Appendix 3

Parameter Estimation

There is much theory about parameter estimation, dealing with properties such as the bias and variance of the estimate. This theory is based on analysis of the probability density functions of the measurements and the parameter space. In this appendix, we discuss such topics as bias of an estimator, the variance, the Cramér-Rao lower bound on the variance, and the posterior distribution. The treatment will be largely informal, based on examples, and exploring these concepts in the context of reconstruction.

The general lesson to be learnt from this discussion is that many of these concepts depend strongly on the particular parametrization of the model. In problems such as 3D projective reconstruction, where there is no preferred parametrization, these concepts are not well defined, or depend very strongly on assumed noise models.

A simple geometric estimation problem. The problem we shall consider is related to the triangulation problem of determining a point in space from its projection into two images. To simplify this problem, however, we consider its 2-dimensional analog. In addition, we fix one of the rays reducing the problem to one of estimating the position of a point along a known line from observing it in a single image.

Thus, consider a line camera (that is, one forming a 1D image as in section 6.4.2-(*p*175)) observing points on a single line. Let the camera be located at the origin $(0, 0)$ and point in the positive Y direction. Further, assume that it has unit focal length. Thus, the camera matrix for this camera is simply $[\mathtt{I}_{2\times 2}|\mathbf{0}]$. Now, suppose that the camera is observing points on the line $\mathrm{Y} = \mathrm{X}+1$ (the "world line"). A point $(\mathrm{X}, \mathrm{X}+1)$ on this line will be mapped to the image point $x = \mathrm{X}/(\mathrm{X}+1)$. However, we assume that a point is measured with a certain inaccuracy, which may be modelled with a probability density function (PDF). The usual practice is to model the noise using a Gaussian distribution. Let us assume at least that the mode (maximum) of the distribution is at zero. The imaging setup is illustrated in figure A3.1.

The estimation problem we consider is the following: given the image coordinates of a point, estimate the position of the "world point" on the line $y = x + 1$. To consider a specific scenario, we may think of the line as a scintillator being flooded with gamma-rays. A camera is used to measure the location of each scintillation and determine its position along the line. The problem seems to be ridiculously simple, but it will turn out that there are some surprises.

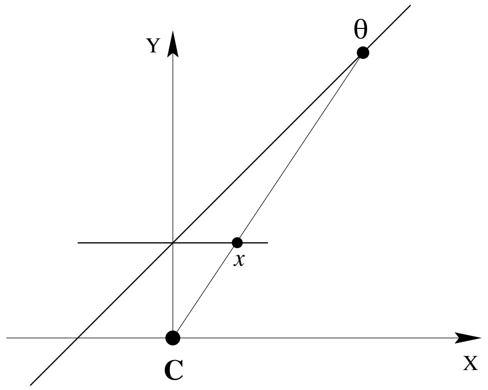

Fig. A3.1. *The imaging setup for a simple estimation problem. A point on the line* $\mathrm{Y} = \mathrm{X} + 1$ *is imaged by a line camera. The projection mapping is given by* $f : \theta \mapsto x = \theta/(1+\theta)$, *where* θ *parametrizes the points on the line* $\mathrm{Y} = \mathrm{X} + 1$. *Measurement is subject to noise with a zero-mode distribution.*

Probability density function. We start by parametrizing the world line $\mathrm{Y} = \mathrm{X} + 1$ by a parameter θ, where the most convenient parametrization is $\theta = \mathrm{X}$ so that the 2D point parametrized by θ is $(\theta, \theta + 1)$. This point projects to $\theta/(\theta + 1)$. We denote this projection function from the world line to the image line by f, so that $f(\theta) = \theta/(\theta+1)$. The measurement of this point is corrupted by noise, resulting in a random variable x with probability distribution given by $p(x|\theta) = g(x - f(\theta))$. For instance, if g is a zero-mean Gaussian distribution with variance σ^2, then

$$p(x|\theta) = (2\pi\sigma^2)^{-1/2} \exp\left(-(x - f(\theta))^2/2\sigma^2\right).$$

Maximum Likelihood estimate. An estimate of the parameter vector θ given a measured value x is a function denoted $\hat{\theta}(x)$, assigning a parameter vector θ to a measurement x. The maximum likelihood (ML) estimate is given by

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(x|\theta).$$

In the current estimation problem it is easily seen that the ML estimate is obtained simply by back-projecting the measured point x and selecting its intersection with the world line, according to the formula

$$\hat{\theta}(x) = f^{-1}(x) = x/(1 - x).$$

This is the ML estimate, because the resulting point, with parameter $\hat{\theta}(x)$, projects forward to x, hence $p(x|\hat{\theta}) = g(x-x) = g(0)$ which by assumption gives the maximum (mode) of the probability density function g. Any other choice of parameter θ will give a smaller value of $p(x|\theta)$.

A3.1 Bias

A desirable property of an estimator is that it can be expected to give the right answer on the average. Given a parameter θ, or equivalently in our case a point on the world line, we consider all possible measurements x and from them reestimate the parameter θ, namely $\hat{\theta}(x)$. The estimator is known as *unbiased* if on the average we obtain

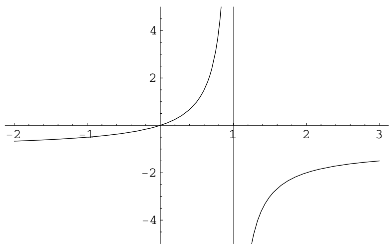

Fig. A3.2. *The ML estimate of the world-point position $\hat{\theta}(x) = f^{-1}(x) = x/(1-x)$ for different measurements of the image point x. Note that for values of x greater than 1, the ML estimate switches to "behind" the camera.*

the original parameter θ (the true value). In forming this average, we weight the measurements x according to their probability. More formally, the bias of the estimator is defined as

$$E_\theta[\hat{\theta}(x)] - \theta = \int_x p(x|\theta)\, \hat{\theta}(x) dx - \theta$$

and the estimator is unbiased if $E_\theta[\hat{\theta}(x)] = \theta$ for all θ. Here E_θ stands for the *expected value* given θ, defined as shown.

Another way of thinking of bias is in terms of repeating an experiment many times with identical model parameters and a different instance of the noise at each trial. The bias is the difference between the average value of the estimated parameters and the true parameter value. It is worth noting that for the bias to be defined, it is not necessary that the parameter space have an *a priori* distribution defined on it, not even that it be a measure space. It is necessary, however that it have some affine structure so that the average (or integral) can be formed.

Now, we determine whether the ML estimate of θ is unbiased in the case where $f(\theta) = \theta/(\theta+1)$. The integral becomes

$$\int_x p(x|\theta)\, \hat{\theta}(x) dx = \int_x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x - \theta/(\theta+1))^2}{2\sigma^2}\right) \frac{x}{1-x} dx$$

It turns out that this integral diverges, and hence the bias is undefined. The difficulty is that with an assumed Gaussian distribution of noise, for any value of θ, there is always a finite (though perhaps very small) probability $p(x|\theta)$ that $x > 1$. For values of $x > 1$, the corresponding ray does not meet the world line in front of the camera (since the ray is parallel to the world line at $x = 1$). The estimate $\hat{\theta}(x)$ as a function of x is shown in figure A3.2, showing how it results in estimated world-points behind the camera. Even if values of $\hat{\theta}(x)$ behind the camera are ignored, the ML estimator has infinite bias, as is explained in figure A3.3.

Limiting the range of parameters. Since the range of the parameter θ is from -1 to ∞, it makes sense to limit its range. In fact, we may have knowledge that all "events" on the world-line lie in a more restricted range. As an example, suppose that we assume that θ lies in the range $-1 \leq \theta \leq 1$, and hence noise-free projected points are in the range $-\infty < x < 1/2$. In this case, the ML estimate for any image point $x > 1/2$ will

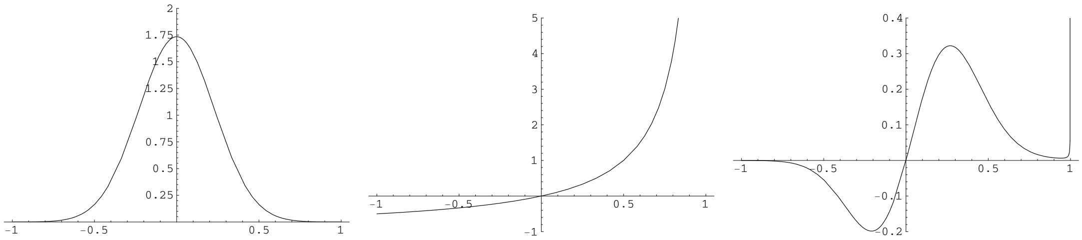

Fig. A3.3. **The reason that the ML estimate with Gaussian noise model has infinite bias.** *(a) The distribution of possible values of the image measurement given a world-point at $x = 0, y = 1$ assuming Gaussian noise distribution with $\sigma = 0.4$ – in symbols $p(x|\theta = 0)$. (b) The ML estimate of the world point for different values of the image point, $\hat{\theta}(x) = x/(1-x)$. Note that as the image point approaches 1, the estimated point on the world-line recedes to infinity. (c) Product $\hat{\theta}(x)\ p(x|\theta = 0)$. The integral of this function from $x = -\infty$ to $x = 1$ gives the bias of the ML estimator. Note that as x approaches 1, the graph increases abruptly to infinity. The integral is unbounded, meaning that the estimator has infinite bias.*

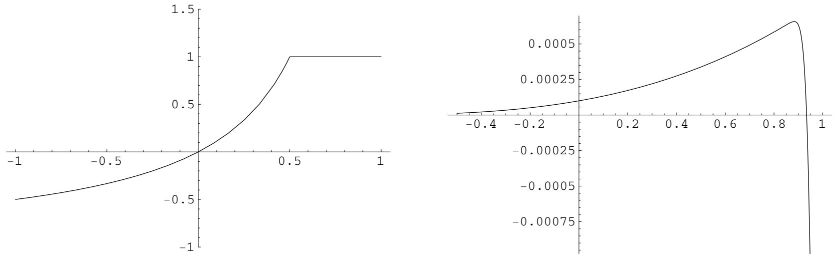

Fig. A3.4. *(a) If the range of possible values of the world-point parameter θ is limited to the range $\theta \leq 1$, then the ML estimate of any point $x > 1/2$ will be $\theta = 1$. This will prevent infinite bias in the estimate, but there will still be bias. (b) The bias $E_\theta[\hat{\theta}(x)] - \theta = \int_x \hat{\theta}(x)p(x|\theta)dx - \theta$ as a function of θ. Measurement noise is Gaussian with $\sigma = 0.01$.*

be at $\hat{\theta}(x) = 1$. With this restriction on the parameter θ, the ML estimate is still biased, as shown in figure A3.4.

If the noise-distribution has finite support, then the bias will also be finite for most values of θ, even if the range of the parameter θ is unrestricted. This is shown in figure A3.5. One learns from this that the bias of the estimator can be very dependent on the noise-model – a factor that is usually not within our control.

Dependency of bias on parametrization. The reason for the infinite bias for the ML estimator with Gaussian noise model in this example is the projective mapping between the world line and the image line. It is possible to parametrize the world line differently in a way that will change the bias.

Let the world line $\mathrm{Y} = \mathrm{X} + 1$ be parametrized in such a way that parameter θ represents the point $(\theta/(1-\theta), 1/(1-\theta))$ on the line. The part of the line in front of the camera (having positive y coordinate) is parametrized by θ in the range $-\infty < \theta < 1$. Under the projection $(\mathrm{X}, \mathrm{Y}) \mapsto \mathrm{X}/\mathrm{Y}$, the projection map f is given by $f(\theta) = \mathrm{X}/\mathrm{Y} = \theta$, thus the point with parameter θ maps to θ. In other words, points on the world line are

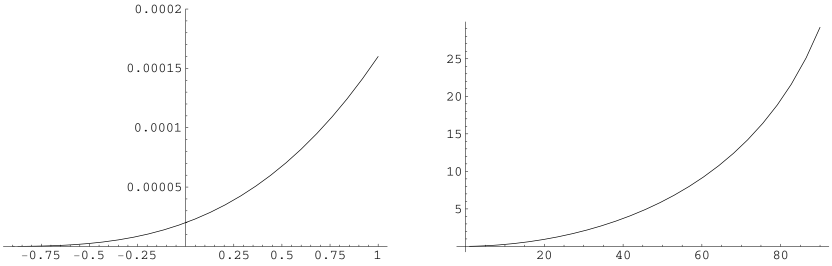

Fig. A3.5. *If the noise-model has finite support then the bias will be finite. In this example the noise model is* $3(1-(x/\sigma)^2)/4\sigma^2$, *for* $-\sigma \leq x \leq \sigma$, *where* $\sigma = 0.01$. *(a) Bias as a function of* θ *for small values of* θ. *(b) Percentage bias* $(E_\theta[\hat{\theta}(x)] - \theta)/\theta$. *The bias is relatively small for small values of* θ, *but larger (up to 20%) for large values. The position of the point along the world-line is always over-estimated.*

parametrized by the coordinate of the point that they project to under the camera mapping.

Now, in this case, the ML estimate is given by $\hat{\theta}(x) = f^{-1}(x) = x$, and the bias is

$$\begin{aligned}
\int_x \hat{\theta}(x)p(x|\theta)dx - \theta &= \int_x xg(x - f(\theta))dx - \theta \\
&= \int_x (x + f(\theta))g(x)dx - \theta \\
&= \int_x xg(x)dx + f(\theta)\int_x g(x)dx - \theta \\
&= 0 + f(\theta) - \theta = 0
\end{aligned}$$

Assuming that the distribution $g(x)$ is zero-mean. Note, this shows that the estimate of the *original* measurement x is unbiased.

Lessons about bias. By a change of parametrization of the world line, the bias has been changed from infinite to zero. In the present example, there is a natural affine parametrization of the world line, for which we have seen that the bias is infinite. However, if we are working in a projective context, then the parameter space has no natural affine parametrization. In this case, it is somewhat meaningless to speak of any absolute measurement of bias. A second lesson from the above example is that the ML estimate of the corrected measurement (as opposed to the world point) is unbiased.

It was also seen that the value of bias is strongly dependent on the noise distribution. Even the very small tails of the Gaussian distribution have a very large effect on the computed bias. Of course, a Gaussian distribution of the noise is only a convenient model for modelling image measurement errors. The exact distribution of image measurement noise is generally unknown, and the conclusion is inescapable that one can not theoretically compute an exact value for the bias of a given estimator, except for synthetic data with known noise model.

A3.2 Variance of an estimator

The other important attribute of an estimator is its variance. Consider an experiment being repeated many times with the same model parameters, but a different instantiation of the noise at each trial. Applying our estimator to the measured data, we obtain an estimate for each of these trials. The variance of the estimator is the variance (or covariance matrix) of the estimated values. More precisely, we can define the variance for an estimation problem involving a single parameter as

$$\mathrm{Var}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta}(x) - \bar{\theta})^2] = \int_x (\hat{\theta}(x) - \bar{\theta})^2 p(x|\theta)dx$$

where

$$\bar{\theta} = E_\theta[\hat{\theta}(x)] = \int_x \hat{\theta}(x)p(x|\theta)dx = \theta + \mathrm{bias}(\hat{\theta})$$

In the case where the parameters θ form a vector, $\mathrm{Var}_\theta(\hat{\theta})$ is the covariance matrix

$$\mathrm{Var}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta}(x) - \bar{\theta})(\hat{\theta}(x) - \bar{\theta})^\mathsf{T}] \tag{A3.1}$$

In many cases we might be more interested in the variability of the estimate with respect to the original parameter θ, which is the mean-squared error of the estimator. This is easily computed from

$$E_\theta[(\hat{\theta}(x) - \theta)(\hat{\theta}(x) - \theta)^\mathsf{T}] = \mathrm{Var}_\theta(\hat{\theta}) + \mathrm{bias}(\hat{\theta})\ \mathrm{bias}(\hat{\theta})^\mathsf{T}.$$

It should be noted that, as with the bias, the variance of an estimator makes good sense only when there is a natural affine structure on the parameter set, at least locally.

Most estimation algorithms will give the right answer if there is no noise. If an algorithm performs badly when noise is added, this means that either the bias or variance of the algorithm is high. This is the case, for instance with the DLT algorithm 4.1(*p*91), or the unnormalized 8-point algorithm discussed in section 11.1(*p*279). The variance of the algorithm grows quickly with added noise.

The Cramér-Rao lower bound. It is evident that by adding noise to a set of measurements information is lost. Consequently, it is not to be expected that any estimator can have zero bias and variance in the presence of noise on the measurements. For unbiased estimators, this notion is formalized in the Cramér-Rao lower bound, which is a bound on the variance of an unbiased estimator. To explain the Cramér-Rao bound, we need a few definitions. Given a probability distribution $p(x|\theta)$, the Fisher score is defined as $V_\theta(x) = \partial_\theta \log p(x|\theta)$. The Fisher Information Matrix is defined to be

$$\begin{aligned} F(\theta) &= E_\theta[V_\theta(x)V_\theta(x)^\mathsf{T}] \\ &= \int_x V_\theta(x)V_\theta(x)^\mathsf{T} p(x|\theta)dx. \end{aligned}$$

The relevance of the Fisher Information Matrix is expressed in the following result.

Result A3.1. Cramér-Rao lower bound. *For an unbiased estimator $\hat{\theta}(x)$,*

$$\det(E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\mathsf{T}]) \geq 1/\det F(\theta).$$

A Cramér-Rao lower bound may also be given in the case of a biased estimator.

A3.3 The posterior distribution

An alternative to the ML estimate is to consider the probability distribution for the parameters, given the measurements, namely $p(\theta|x)$. This is known as the posterior distribution, namely the distribution for the parameters *after* the measurements have been taken. To compute it, we need a *prior* distribution $p(\theta)$ for the parameters *before* any measurement has been taken. The posterior distribution can then be computed from Bayes Law

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}.$$

Since the measurement x is fixed, so is its probability $p(x)$, so we may ignore it, leading to $p(\theta|x) \approx p(x|\theta)\, p(\theta)$. The maximum of the posterior distribution is known as the Maximum A Posteriori (MAP) estimate.

Note. Though the MAP estimate may seem like a good idea, it is important to realize that it depends on the parametrization of the parameter space. The posterior probability distribution is proportional to $p(x|\theta)p(\theta)$. However, $p(\theta)$ is dependent on the parametrization of θ. For instance if $p(\theta)$ is a uniform distribution in one parametrization, it will not be a uniform distribution in a different parametrization that differs by a non-affine transformation. On the other hand, $p(x|\theta)$ does not depend on the parametrization of θ. Therefore, the result of a change of parametrization is to alter the posterior distribution in such a way that its maximum will change. If the parameter space does not have a natural affine coordinate system (for instance if the parameter space is projective) then the MAP estimate does not really make a lot of sense.

Other estimates based on the posterior distribution are also possible. Given a measurement x, and the posterior distribution $p(\theta|x)$, we may wish to make a different estimate of the parameter θ. One sensible choice is the estimate that minimizes the expected squared error in the estimate, namely

$$\hat{\theta}(x) = \mathrm{argmin}_Y E[\|Y - \theta\|^2] = \mathrm{argmin}_Y \int \|Y - \theta\|^2 p(\theta|x) d\theta,$$

which is the mean of the posterior distribution.

A further alternative is to minimize the expected absolute error

$$\hat{\theta}(x) = \mathrm{argmin}_Y E[\|Y - \theta\|] = \mathrm{argmin}_Y \int \|Y - \theta\| p(\theta|x) d\theta,$$

which is the median of the posterior distribution. Examples of these estimates are shown in figure A3.6 and figure A3.7.

More properties of these estimates are listed in the notes at the end of this appendix.

A3.4 Estimation of corrected measurements

We have seen that in geometric estimation problems, particularly those that involve projective models, concepts such as bias and variance of the estimator are dependent on the particular parametrization of the model, for instance a particular projective coordinate frame chosen. Even in cases where a natural affine parametrization of the model

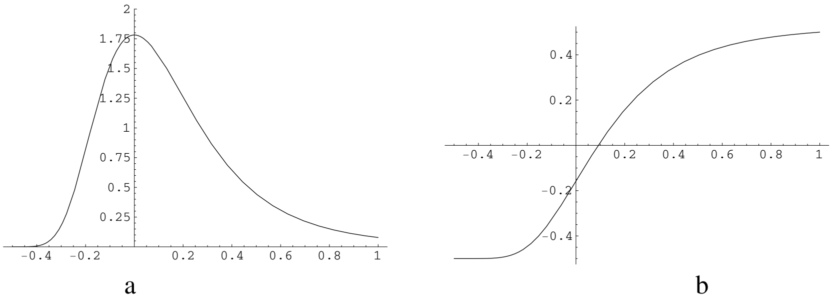

Fig. A3.6. **Different estimators for θ.** *(a) For the imaging setup of figure A3.1: the* a posteriori *distribution $p(\theta|x = 0)$ assuming a Gaussian noise distribution with $\sigma = 0.2$, and a prior distribution for θ, uniform on the interval $[-1/2, 1]$. The mode (maximum) of this distribution is the Maximal Apriori (MAP) estimate of θ, which is identical with the ML estimate, because of the assumed uniform distribution for θ. The mean of this distribution ($\theta = 0.1386$) is the estimate that minimizes the expected squared error $E[(\hat{\theta}(x) - \bar{\theta})^2]$ with respect to the true measurement $\bar{\theta}$. (b) The cumulative a posteriori distribution (offset by -0.5). The zero point of this graph is the median of the distribution, namely the estimate that minimizes $E[|\hat{\theta}(x) - \bar{\theta}|]$. The median lies at $\theta = 0.09137$.*

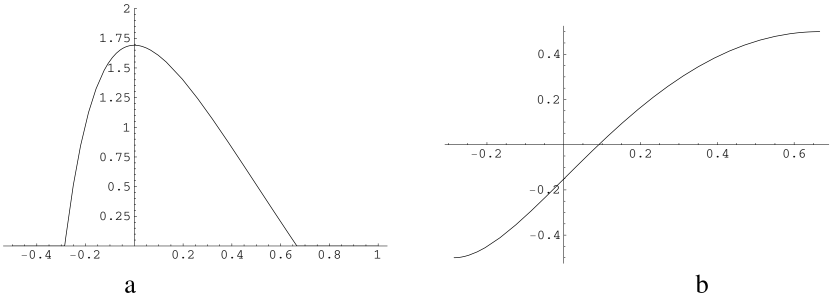

Fig. A3.7. **Different estimators with the parabolic noise model**. *The two graphs show the* a posteriori *distribution of θ and its cumulative distribution. In this example, the noise model is $3(1 - (x/\sigma)^2)/4\sigma^2$, for $-\sigma \leq x \leq \sigma$, where $\sigma = 0.4$. The mode of the distribution ($\theta = 0$) is the MAP estimate, identical with the ML estimate, the mean ($\theta = 0.1109$) minimizes the expected squared error in θ and the median ($\theta = 0.0911$) minimizes the expected absolute error.*

exists, it may be difficult to find an unbiased estimator. For instance, the ML estimator for problems such as triangulation is biased.

We saw however in the example of 1D back-projection discussed in section A3.1, that if instead of attempting to compute the model (namely the back-projected point) we estimate the corrected measurement instead, then the ML estimator is unbiased. We explore this notion further in this section, and show that in a general setting, the ML estimator of the corrected measurements is not only unbiased but attains the Cramér-Rao lower bound, when the noise-model is Gaussian.

Consider an estimation problem that involves fitting a parametrized model to a set of image measurements. As seen in section 5.1.3(p134) this problem may be viewed as an estimation problem in a high-dimensional Euclidean space $\mathbb{R}^N$, which is the space of all image measurements. This is illustrated in figure 5.2(p135). The estimation problem is, given a measurement vector $\mathbf{X}$, to find the closest point lying on a surface representing the set of all allowable exact measurements. The vector $\widehat{\mathbf{X}}$ represents the set of "corrected" image measurements that conform to the given model. The model itself may depend on a set of parameters θ, such as the fundamental matrix, hypothesized 3D points or other parameters appropriate to the problem.

The estimation of the parameters θ is subject to bias in the same way as we have seen with the simple problem discussed in section A3.1, and the exact degree of bias is dependent on the precise parametrization. Generally (for instance in projective-reconstruction problems) there is no natural affine coordinate system for the model parameters, though there is a natural affine coordinate system for the image plane.

If we think of the problem differently, as the problem of directly finding the corrected measurement vector $\widehat{\mathbf{X}}$, then we find a more favourable situation. The measurements are carried out in the images, which have a natural affine coordinate system, and so questions of bias in estimating the corrected measurements make more sense. We will show that, provided the measurement surface is well approximated by its tangent plane, the ML estimate of the corrected measurement vector is unbiased, provided the noise is zero-mean. In addition, if the noise is Gaussian and isotropic, then the ML estimate meets the Cramér-Rao lower-bound.

The geometric situation is follows. A point $\overline{\mathbf{X}}$ lies on a measurement surface, as shown in figure 5.2(p135). Noise is added to this point to obtain a measured point $\mathbf{X}$. The estimate $\widehat{\mathbf{X}}$ of the true point $\overline{\mathbf{X}}$ is obtained by selecting the closest point on the measurement surface to the measured point. We make an assumption that the measurement surface is effectively planar near $\mathbf{X}$.

We may choose a coordinate system in which the measurement surface close to $\overline{\mathbf{X}}$ is spanned by the first d coordinates. We may write $\overline{\mathbf{X}} = (\overline{\mathbf{X}}_1^{\mathsf{T}}, \mathbf{0}_{N-d}^{\mathsf{T}})^{\mathsf{T}}$, where $\overline{\mathbf{X}}_1$ is a d-vector. The measured point may similarly be written as $\mathbf{X} = (\mathbf{X}_1^{\mathsf{T}}, \mathbf{X}_2^{\mathsf{T}})^{\mathsf{T}}$, and its projection onto the tangent plane is $\widehat{\mathbf{X}} = (\mathbf{X}_1^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}})^{\mathsf{T}} = (\widehat{\mathbf{X}}_1^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}})^{\mathsf{T}}$. We suppose that the noise-distribution is given by $p(\mathbf{X}|\overline{\mathbf{X}}) = g(\mathbf{X} - \overline{\mathbf{X}})$. Now, the bias of the estimate $\widehat{\mathbf{X}}$ is

$$\begin{aligned} E[(\widehat{\mathbf{X}} - \overline{\mathbf{X}})] &= \int_{\mathbf{X}} (\widehat{\mathbf{X}} - \overline{\mathbf{X}}) p(\mathbf{X}|\overline{\mathbf{X}}) d\mathbf{X} = \int_{\mathbf{X}} (\widehat{\mathbf{X}} - \overline{\mathbf{X}}) g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \\ &= \int_{\mathbf{X}} \mathsf{J}(\mathbf{X} - \overline{\mathbf{X}}) g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \\ &= \mathsf{J} \int_{\mathbf{X}} (\mathbf{X} - \overline{\mathbf{X}}) g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \\ &= \mathbf{0} \end{aligned}$$

where J is the matrix $[\mathtt{I}_{d\times d} | \mathtt{0}_{d\times(N-d)}]$. This shows that the estimate of $\mathbf{X}$ is unbiased as long at g has zero-mean. The variance of the estimate is equal to

$$\begin{aligned} E[(\widehat{\mathbf{X}} - \overline{\mathbf{X}})(\widehat{\mathbf{X}} - \overline{\mathbf{X}})^{\mathsf{T}}] &= \int_{\mathbf{X}} (\widehat{\mathbf{X}} - \overline{\mathbf{X}})(\widehat{\mathbf{X}} - \overline{\mathbf{X}})^{\mathsf{T}} p(\mathbf{X}|\overline{\mathbf{X}}) d\mathbf{X} = \int_{\mathbf{X}} (\widehat{\mathbf{X}} - \overline{\mathbf{X}})(\widehat{\mathbf{X}} - \overline{\mathbf{X}})^{\mathsf{T}} g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \\ &= \int_{\mathbf{X}} \mathsf{J}(\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^{\mathsf{T}} \mathsf{J}^{\mathsf{T}} g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \\ &= \mathsf{J} \int_{\mathbf{X}} (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^{\mathsf{T}} g(\mathbf{X} - \overline{\mathbf{X}}) d\mathbf{X} \mathsf{J}^{\mathsf{T}} \\ &= \mathsf{J} \Sigma_g \mathsf{J}^{\mathsf{T}} \end{aligned}$$

where Σ_g is the covariance matrix of g.

We now compute the Cramér-Rao lower bound for this estimator, supposing that the distribution $g(\mathbf{X})$ is Gaussian defined by $g(\mathbf{X}) = k\exp(-\|\mathbf{X}\|^2/2\sigma^2)$. In this case, the variance of the estimator is simply $\sigma^2 \mathtt{I}_{d\times d}$.

We next compute the Fisher information matrix. The probability distribution is

$$\begin{aligned} p(\mathbf{X}|\overline{\mathbf{X}}) &= g(\mathbf{X}-\overline{\mathbf{X}}) = k\exp(-\|\mathbf{X}-\overline{\mathbf{X}}\|^2/2\sigma^2) \\ &= k\exp(-\|\mathbf{X}_1-\overline{\mathbf{X}}_1\|^2/2\sigma^2)\exp(-\|\mathbf{X}_2\|^2/2\sigma^2). \end{aligned}$$

Taking logarithms and derivatives with respect to $\overline{\mathbf{X}}_1$ gives

$$\partial_{\overline{\mathbf{X}}_1} \log p(\mathbf{X}|\overline{\mathbf{X}}) = -(\mathbf{X}_1 - \overline{\mathbf{X}}_1)/\sigma^2.$$

The Fisher information matrix is then

$$1/\sigma^2 \int (\mathbf{X}_1-\overline{\mathbf{X}}_1)(\mathbf{X}_1-\overline{\mathbf{X}}_1)^{\mathsf{T}} g(\mathbf{X}_1-\overline{\mathbf{X}}_1) g(\mathbf{X}_2) d\mathbf{X}_1 d\mathbf{X}_2 \quad = \mathtt{I}_{d\times d}/\sigma^2.$$

Thus, the Fisher information matrix is the inverse of the covariance matrix for the estimator. Thus, for the case where the noise distribution is Gaussian, the ML estimator meets the Cramér-Rao lower bound, to the extent that the measurement surface is flat.

It should be noticed that the Fisher Information Matrix does not depend on the specific shape of the measurement surface, but only on its first-order approximation, the tangent plane. The properties of the estimate does however depend on the shape of the measurement surface, both as regards its bias and variance. It may also be shown that if the Cramér-Rao bound is met, then the noise distribution must be Gaussian. In other words, if the noise distribution is not Gaussian, then we can not meet the lower bound.

A3.5 Notes and exercises

(i) Show by a specific example that the a posteriori distribution is altered by a change of coordinates in the parameter space. Show also that the mean of the distribution may be altered by such a coordinate change. Thus the mode (MAP estimate) and mean of the posterior distribution are dependent on the choice of coordinates for the parameter space.

(ii) Show for any PDF $p(\theta)$ defined on $\mathbb{R}^n$, that $\mathrm{argmin}_Y \int \|Y-\theta\|^2 p(\theta)d\theta$ is the mean of the distribution.

(iii) Show for any PDF $p(\theta)$ defined on $\mathbb{R}$, that $\mathrm{argmin}_Y \int |Y-\theta| p(\theta)d\theta$ is the median of the distribution. In higher dimensions, show that $\int \|Y-\theta\| p(\theta)d\theta$ is a convex function of Y, and hence has a single minimum. The value of Y that minimizes this is a higher-dimensional generalization of the median of a 1-dimensional distribution.

(iv) Show that the median of a PDF defined on $\mathbb{R}$ is invariant to reparametrization of $\mathbb{R}$. Show by an example that this is not true in higher dimensions, however.