principle) compute the average bias

$$\overline{\text{Bias}} = \int_{\mathcal{M}} (\text{E}\widehat{m} - m)\ dm$$

where $\mathcal{M}$ is the space of possible models. Such space can be

- A convex set, for example $\|m\|_W \leq 1$
- A distribution, (e.g. Gaussian)
- The convex hull of many examples $\mathcal{M} = \text{span}(m_1, \ldots, m_q)$.

Estimating the bias can be done *prior* to obtaining any data! One way to do it is to compute the integral using Monte-Carlo methods. For some simple (yet important) cases one can avoid Monte-Carlo methods. Consider the case of Tikhonov regularization with a fixed regularization parameter. We saw that the bias can be written as

$$\|\text{Bias}\|^2 = \alpha^2 \|C^{-1}L^\top Lm\|^2$$

Assume now that m has a PDF with mean m_0 and covariance matrix C_m. Then, it is easy to verify that

$$\|\overline{\text{Bias}}\|^2 = \alpha^2 \|C^{-1}L^\top Lm_0\|^2 \ + \ \alpha^2 \text{trace}\,(LL^\top C^{-1} C_m C^{-1} L^\top L)$$

The interesting point is that even if m has a very complicated distribution then we require only the covariance. Thus, we may be able to estimate the covariance of m which is much easier than the whole distribution.

The estimation of the variance is straight forward. In this case we require to estimate

$$\text{Var} = \sigma^2\, \text{trace}(AC^{-2}A^\top)$$

For large scale problems, estimating both bias and variance can be rather difficult. Stochastic trace estimators are useful tools for these problems. A common approximation for the trace is

$$\text{trace}(H) \approx \sum_i v_i^\top H v_i$$

It turns out that even with a choice of a single vector the approximation can be reasonable.

It is possible to use the same ideas as above in order to estimate the model in an area. For example, consider any window w and the product $w^\top m$. The difference

$$\mathsf{MSE}_w = \|w^\top(\widehat{m} - m)\|^2$$

can be thought of as a semi-norm. In this case, the weighted bias and variance are

$$\|\overline{\text{Bias}}\|^2 = \alpha^2\|w^\top C^{-1}L^\top L m_0\|^2 \ + \ \alpha^2 \text{trace}\,(ww^\top LL^\top C^{-1}C_m C^{-1}L^\top L)$$
$$\text{Var} = \sigma^2\,\text{trace}(ww^\top AC^{-2}A^\top)$$

7.2 Bayesian approach

In the frequentist approach we treat the problem of recovering the model as a deterministic quantity. The way to obtain a stable estimate was to use regularization and the way to pick a regularization operator was to reduce the bias associated with the models we expect to have. A different approach all-together is to use Bayesian techniques for the estimation of the model. The advantage of Bayesian methods is that they use rigorous probabilistic tools to estimate the model and its uncertainty. If one "buys into" this framework then it is easy to come with answers (at least in principle) for most of the estimation problems that are associated with the inverse problem. Here we try to review the Bayesian framework and discuss its advantages and its faults.

7.2.1 Bays formula and its implication

In the Bayesian framework we generate a probabilistic model of m. This is sometimes presented in a somewhat confusing manor and (in my opinion) without real reason.

The idea is that we treat m as a random variable. This does not mean that the "true" model is random. It does mean that the information about the model is modeled as a probability and before we know what the true model is, we assign a probability density function to its distribution. Let $\pi(m)$ be the probability density function that is associated with the mode prior to conducting any experiment and collecting any data. Since this probability is not related to the data it is referred to as, the prior.

Now assume that we are given a model m. The question is, what is the probability of having a data vector d. This is a conditional probability, which we mark as $p(d|m)$. In the case of Gaussian, iid noise we can write the prior as

$$p(d|m) \sim \exp\left(-\frac{\|d - Am\|^2}{\sigma^2}\right)$$

The probability $p(d|m)$ is referred to as, the likelihood.

Figure 7.1. *The prior (left) the likelihood (middle) and the posterior (right)*

Next, we collect data d and ask, what is the probability of the model m given the data vector d. Bayes formula is

$$p(m|d) = \frac{\pi(m)\, p(d|m)}{p(d)} \tag{7.1}$$

The probability $p(m|d)$ is referred to as the posterior. This simple formula entails in it all the information about the model. In fact, one may claim that this is the "answer" to our inference problem of recovering the model given the data.

Let us demonstrate the idea using a very simple example. Assume that $m = [m_1, m_2]$ and that we have a single datum of the form

$$d = m_1 + m_2 + \epsilon$$

If ϵ is normal iid with standard deviation σ^2 then d is Gaussian with mean $m_1 + m_2$ and standard deviation σ^2, that is

$$p(d|m) \sim \exp\left(-\frac{(d - m_1 + m_2)^2}{\sigma^2}\right).$$

Now assume that the prior is also Gaussian and that

$$\pi(m) \sim \exp\left(-\frac{m_1^2 + m_2^2}{\sigma_m^2}\right).$$

Finally, the posterior is

$$p(m|d) \sim \exp\left(-\frac{(d - m_1 + m_2)^2}{\sigma^2} - \frac{m_1^2 + m_2^2}{\sigma_m^2}\right).$$

We plot the probability of the prior, the likelihood and the posterior in Figure 7.1.

The question, what is the model that yields a particular set of data is now meaningless. There is only a distribution of models, some are more likely than others to recover the data. For example, we can ask, which model maximizes the posterior. This model is the maximum a posteriori (MAP) model. Since $p(m|d)$ is a distribution, the chances of having this particular model are zero and therefore, the model cannot be treated as "the answer to the estimation problem".

Now consider a slightly more general case where

$$\begin{aligned} \pi(m) &\sim \exp(-m^\top C_m^{-1} m) \\ p(d|m) &\sim \exp(-(d - Am)^\top C_d^{-1}(d - Am)) \end{aligned}$$

with C_m and C_d are covariance matrices for the prior and the likelihood. Then, the posterior is

$$p(m|d) = \exp(-m^\top C_m^{-1} m - (d - Am)^\top C_d^{-1}(d - Am))$$

The MAP model maximizes the posterior or minimizes the exponent, that is

$$m_{\text{MAP}} = \operatorname{argmin} m^\top C_m^{-1} m + (d - Am)^\top C_d^{-1}(d - Am))$$

Assume now that $C_m^{-1} = L^\top L$ and that $C_d^{-1} = \alpha^{-1} I$ then

$$m_{\text{MAP}} = \operatorname{argmin} \alpha \|Lm\|^2 + \|d - Am\|^2$$

which is equivalent to the Tikhonov estimate.

The equivalence between a particular MAP estimate and the Tikhonov estimate can be misleading. One may believe that the Bayesian estimators are equivalent to their deterministic counterparts. This is one of the most common mistakes. In fact, the frequentist approach is very different compared with the Bayesian approach for most other aspect of the model. Maybe, the most important aspect of the Bayesian approach is the fact that the bias does not exist. Since there is no "true" model there is no meaning to discuss $\mathsf{E}m - m$. In fact, for any symmetric posterior we have that $\mathsf{E}m$ is the MAP estimate. Thus, the MAP estimate can be interpreted as an average of all the models that yield the data (in a distribution form).

We can however ask, what is the difference between the MAP estimate m_{MAP} and the particular realization m^r given a realization of the noise ϵ^r. The usual formula shows that

$$m_{\text{MAP}} - m^r = -\alpha^2 (A^\top A + \alpha L^\top L)^{-1} L^\top L m^r + (A^\top A + \alpha L^\top L)^{-1} A^\top \epsilon^r$$

Taking the expected value over the noise we have

$$\mathsf{E}_\epsilon \left(m_{\text{MAP}} - m^r \right) = -\alpha^2 (A^\top A + \alpha L^\top L)^{-1} L^\top L m^r$$

Thus the bias in the frequentist approach is interpreted as the difference between a particular realization to the mean. If we average over all models we simply get that

$$\mathsf{E}_m \mathsf{E}_\epsilon \left(m_{\text{MAP}} - m \right) = 0.$$

(assuming a symmetric probability density function).

Given the posterior, one can now discuss any statistical quantity of the model including confidence intervals. An α confidence interval, I is an interval for which

$$P(m \in I) = 1 - \alpha.$$

It is easy to obtain a formula for this interval and assuming that the bias is correct, one can actually expect to obtain a correct estimate for the uncertainty involved with the MAP estimate.

Another interesting feature of classical Bayesian estimation is that there is no need to compute a regularization parameter. Given the covariance matrices of the prior and the likelihood determine the relative weight between the "regularization" term and the data fitting term. Although this may look like an advantage it is a serious disadvantage as it is well known that it is better to choose the regularization parameter taking the actual noise into consideration. Although there are Bayesian approaches that allow for the use of GCV or any other criteria, such an approach is less natural within a Bayesian framework.