Estimation

1 Introduction to Estimation

One of the powers of probability is that you can estimate the behavior of phenomena. How can we do that? That's something we will look at in this chapter.

1.1 Definitions

It often occurs that there is some phenomenon of which we want to know the behavior. Such a phenomenon can be modeled as a random vector $\underline{\mathbf{y}}$, with a certain size m. However, we usually don't know the distribution of such a random variable. The PDF just isn't known to us. But, given certain parameters, we can find it. In this case we can put the n unknown parameters into a vector $\mathbf{x}$. Then, once $\mathbf{x}$ is known, we can find the PDF of $\underline{\mathbf{y}}$. This PDF is written as $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$. (An example could be where we know $\underline{\mathbf{y}}$ is normally distributed, but we don't know the mean $\overline{\mathbf{y}}$ and the standard deviation σ_y.)

Now how can we find $\mathbf{x}$? The truth is, we can't. However, by observing the phenomenon described by $\underline{\mathbf{y}}$, we can guess it. Let's say our guess $\hat{\mathbf{x}}$ (called the **estimate** of x) is given by some function $\hat{\mathbf{x}} = G(\mathbf{y})$. Our task is to set this function $G(\mathbf{y})$. Once we have done so, we can also define a new random variable, called the **estimator**, as $\underline{\hat{\mathbf{x}}} = G(\underline{\mathbf{y}})$.

1.2 Finding a good estimator

So we now know we have to choose an estimator $\underline{\hat{\mathbf{x}}} = G(\underline{\mathbf{y}})$. How would we know what would be a good one? There are three criteria for that. First there is the **estimation error** $\underline{\hat{\epsilon}} = \underline{\hat{\mathbf{x}}} - \mathbf{x}$, which is also a random variable. The estimator $\hat{x}$ is said to be an **unbiased estimator** of x if, and only if $E(\underline{\hat{\epsilon}}) = E(\underline{\hat{\mathbf{x}}} - \mathbf{x}) = \mathbf{0}$, or, equivalently, $E(\underline{\hat{\mathbf{x}}}) = \mathbf{x}$. If the estimator $\underline{\hat{\mathbf{x}}}$ is not unbiased, then we say that its **bias** is $E(\underline{\hat{\epsilon}})$.

Another measure of quality is the **mean squared error** (MSE), defined as $E\left(|\underline{\hat{\mathbf{x}}} - \mathbf{x}|^2\right)$, or, equivalently, $E\left(\underline{\hat{\epsilon}}^2\right)$. Of course the mean squared error should be as small as possible.

Finally, let's look at the third measure of quality. It is defined as $P(|\underline{\hat{\mathbf{x}}} - \mathbf{x}|^2 \leq r^2)$, for some radius r. In words, this is the probability that the vector $\underline{\hat{\epsilon}}$ is in the (hyper-)sphere with radius r. This chance should be as big as possible.

There are three common ways of finding an estimator. Which one to use depends on the data that you have and the accuracy that you want. We will take a look at them in the rest of this chapter.

2 Least-Squares Estimation

One of the most well-known methods of determining an estimation is the least-squares estimation method. Let's take a look at how it works.

2.1 The consistent case

Let's suppose we have a set of measurements $\mathbf{y}$, having size m, and a set of unknown parameters $\mathbf{x}$, having size n. To apply the **least-squares method**, we assume that

$$\mathbf{y} = A\mathbf{x}, \tag{2.1}$$

where the $m \times n$ matrix A is known. In words, this means that the measured parameters (in $\mathbf{y}$) are linear functions of the unknown variables (in $\mathbf{x}$). However, now the question arises whether, given a (measured) $\mathbf{y}$, there is an $\mathbf{x}$ which satisfies the above equation. If there is, then the system is **consistent**. Otherwise it is **inconsistent**. The inconsistent case will be treated in the next paragraph. Now we'll take a closer look at the consistent case.

So suppose the system $\mathbf{y} = A\mathbf{x}$ is consistent. In this case we know that there is at least one $\mathbf{x}$ satisfying $\mathbf{y} = A\mathbf{x}$. If there is exactly one solution, then this solution is our estimate $\hat{\mathbf{x}}$. It can be found using

$$\hat{\mathbf{x}} = A^{-1}\mathbf{y}. \tag{2.2}$$

The corresponding **least-squares solution** $\hat{\mathbf{y}}$ can be found using $\hat{\mathbf{y}} = A\hat{\mathbf{x}} = \mathbf{y}$.

It is, however, also possible that there are infinitely many solutions $\mathbf{x}$. In this case we can't be sure which $\mathbf{x}$ to choose. This often means we need more measurement data. This is the case if the columns of A aren't all linearly independent.

2.2 The inconsistent case

Now let's suppose the system $\mathbf{y} = A\mathbf{x}$ is inconsistent. In this case there is no solution $\mathbf{x}$. We now refer to the system as an **overdetermined system**, denoted as $\mathbf{y} \approx A\mathbf{x}$. Assuming that there are no linearly dependent columns in A, we define the **redundancy** of the system as $m - n$.

To make sure there are solutions, we add a measurement error vector $\mathbf{e}$, such that $\mathbf{y} = A\mathbf{x} + \mathbf{e}$. We now want to choose $\mathbf{e}$ such that $\mathbf{e}^2 = \mathbf{e}^T\mathbf{e}$ is minimal. This is the **least-squares principle**. The minimal $\mathbf{e}$ is denoted by $\hat{\mathbf{e}}$. With $\hat{\mathbf{e}}$ chosen correctly, there is a solution $\mathbf{x}$, called the estimate $\hat{\mathbf{x}}$. It can be found using

$$\hat{\mathbf{x}} = \left(A^T A\right)^{-1} A^T\mathbf{y}. \tag{2.3}$$

The corresponding least-squares solution can once more be found using $\hat{\mathbf{y}} = A\hat{\mathbf{x}}$. The difference $\hat{\mathbf{e}} = \mathbf{y} - A\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{y}}$ is the **least-squares residual vector**. The value of $\hat{\mathbf{e}}^2$ is a measure of the inconsistency of the system.

The vector $\mathbf{y}$ generally consists of measurement data. Sometimes we know that some measurement data is more accurate than other. That data should thus be taken into account more. For this, there is the **weighted least-squares method**. In this case we don't want to minimize $\mathbf{e}^2 = \mathbf{e}^T\mathbf{e}$. This time we should minimize $\mathbf{e}^T W\mathbf{e}$, where W is the **weight matrix**. The minimum value $\hat{\mathbf{e}}^T W\hat{\mathbf{e}}$ is now the measure of inconsistency of the system. The corresponding estimate can then be found using

$$\hat{\mathbf{x}} = \left(A^T W A\right)^{-1} A^T W\mathbf{y}. \tag{2.4}$$

Generally W is a positive diagonal matrix. This is, however, not always the case.

2.3 Orthogonal projectors

Let's take a closer look at what variables we got now. We have a measurement vector $\mathbf{y}$. In the inconsistent case $\mathbf{y}$ can't be written as $A\mathbf{x}$. So we search for a $\hat{\mathbf{y}}$ close to $\mathbf{y}$ that can be written as $A\hat{\mathbf{x}}$. The least-squares residual vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ is as small as possible.

It can now be shown that $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are orthogonal vectors. While $\hat{\mathbf{y}}$ lies in the column space of A, $\hat{\mathbf{e}}$ is orthogonal to the column space of A. (So $\hat{\mathbf{e}}$ lies in the column space of $A^\perp$.) Now let's define the two matrices P_A and $P_A^\perp$ as

$$P_A = A\left(A^T W A\right)^{-1} A^T W \qquad \text{and} \qquad P_A^\perp = I_m - P_A. \tag{2.5}$$

These two matrices are **orthogonal projectors**. What they do is, they project vectors on the column spaces of A and $A^\perp$. We therefore have

$$\hat{\mathbf{y}} = P_A\mathbf{y} \qquad \text{and} \qquad \hat{\mathbf{e}} = P_A^\perp\mathbf{y}. \tag{2.6}$$

2.4 Implementing random vectors

Previously we saw the system of equations $\mathbf{y} = A\mathbf{x}+\mathbf{e}$. Here, the vector $\mathbf{y}$ represented a series of measurements. However, we can take more measurements from a phenomenon. Every time, these measurements can take different values. So it would be wise to represent $\mathbf{y}$ by a random vector $\underline{\mathbf{y}}$. Equivalently, the vector $\mathbf{e}$ also has a different value every time. So it should also be a random vector $\underline{\mathbf{e}}$. We thus get

$$\underline{\mathbf{y}} = A\mathbf{x} + \underline{\mathbf{e}}. \tag{2.7}$$

We assume we have chosen our estimate $\mathbf{x}$ (which does stay constant during different measurements) such that $E(\underline{\mathbf{e}}) = \mathbf{0}$ or, equivalently, $E(\underline{\mathbf{y}}) = A\mathbf{x}$. If this is indeed the case, then we say that the above linear system is called a **linear model** of $E(\underline{\mathbf{y}})$.

From $\underline{\mathbf{y}}$, we can derive the random variables $\underline{\hat{\mathbf{x}}}$, $\underline{\hat{\mathbf{y}}}$ and $\underline{\hat{\mathbf{e}}}$. For that, we can use relations we actually already know. They are

$$\underline{\hat{\mathbf{x}}} = \left(A^TWA\right)^{-1}A^TW\underline{\mathbf{y}} = \mathbf{x} + \left(A^TWA\right)^{-1}A^TW\underline{\mathbf{e}}, \tag{2.8}$$

$$\underline{\hat{\mathbf{y}}} = P_A\underline{\mathbf{y}} = A\mathbf{x} + P_A\underline{\mathbf{e}} \qquad \text{and} \qquad \underline{\hat{\mathbf{e}}} = P_A^{\perp}\underline{\mathbf{y}} = \mathbf{0} + P_A^{\perp}\underline{\mathbf{e}}. \tag{2.9}$$

And that's all we need to know if we want to estimate what the outcome of the experiment will be next.

3 Best linear unbiased estimation

The second method of finding $\underline{\hat{\mathbf{x}}}$ we will look at is the BLUE method. To use it, you also need the variance matrix Q_{yy} of $\underline{\mathbf{y}}$. Because the BLUE method considers Q_{yy}, it can be a rather accurate estimation method. Let's find out how it works.

3.1 Definitions and conditions

Let's consider the linear system of equations

$$E(\underline{\mathbf{y}}) = A\mathbf{x}. \tag{3.1}$$

Just like in the previously discussed method, we want to find an estimate $\mathbf{x}$ (or more precise, an estimator $\underline{\mathbf{x}}$) that corresponds to certain conditions. Before we look at those conditions, we first make some definitions.

Let's define the vector $\mathbf{z}$ (having size k) as a linear combination of $\mathbf{x}$. So, $\mathbf{z} = F^T\mathbf{x} + \mathbf{f_0}$, for some known $n \times k$ matrix F and k-vector $\mathbf{f_0}$. Just like we want to find an estimator $\underline{\hat{\mathbf{x}}}$ for $\mathbf{x}$, we can also be looking for an estimator $\underline{\mathbf{z}}$ for $\mathbf{z}$. This gives us also the relation $\underline{\mathbf{z}} = F^T\underline{\mathbf{x}} + \mathbf{f_0}$. So to find $\underline{\mathbf{x}}$ we might as well try to find $\underline{\mathbf{z}}$. The estimator $\underline{\mathbf{z}}$ depends on $\underline{\mathbf{y}}$. So let's define $G(\underline{\mathbf{y}})$ such that $\underline{\mathbf{z}} = G(\underline{\mathbf{y}})$.

Enough definitions. Let's look at what conditions we want the estimator $\underline{\mathbf{z}}$ to have. First of all we want it to be **unbiased**. This means that $E(\underline{\mathbf{z}}) = \mathbf{z}$ or, equivalently, $E(G(\underline{\mathbf{y}})) = F^T\mathbf{x} + \mathbf{f_0}$. For reasons of simplicity, we also want it to be **linear**. This is the case if $G(\underline{\mathbf{y}})$ is a linear function, and can thus be written as $G(\underline{\mathbf{y}}) = L^T\underline{\mathbf{y}} + \mathbf{l_0}$ for some $m \times k$ matrix L and a k-vector $\mathbf{l_0}$. This linearity condition can be rewritten to two conditions, being

$$A^TL = F \qquad \text{and} \qquad \mathbf{l_0} = \mathbf{f_0}. \tag{3.2}$$

Any estimator $\underline{\mathbf{z}}$ of $\mathbf{z}$ that is both linear and unbiased is called a **linear unbiased estimator** (LUE).

3.2 Finding the best linear unbiased estimator

There is one slight problem. There are many LUEs. We want only the so-called **best linear unbiased estimator** (BLUE), denoted by $\underline{\hat{\mathbf{z}}}$. But what is the best LUE? We now define the best LUE (the BLUE) to be the LUE with the smallest mean squared error (MSE). So we have

$$E\left(|\underline{\hat{\mathbf{z}}}-\mathbf{z}|^2\right) \leq E\left(|\underline{\mathbf{z}}-\mathbf{z}|^2\right) \tag{3.3}$$

for every LUE $\underline{\mathbf{z}}$. Now the question arises how we can find this BLUE $\underline{\hat{\mathbf{z}}}$ and the corresponding best estimator $\underline{\hat{\mathbf{x}}}$ for $\mathbf{x}$. For that, we use the **Gauss-Markov** theorem, which states that

$$\underline{\hat{\mathbf{x}}} = \left(A^T Q_{yy}^{-1} A\right)^{-1} A^T Q_{yy}^{-1} \underline{\mathbf{y}} \qquad \text{and} \qquad \underline{\hat{\mathbf{z}}} = F^T \underline{\hat{\mathbf{x}}} + \mathbf{f_0}, \tag{3.4}$$

where Q_{yy} is the variance matrix of $\underline{\mathbf{y}}$ (so $Q_{yy} = D(\underline{\mathbf{y}})$). It is interesting to note that the final value of $\underline{\hat{\mathbf{x}}}$ does not depend on the matrix F or the vector $\mathbf{f_0}$ at all. It just depends on A and $\underline{\mathbf{y}}$.

Another interesting fact to note is the link with the weighted least-squares estimation method. If the weight matrix W in the WLSE method is equal to the inverse of the variance matrix Q_{yy} (so $W = Q_{yy}^{-1}$) in the BLUE method, you would find exactly the same estimator $\underline{\mathbf{x}}$.

4 Maximum Likelihood Estimation and Confidence Regions

The third method of finding $\underline{\hat{\mathbf{x}}}$ uses the PDF of $\underline{\mathbf{y}}$. It can therefore not always be applied. But its advantage is that it can be applied in the case where $\underline{\mathbf{y}}$ can't be written as $A\mathbf{x}$.

4.1 The condition

To make a **maximum likelihood estimation** (MLE) we need to know the PDF of $\underline{\mathbf{y}}$, given a certain unknown vector $\mathbf{x}$. We write this as $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$.

Now suppose we have some measurement $\mathbf{y}$. The idea behind the MLE is to choose the value of $\mathbf{x}$ for which it is most likely that $\mathbf{y}$ occurred. The **likelihood of y** then is $f_{\mathbf{y}}(\mathbf{y}|\mathbf{x})$. Note that this likelihood is a function of the unknown vector $\mathbf{x}$. This likelihood should be maximized. The corresponding $\mathbf{x}$, now denoted as $\hat{\mathbf{x}}$, is the maximum likelihood estimation.

4.2 Finding the maximum likelihood estimation

There is no general method of finding the MLE. For relatively easy PDFs of $\underline{\mathbf{y}}$, simple logic can often lead to the MLE. For more difficult PDFs, finding the MLE might even require complicated (numerical) computation.

If, however, $f_{\underline{\mathbf{y}}}(\mathbf{y}|\mathbf{x})$ is sufficiently smooth, then $\hat{\mathbf{x}}$ can be found using the conditions

$$\partial_x f_{\underline{\mathbf{y}}}(\mathbf{y}|\hat{\mathbf{x}}) = \mathbf{0} \qquad \text{and} \qquad \partial^2_{xx^T} f_{\underline{\mathbf{y}}}\left(\mathbf{y}|\hat{\mathbf{x}}\right) < 0. \tag{4.1}$$

If the PDF simply gives a scalar result, then the above states that the first derivative must be zero, indicating that there is a local minimum/maximum. The second derivative must be smaller than zero, indicating it is in fact a maximum.

If, however, the PDF returns a vector, then things are a bit more difficult. Then the first condition requires that the gradient has to be zero. The second condition states that the so-called **Hessian matrix** (the matrix of derivatives) needs to be negative definite. In other words, all its eigenvectors need to be negative.

Finally, when multiple values satisfy the above conditions, just insert their values $\hat{\mathbf{x}}$ into $f_{\underline{\mathbf{y}}}(\mathbf{y}|\hat{\mathbf{x}})$ and see which value gives the highest likelihood of $\mathbf{y}$.

4.3 Confidence Regions

Suppose we have finally acquired a good estimate $\hat{\mathbf{y}}$ of $\mathbf{y}$ (using any of the three discussed methods). How can we indicate how good this estimate actually is? We can do this, using **confidence regions**.

Suppose we have a region S. For example, S can be defined as the interval $[\hat{\mathbf{y}} - \epsilon, \hat{\mathbf{y}} + \epsilon]$. We now take a new measurement $\mathbf{y}$. Let's examine the chance that $\mathbf{y}$ is in the region S. This chance then is

$$P(\mathbf{y} \in S) = 1 - \alpha. \tag{4.2}$$

We now say that S is a $100(1 - \alpha)$ **percent confidence region**. (For example, if $\alpha = 0.01$, then it is a 99% confidence region.) Also $1 - \alpha$ is called the **confidence coefficient**.