10. Maximum Likelihood Estimation

While the term *likelihood* is often used as a synonym of the term *probability* in ordinary conversation, in the extraordinary world of statistics the two terms mean similar but clearly distinct things. To see this, consider a problem where we are given a data vector x and asked to estimate a parameter vector θ. The ideal solution would be to find the parameter vector θ that maximizes the probability of θ given data x, the so called *a posteriori* probability, which is written as

$$p(\theta|x).$$

While this is in many respects indeed the best approach, we will consider it later, and focus for now on a seemingly lesser approach that maximizes

$$p(x|\theta)$$

which is called the likelihood of θ and is written as

$$L(\theta|x)$$

or more often as

$$L(\theta).$$

The difference between the probability of θ given the data and the likelihood is quite obvious after these definitions. One of the subtler differences that occasionally manifest themselves is that while the probability integrates to unity by definition, the likelihood does not. In fact it might not even be integrable. This means that it can quite forgiving to some abuses and occasionally tolerates some significant mathematical acrobatics and we still maintain mathematical consistency.

Seems that this technique of setting up the formula for the likelihood and maximizing it allows one to manipulate the mathematical expressions quite freely. The practical significance of this freedom is that quite often we get simple and elegant closed form solutions, even to complex problems. Some of these problems would have much more complex solutions if we used other methods.

Before we show some examples, we should list the advantages of the Maximum Likelihood method. It is truly essential that the reader is indoctrinated in the solidly founded theory of the advantages of a statistical method before the reader becomes intimately acquainted with the method and realize that some of them have no redeemable value. Here we go:

(1) Often leads to simple equations without approximations or further compromises. Of course this depends on how one sets up the problem, but chances are clearly higher with Maximum Likelihood than many other methods (well there are not that many)

(2) The method often cannot incorporate any prior knowledge without losing the above advantage. But in many cases the advantage of a simple, efficient and exact solution allows one to use more data for the same amount of computation and

thus outweigh the lack of a prior.

(3) It is asymptotically unbiased.

(4) It has asymptotically minimum variance.

(5) It is asymptotically Gaussian

The first two advantages are vague promises, but they represent real advantages when they materialize. The rest are well known properties that are very briefly stated in this modest document, but rest assured that mathematicians, hordes of them, have worked hard for many long nights to clad them in stainless steel rigor and obfuscated mathematical lore in a vain attempt to hide the fact that these properties are only warm and fuzzy. Just this and nothing more.

But to be fair to the method, many well known algorithms have their roots in (or at least they can be proved with) the Maximum Likelihood method. Among them the Kalman filter, the Wiener filter, EM Clustering, K-means, etc. And the warm and fuzzy feeling of some of the properties of Maximum Likelihood kept researchers going in their cold, dump labs until they discovered some great things.

10.1. A Simple Example

We can demonstrate the abilities of the method with a very simple example, the one that most expositions of the method present as a first example.

Assume that we re given N vectors x_i for $i = 1..N$, we are told that these x_i's are i.i.d. (independent identically distributed) random variables that follow a Gaussian distribution and our job is to find the mean μ and covariance matrix C of this Gaussian distribution. We already know the answer of course, so this exercise will help us build confidence towards the method.

The probability density of a data vector x_i is

$$p(x_i|\mu, C) = \frac{1}{\sqrt{2\pi|C|}} e^{-\frac{(x-\mu)^T C^{-1}(x-\mu)}{2}}$$

and since these x_i's are i.i.d. the probability of the whole collection of them is

$$p(x_1, x_2, \cdots, x_N|\mu, C) = \prod_{i=1}^{N} p(x_i|\mu, C)$$

and so the likelihood of the parameters is

$$L(\mu, C) = \prod_{i=1}^{N} p(x_i|\mu, C).$$

Unfortunately typesetting monstrous equations is not the authors forte, so some simplifications are in order. The most obvious, most universal and most effective is to minimize the logarithm of the likelihood, the log-likelihood, instead of the straight likelihood. This way the products turn into sums and taking derivatives is much simpler. So, shall we start?

$$\ln L(\mu, C) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln|C| - \frac{1}{2}\ln 2\pi - \frac{1}{2}(x_i-\mu)^T C^{-1}(x_i-\mu)\right)$$

and taking the derivative first with respect to μ

$$\frac{\partial \ln L(\mu, C)}{\partial \mu} = \sum_{i=1}^{N}\left(C^{-1}(x_i-\mu)\right) = 0$$

which gives us the very familiar expression

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

We next take the derivative with respect to the covariance matrix C, but this time we are have to do some real work

$$\frac{\partial \ln L(\mu, C)}{\partial C} = \sum_{i=1}^{N}\left(-\frac{1}{2}\frac{\partial \ln|C|}{\partial C} - \frac{1}{2}\frac{\partial (x_i-\mu)^T C^{-1}(x_i-\mu)}{\partial C}\right)$$

since we have no idea how to take derivatives of determinants and inverse matrices with respect to matrices. If one observes that the determinant is just the sum of all the products of the elements of the matrices with their corresponding minor matrices and that the inverse times the matrix is the identity matrix, one is definitely a second millennium fossil. Now-A-Days we google. After we skip over pages referring to similarly named movies, we get

$$\frac{\partial \ln|C|}{\partial C} = C^{-1}$$
$$\frac{\partial (x_i-\mu)^T C^{-1}(x_i-\mu)}{\partial C} = -C^{-1}(x_i-\mu)(x_i-\mu)^T C^{-1}$$

and thus

$$\frac{\partial \ln L(\mu, C)}{\partial C} = \frac{1}{2}\sum_{i=1}^{N}\left(-C^{-1} + C^{-1}(x_i-\mu)(x_i-\mu)^T C^{-1}\right) = 0$$

and after pre and post multiplying with matrix C we get the equally familiar

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T.$$

Anybody good with statistics knows that this is a biased estimate of the covariance which proves by example that Maximum Likelihood does not always produce an unbiased estimate. In fact it rarely does in practice but then unbiasedness is such a futile thing to attempt.

Some observations are worth mentioning now. First of all this is one of the few cases where the normalizing factor of a Gaussian plays a role in some computation or derivation. Most often, but not always, we can just ignore it. Second, it seems that this example is rather unusual in another way. We rarely care about parameters of probability distributions, what we prefer to estimate is motions, shapes etc, what real people want.

This is not strictly true! Most of the times what we estimate is a quantity directly related to the mean of some distribution and occasionally to its variance.

Although computing the covariance matrix does not seem much work, we might get more accurate results if someone gave us the values of a few components of this matrix. It is not unusual in practice to know either the values of a few components, know that this matrix has a special form or have preferred ranges of values for some particular components. So let us assume that a fairy told us that the off diagonal component $C[k,l] = \alpha$ and what we have now is a constrained minimization. We add the terms

$$\lambda_1(C[k,l] - \alpha) + \lambda_2(C[l,k] - \alpha)$$

to the log-likelihood in classic Lagrangian fashion and the derivative with respect to the covariance becomes

$$\frac{\partial \ln L(\mu, C)}{\partial C} = \frac{1}{2}\sum_{i=1}^{N}\left(-C^{-1} + C^{-1}(x_i - \mu)(x_i - \mu)^T C^{-1} + \lambda_1 \hat{e}_l \hat{e}_k^T + \lambda_2 \hat{e}_k \hat{e}_l^T\right) = 0$$

for which no combination of pre and post multiplications is going to give us a simple closed form solution. So, the moral of the story is: Do not believe in fairies.

10.2. A Not-that-simple Example

Ready for a deep water test? Great, let us all grab an anchor and jump in the water! We are going to compute the maximum likelihood estimate of the depth of a point given its correspondence in two frames and the rotation and translation between the frames.

Let the position of our point in the camera coordinate system at time $N-1$ be X_{N-1} and at time N be X_N. The corresponding projections on the image plane will be

$$x_{N-1} = K\,\frac{X_{N-1}}{\hat{Z}^T X_{N-1}}$$

$$x_N = K\,\frac{X_N}{\hat{Z}^T X_N}$$

where K is a 2×3 calibration camera that takes care of the scaling and shifting that is involved in the projection. Further, let R_N and T_N be the known motion parameters, rotation matrix and translation vector, of the object that contains our point so that

$$X_N = RX_{N-1} + T.$$

We all know that we can project a 3-D point to the image plane but we cannot project an image point back to the 3-D world without some additional information namely the depth (or the Z component) of the point. Since it is slightly more convenient we will use the inverse depth but everything could be done almost as well with straight depths. The inverse depth at time $N-1$ is

$$\zeta_{N-1} = \frac{1}{\hat{Z}^T X_{N-1}}$$

and this is our unknown. We do not know the depth at time N either but we do not care to recover it and so do not bother with it at all. We are given the image position of the point at time $N-1$ but its corresponding point at time N is only known with some variance ${}^{x}C_N$. This means that the image point x_N has a probability density

$$p(x_N|\zeta_{N-1}) = \frac{1}{\sqrt{(2\pi)^2|C|}} e^{-\frac{(x_N-{}^{x}\mu_N)^T {}^{x}C_N^{-1}(x_N-{}^{x}\mu_N)}{2}}.$$

This probability density is our likelihood, and true to the ideals of statistics, we maximize its logarithm instead of the likelihood itself. This log-likelihood is, after omitting all the constants that do not interfere with the maximization

$$-(x_N - {}^{x}\mu_N)^T \ {}^{x}C_N^{-1}(x_N - {}^{x}\mu_N) \tag{10.1}$$

a true least squares expression. Alas, though! We do not know this ${}^{x}\mu_N$. The secret here is: Do Not Panic. This ${}^{x}\mu_N$ is the mean of the distribution and if we knew the depth, since we are given everything else that is relevant, we could find where the point is in 3-D and project it to the camera at time N. This projection is the mean of the distribution. Let's compute it then.

The inverse depth and image point at time $N-1$ along with the camera calibration matrix

$$\zeta_{N-1} = \frac{1}{\hat{Z}^T X_{N-1}}$$
$$x_{N-1} = K \frac{X_{N-1}}{\hat{Z}^T X_{N-1}}$$

can easily give us the 3-D vector of the point with respect to the coordinate system. There are half a dozen ways to write the expression, so we roll a dice (which conveniently has half a dozen facets) and write

$$X_{N-1} = \frac{K^+ \begin{bmatrix} x_{N-1} \\ \cdots\cdots \\ 1 \end{bmatrix}}{\zeta_{N-1}}$$

where

$$K^+ = \begin{bmatrix} K \\ \cdots\cdots \\ 0\ 0\ 1 \end{bmatrix}^{-1}$$

and by using the rigidity equation we write

$$X_N = R \frac{K^+ \begin{bmatrix} x_{N-1} \\ \cdots\cdots \\ 1 \end{bmatrix}}{\zeta_{N-1}} + T$$

and because ${}^x\mu_N$ is really x_N if we knew the depth

$${}^x\mu_N = K \frac{R \frac{K^+ \begin{bmatrix} x_{N-1} \\ \cdots\cdots \\ 1 \end{bmatrix}}{\zeta_{N-1}} + T}{\hat{Z}^T (R \frac{K^+ \begin{bmatrix} x_{N-1} \\ \cdots\cdots \\ 1 \end{bmatrix}}{\zeta_{N-1}} + T)}.$$

To avoid typesetting and keep the math legible we rewrite it as

$${}^x\mu_N = \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2}$$

where vectors V_1 and V_2 and scalars s_1 and s_2 correspond to their bulkier counterparts above. Going back to Eq. (10.1) we have to minimize

$$\left(x_N - \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2} \right)^T {}^x C_N^{-1} \left(x_N - \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2} \right)$$

which we do by taking derivatives

$$\frac{\partial}{\partial \zeta_{N-1}} \left(x_N - \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2} \right)^T {}^x C_N^{-1} \left(x_N - \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2} \right) =$$
$$-2 \frac{V_2(s_1 + \zeta_{N-1} s_2) - (V_1 + \zeta_{N-1} V_2) s_2}{(s_1 + \zeta_{N-1} s_2)^2} {}^x C_N^{-1} \left(x_N - \frac{V_1 + \zeta_{N-1} V_2}{s_1 + \zeta_{N-1} s_2} \right) = 0$$

and that after several obvious simplifications we get

$$\zeta_{N-1} = \frac{(V_2 s_1 - V_1 s_2)^T {}^x C_N^{-1} (s_1 x_{N-1} - V_1)}{(V_2 s_1 - V_1 s_2)^T {}^x C_N^{-1} (s_2 x_{N-1} - V_2)}$$

and we can plug back the definitions for vectors V_1 and V_2 and scalars s_1 and s_2 and squeeze a tiny bit of extra simplifications.

Now the question in everybody's mind is whether we can introduce a prior of any kind to the minimization without an exorbitant penalty. The answer is no, but most often we do not have a choice for motion problems. We despairfully need any kind of information to make it work, and beggars are not choosers.

10.3. Applying Maximum Likelihood

We have seen two examples of application of Maximum Likelihood and both are kind of different. How is Maximum Likelihood applied in practice? Is there a common pattern? It turns out that the vast majority of the applications of this technique fall into the style of the last example above. We have a set of parameters θ that we want to estimate and our model connects them to the mean of some measurements $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$. Almost always, the underlying probability distribution is either a gaussian or a variant of it and maximizing the likelihood is equivalent to minimizing

$$\ln L(\theta) = (x - \mu(\theta))^T C^{-1} (x - \mu(\theta)) \qquad \text{pplMaxLikeEq}$$

where the covariance matrix C is assumed known and constant (independent of θ). This looks like a least squares problem and unless $\mu(\theta)$ is linear, it is a non-linear least squares problem. Taking derivatives we obtain the normal equations

$$-\nabla \ln L(\theta) = (\nabla \mu(\theta))^T C^{-1} (x - \mu(\theta)) = 0.$$

This all sounds very simple, but how does one do it? Well, all the "computer vision" of the process is hidden inside the function $\mu(\theta)$. The parameter vector θ could contain the motion parameters, the depth (as above) or a parametric form of the motion or depth. Vector θ could also contain the set of parameters that define an object in a recognition problem (e.g. given that the object in front of the camera is an orthogonal prism, what is its orientation in space or its proportions), or even some intermediate result of some estimation that does not have a an immediate utility of its own.

Sometimes, though, we are not so lucky and end up with models that have both the mean and covariance matrix dependent on θ. In such unfortunate cases one can attempt a fine act of desperation and mount a heroic frontal attack by differentiating the log likelihood

$$-\frac{\partial \ln L(\theta)}{\partial \theta_i} = \frac{\partial \mu(\theta)}{\partial \theta_i}^T C^{-1}(x - \mu(\theta)) + \frac{1}{2} tr\left(C^{-1} \frac{\partial C}{\partial \theta_i}\right) + \frac{1}{2}(x - \mu(\theta))^T C^{-1} \frac{\partial C}{\partial \theta_i} C^{-1}(x - \mu(\theta))$$

where we do not even dare take the derivatives with respect to vector θ and instead settled for a derivative with respect to an element θ_i of θ.

10.4. Multiple Sources of Information

Whenever we have two independent measurements x and y we tend to write the likelihood as

$$L(\theta) = p(x, y|\theta) = p(x|\theta)p(y|\theta)$$

without blinking. This is essentially playing with the assumptions a bit. If x and y were independent, then how come they both depend on θ? What we really mean here is that the measurements x and y are both related to θ but any randomness in x is independent of

any randomness in y. Someone had the excellent idea to call such a thing *conditional independence*. And two random variables are conditionally independent on θ iff

$$p(x, y|\theta) = p(x|\theta)p(y|\theta)$$

which is exactly what we need. This assumption is routinely used to handle "independent" measurements and several other things, but it is not always a reasonable (i.e. not very accurate) or productive (i.e. might lead to a more complex algorithm, or even worse, to a preexisting one) assumption.

One further reason to love this log in front of the likelihood is that the log-likelihoods from different sources are additive. Writing the log-likelihood with explicit reference to the source data

$$\ln L(\theta|x, y) = \ln p(x, y|\theta) = \ln p(x|\theta) + \ln p(y|\theta) = \ln L(\theta|x) + \ln L(\theta|y). \quad (10.3)$$

an observation that will save us trouble later.

We can try our luck with what we learned in Sec. 10.3 and 10.4. Our first take will be simple. We consider $\mu(\theta) = \theta$ in Eq. (10.2) and x and y are gaussian with mean θ and variances C_x and C_y respectively. So Eq. (10.3) becomes

$$\begin{aligned}
&2(\ln L(\theta|x) + \ln L(\theta|y)) = \\
&(x-\theta)^T C_x^{-1}(x-\theta) + (y-\theta)^T C_y^{-1}(y-\theta) = \\
&x^T C_x^{-1} x + y^T C_y^{-1} y - 2\theta^T C_x^{-1} x - 2\theta^T C_y^{-1} y + \theta^T C_x^{-1}\theta + \theta^T C_y^{-1}\theta = \\
&x^T C_x^{-1} x + y^T C_y^{-1} y - 2\theta^T\left(C_x^{-1} x + C_y^{-1} y\right) + \theta^T\left(C_x^{-1} + C_y^{-1}\right)\theta = \\
&x^T C_x^{-1} x + y^T C_y^{-1} y - \mu_\theta^T C_\theta^{-1}\mu_\theta + \\
&\mu_\theta^T C_\theta^{-1}\mu_\theta - 2\theta^T C_\theta^{-1}\mu_\theta + \theta^T C_\theta^{-1}\theta
\end{aligned}$$

where

$$\begin{aligned}
C_\theta &= \left(C_x^{-1} + C_y^{-1}\right)^{-1} \\
\mu_\theta &= C_\theta\left(C_x^{-1} x + C_y^{-1} y\right)
\end{aligned} \quad (10.4)$$

and thus

$$\begin{aligned}
&2(\ln L(\theta|x) + \ln L(\theta|y)) = \\
&const + \mu_\theta^T C_\theta^{-1}\mu_\theta - 2\theta^T C_\theta^{-1}\mu_\theta + \theta^T C_\theta^{-1}\theta = \\
&const + (\theta-\mu_\theta)^T C_\theta^{-1}(\theta-\mu_\theta).
\end{aligned}$$

The result above is interesting because it tells us that the likelihood of the unknown parameters that we gleaned from two Gaussian sources looks like a Gaussian itself, although it is not necessarily a Gaussian since it involves a constant that prevents it from

integrating to unity. Eq. (10.4) gives us a nice way to combine information from the two Gaussian sources which is not too complicated. But it has a disadvantage that we can easily fix. In a situation that new information comes with a regularity each time we have to invert the old covariance matrix, invert the new covariance matrix, add them and invert the result. Similarly with the mean. This is a lot of inversions. We can avoid most of them if we simply keep track of the inverse of the covariance instead of the covariance and the product of the inverse of the covariance and the mean instead of the mean. The first we call information and the latter we call score:

$$
\begin{aligned}
I_x &= C_x{}^{-1} \\
I_y &= C_y{}^{-1} \\
S_x &= C_x{}^{-1}\mu_x \\
S_y &= C_y{}^{-1}\mu_y \\
I_\theta &= I_x + I_y \\
S_\theta &= S_x + S_y
\end{aligned}
$$

and as to why we call them information and score, you have to wait until the next section. Anyhow the log likelihood of θ can be written

$$const - 2\theta^T S_\theta + \theta^T I_\theta \theta.$$

But before we get there it would be nice to explore a more general case rather than $\mu(\theta) = \theta$. Unfortunately, we can only explore the situation $\mu_i(\theta) = \Lambda_i\theta$ where Λ_i for $i = x, y$, is an arbitrary matrix. The (almost) only way we can attack non-linear μs is with linear approximations, so this will be the most general we will go.

If Λ_i were invertible, then this further exploration would be trivial. The most interesting situations though involve matrices Λ_i that are not even square which while non trivial, are certainly not hard. The log likelihood of θ given x is

$$
\begin{aligned}
&2\ln L(\theta\,|x) = \\
&(x - \Lambda_x\theta)^T C_x{}^{-1}(x - \Lambda_x\theta) = \\
&x^T C_x{}^{-1} x - 2\theta^T \Lambda_{x_T} C_x{}^{-1} x + \theta^T \Lambda_{x_T} C_x{}^{-1} \Lambda_x \theta + \\
&const - 2\theta^T S_x + \theta^T I_x \theta
\end{aligned}
$$

where

$$
\begin{aligned}
S_x &= \Lambda_{x_T} C_x^{-1} x \\
I_x &= \Lambda_{x_T} C_x^{-1} \Lambda_x
\end{aligned}
$$

and if we do the same with the log likelihood of y we get if we add the two log likelihoods

$$2\ln L(\theta\,|x) + 2\ln L(\theta\,|y) =$$
$$const - 2\theta^T(S_x + S_y) + \theta^T(I_x + I_y)\theta$$
$$const - 2\theta^T S_\theta + \theta^T I_\theta \theta$$

where

$$S_\theta = S_x + S_y = \Lambda_x^T {C_x}^{-1} x + \Lambda_y^T {C_y}^{-1} y$$
$$I_\theta = I_x + I_y = \Lambda_x^T {C_x}^{-1} \Lambda_x + \Lambda_y^T {C_y}^{-1} \Lambda_y \,.$$

It is worth noticing that although I_x exists, its inverse may not. The same for I_y. They are both non negative definite but not always positive definite. Their sum is also non negative definite but not always provably positive definite. In fact, before we incorporate all the data in our likelihood, the partial sum of the I_is may not be invertible. In practice this non-invertibility means that we have infinite uncertainty along some dimensions of our problem.

10.5. Variance of the Estimate

No estimate is good unless you know how much you can trust it. And this is not just curiosity. Unless we know the variance we cannot combine information from two different sources even if everything is gaussian. So having a measure of our confidence is not a luxury, it is a necessity. And as always the most common and convenient measure of confidence is the variance for unidimensional estimates and the covariance matrix for multidimensional ones.

We would really want a simple and general way to compute the covariance matrix and here the record is mixed. There is no such general way known either to humans or to mathematicians and, to make things worse, it is not easy to come up with even *ad hoc* solutions to specific problems. On the other hand, there is a very descent approximation to the covariance, one that is rather unlikely to embarrass anybody. It involves the *Fisher Information Matrix*, a concept normally understood only by the initiates to the rites and rituals of Statistics of the most inner circle. The astonishing importance of this matrix lies in exactly three facts:

(1) It is named after Fisher, the most important mathematician of the twentieth century.

(2) Sounds like the thingy also called information which was developed by Shannon, the most important mathematician of the twentieth century

(3) It is defined in a rather cryptic way, to make it opaque to the minds of anybody who is not the most important mathematician of the twentieth century.

And since the reader of this modest text has lost their chance to become the most important mathematician of the twentieth century (unless they are already) we a going to use slightly less cryptic language to present this quite useful, and in fact rather simple, concept. If the reader has any doubt about the simplicity of this concept, they should compare the date of birth of Fisher and the date of the first publication of his ideas. He was almost

a kid really and kids have a taste for simple things, like potato chips and milkshakes.

While there is no theorem or anything that tells us that the approximation using the Fisher information matrix is within some bounds (and usually these theorems come with so many stipulations, provisions, catches and fine print that make them useless), we have good reason to hope that it is good because it is exact in many situations. It is exact when the underlying distribution is gaussian and certain critical functions linear, it is exact at the limit (infinite amount of data), is the lower bound if the estimator is unbiased. But most important it is very intuitive, works quite well in practice and there are too few alternatives (so we are stuck).

The first relevant result is the *Gaussianity of the Likelihood*. It is possible to show that the likelihood $L(\theta)$ looks like a scaled gaussian if we use many-many iid data. This is an almost direct consequence of the *Central Limit Theorem*, but it is easy to recount the basic steps of the proof without invoking the Central Limit Theorem. Let $\hat{\theta}$ be the value of the unknown parameters that maximizes the likelihood, in other words the maximum likelihood estimate of θ that makes the derivative of the likelihood equal to zero. If we do a second order approximation (first order approximation reduces to a constant near a maximum) to the likelihood

$$\ln L(\theta) = \sum_{i=1}^{N} \ln p(x_i|\theta) =$$

$$\sum_{i=1}^{N} \ln p(x_i|\hat{\theta}) + \sum_{i=1}^{N} \left[\frac{\partial}{\partial \theta} \ln p(x_i|\theta) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) +$$

$$\frac{1}{2} (\theta - \hat{\theta})^T \sum_{i=1}^{N} \left[\frac{\partial^2}{\partial \theta \partial \theta^T} \ln p(x_i|\theta) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

and since the first derivative at the maximizing point $\theta = \hat{\theta}$ is zero

$$\ln L(\theta) = \sum_{i=1}^{N} \ln p(x_i|\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^T \sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \theta^T} \ln p(x_i|\theta = \hat{\theta})(\theta - \hat{\theta})$$

and setting

$$I = -\sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \theta^T} \ln p(x_i|\theta = \hat{\theta})$$

the log-likelihood becomes

$$L(\theta) = const \quad e^{-\frac{(\theta-\hat{\theta})^T I (\theta-\hat{\theta})}{2}}$$

and it certainly looks like a scaled gaussian with mean $\hat{\theta}$ and variance I^{-1}. So we have proved that near the maximum at θ the likelihood looks like a gaussian. The result can be extended and prove that for sufficiently large number of data N, the likelihood looks like a gaussian everywhere. This is where the various preconditions of the formal definition of the theorem come into play to ensure that for sufficiently large number of data N, there is

a single maximum, that the maximum grows sharper as N increases, etc.

Let us see now what happens when N grows big. From the law of the large numbers we know that if we average a large number of things we get the expected value, something obvious to everybody except the brightest mathematicians, who had to prove it. So

$$\mathbf{I} = \lim_{N\to\infty} I = \lim_{N\to\infty} I = NE\left[-\frac{\partial^2}{\partial\theta\partial\theta^T}\ln p(x|\theta=\hat{\theta})\right]$$

or, using the latest fashion in notation

$$\mathbf{I} = N\ E\left[-\nabla\nabla^T \ln L(\theta|x)\right]$$

and $\mathbf{I}$ is the famous *Fisher Information Matrix*. If we consider the likelihood of all data, not just a single vector x then we write as

$$\mathbf{I} = E\left[-\nabla\nabla^T \ln L(\theta|\{x_1\cdots x_N\})\right] = E\left[-\nabla\nabla^T \ln L(\theta|\mathbf{x})\right]$$

Up to this point we showed that at the limit the likelihood is a scaled gaussian in θ. We can go on and show that our estimate $\hat{\theta}$ which is a random variable since it is a function of our (randomly corrupted) data, follows a normal distribution. It is easy to do it in the Bayesianist fashion, all we have to do is a Bayesian leap of faith. It is a bit harder to do it in the Frequentist fashion and their muscle snapping rigor, but every bit as enjoyable. Knowing better, we stay out of the crossfire of these two most lustrous warring tribes and prove nothing.

So we have kind of showed that when we have lots of data the variance of our estimate is the inverse of the Fisher Information Matrix. Not only this, it follows a gaussian distribution too. But before we go on to prove a couple more theoretical results regarding the same subject we want to study a few practical aspects of this matrix. The first thing to show is that this matrix has two equivalent definitions. Let us find the second form by starting with the first

$$\mathbf{I} = E\left[-\nabla\nabla^T \ln L(\theta|\mathbf{x})\right] = -\int\left(\nabla\nabla^T \ln L(\theta|\mathbf{x})\right)p(\mathbf{x}|\theta)\mathrm{d}\mathbf{x} =$$

$$-\int\left(\nabla\nabla^T \ln L(\theta|\mathbf{x})\right)L(\theta|\mathbf{x})\mathrm{d}\mathbf{x} = -\int\nabla\left(\frac{\nabla L(\theta)}{L(\theta)}\right)L(\theta)\mathrm{d}\mathbf{x} =$$

after dropping the dependence of the likelihood on the vector $\mathbf{x}$ to save keystrokes and continue

$$\mathbf{I} = -\int\left(\frac{\nabla\nabla^T L(\theta)}{L(\theta)}\ L(\theta)\mathrm{d}\mathbf{x}\right) + \int\frac{(\nabla L(\theta))(\nabla L(\theta))^T}{L^2(\theta)}\ L(\theta)\mathrm{d}\mathbf{x}. \tag{10.5}$$

We notice that the first term is the integral

$$\left(\nabla\nabla^T L(\theta)\right)\mathrm{d}\mathbf{x} = \nabla\nabla^T\left(\quad L(\theta)\mathrm{d}\mathbf{x}\right) = \nabla\nabla^T 1 = 0$$

and so the Fisher information can be written as

$$\mathbf{I} = \quad \left(\frac{\nabla L(\theta)}{L(\theta)}\right)\left(\frac{\nabla L(\theta)}{L(\theta)}\right)^T L(\theta)\mathrm{d}\mathbf{x} = E\left[S(\theta)S^T(\theta)\right.$$

where $S(\theta)$ is the gradient of the log-likelihood and is called the *score*. Impressive! Very impressive. But does it have any use?

Computing the Fisher information, one realizes a common numerical problem. While the Fisher information is the second derivative at the maximum and it is guaranteed to be positive definite in theory, it is not so always in practice for various reasons. We might have not maximized the likelihood yet, either because we do not want to waste the computational resources if we have found a good enough estimate or because we use an algorithm that requires the Fisher information to perform the maximization. We might even have maximized it, but plain old round off error prevents us from obtaining a positive definite matrix. Now look at Eq. (10.5). No matter the value of $S(\theta)$, this matrix is definitely non-negative definite and very probably positive definite.

There are two more reasons why this alternate form is useful. One is that half the related theorems have the one form as starting point and the other half the other form. And second, very often there is no better way to verify an analytical derivation of the Fisher information than do it using a different formula.

Since we touched the topic of numerical computation, we might talk a bit more about it. So let's focus on the numerical maximization of the log-likelihood. The likelihood, being no different than any other function is maximized by finding the parameters $\hat{\theta}$ that make the score (derivative of log-likelihood) zero

$$S(\hat{\theta}) = 0$$

which very often is a non linear equation and practically all non linear equations are solved with an iterative procedure. We start with a good guess θ_0 and then find successively a θ_1, θ_2, etc until we (hopefully) converge to $\hat{\theta}$. The most generic way to get these successive approximations is the *Newton-Raphson* method where we approximate the score with first order Taylor series

$$S(\theta) \quad (\theta_j) + \nabla S(\theta_j)(\theta - \theta_j) = 0$$

which leads to the well known update rule

$$\theta_{j+1} = \theta_j - \left(\nabla S(\theta_j)\right)^{-1} S(\theta_j) \tag{10.6}$$

that works really well if we are really close to the solution. If we are not so close to the solution bad things can happen like the $-\nabla S(\theta_j)$, known as the *Observed Information* not

being positive definite. Such a lack of definiteness spells double-trouble. Not only is it hard to invert such matrices, the solution we get can wonder towards saddle points or minima. No need to despair. We can remember that the same problem appeared above when we computed the Fisher information. Given the uncanny similarity between the names and definitions of the Fisher and observed information we follow a similar derivation and get

$$\nabla S(\theta_j) = \nabla\nabla^T \ln p(\mathbf{x}|\theta_j) =$$
$$\sum_{i=1}^{N} \nabla\nabla^T \ln p(x_i|\theta_j) =$$
$$\left(\sum_{i=1}^{N} \frac{\nabla\nabla^T p(x_i|\theta_j)}{p(x_i|\theta_j)} \right) - \left(\sum_{i=1}^{N} (\nabla \ln p(x_i|\theta_j))(\nabla \ln p(x_i|\theta_j))^T \right)$$

where the first term approximates the expected value

$$\left(\sum_{i=1}^{N} \frac{\nabla\nabla^T p(x_i|\theta_j)}{p(x_i|\theta_j)} \right) \quad NE\left[\frac{\nabla\nabla^T p(x_i|\theta_j)}{p(x_i|\theta_j)} \right] =$$
$$E\left[\frac{\nabla\nabla^T p(x|\theta_j)}{p(\mathbf{x}|\theta_j)} \right] = 0.$$

This looks mostly harmless but it is not. Things that have expected value equal to zero, tend to dance around zero so that the "average" is zero. But when we compute it by averaging a finite number of samples, this quantity is positive about as many times as it is negative. Now it can just happen that it has the wrong sign and large magnitude at a most unfortunate moment (e.g. during a demo), with catastrophic results. But then if we know that it has zero mean, why on earth are we computing it? So we forget about it and compute the second term only which unsurprisingly enough corresponds to the alternative form of the Fisher information.

Replacing the computed value of this summation with its expected value, works very well in practice and it is an integral part of Levenberg--Marquardt, a well known and very effective algorithm. If we have a theory, what do we need the data. Recognizing a beautiful thing when we see it we carry this idea further.

Mark Twain is credited with one of the extremely few quotes about science that display some deeper understanding about the subject:

> There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

which brings us to the scoring method. Since what we need to invert in the vanilla Newton-Raphson is the observed information, how about if we use the Fisher (or expected) information. It turns out that this is a pretty good idea and has a fascinating property: it does not need any data! No pesky facts harassing our beautiful theories!

There is of course some exaggeration in the above statement since the process cannot be totally data free. There are two places where the data is hidden. One is inside the score $S(\theta_j)$ in Eq. (10.6), which we cannot replace with its expected value since it is identically zero, and the other in the guess θ_j itself. Since θ_j is close to the maximizing value of θ, it represents a distilation of the data (or at least approximates it).

Time for a couple of examples. First we can try to compute the variance of the estimated mean of a gaussian distribution. In a previous example we derived the log-likelihood of this as

$$\ln L(\mu) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln|C_x| - \frac{1}{2}\ln 2\pi - \frac{1}{2}(x_i-\mu)^T C_x^{-1}(x_i-\mu)\right)$$

where C_x is the covariance of the data, so that we distinguish it from the covariance of the estimate which we name C_μ. The score, that is the derivative with respect to the mean μ, we computed as

$$S(\mu) = \frac{\partial \ln L(\mu)}{\partial \mu} = \sum_{i=1}^{N}\left(C_x^{-1}(x_i-\mu)\right)$$

There are two ways to compute the Fisher information and we will try them both to see if we get what we think we should get with any of them.

The first attempt will be using the second derivative of the likelihood or, equivalently, the first derivative of the score

$$\mathbf{I} = E\left\lceil \frac{\partial S(\mu)}{\partial \mu^T} \right\rceil = E\left\lceil \sum_{i=1}^{N} C_x^{-1} \right\rceil = N\ C_x^{-1}$$

where we did not even need to take the expected value since the data had already evaporated by the time we were done with the second derivative. So the covariance of the estimate is

$$C_\mu = \mathbf{I}^{-1} = \frac{1}{N}C_x$$

which is what we expected it to be. We can try now the other version where

$$\mathbf{I} = E\left\lceil S(\mu)S(\mu)^T \right\rceil = E\left\lceil \left(\sum_{i=1}^{N} C_x^{-1}(x_i-\mu)\right)\left(\sum_{j=1}^{N} C_x^{-1}(x_j-\mu)\right)^T \right\rceil =$$

$$E\left\lceil \sum_{i=1}^{N}\sum_{j=1}^{N}(C_x^{-1}(x_i-\mu))(C_x^{-1}(x_j-\mu))^T \right\rceil = C_x^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} E\left\lceil (x_i-\mu)(x_j-\mu)^T \right\rceil C_x^{-1}$$

and since $(x_i-\mu)$ and $(x_j-\mu)$ are independent for $i \neq j$, and thus the expected value of their product is the product of their expected values which are both zero

$$\mathbf{I} = C_x^{-1} \sum_{i=1}^{N} E\left[(x_i - \mu)(x_i - \mu)^T \right] C_x^{-1} = N\, C_x^{-1} C_x C_x^{-1} = N\, C_x^{-1}$$

exactly as before. We derived the covariance in two different ways and we got what we expected both times. So we must be doing something right.

The next example will be optical flow. We will derive the Lucas and Kanade algorithm using a Maximum Likelihood approach and find the variance as well. The assumption is that the intensity $I_N[x]$ of a pixel x in frame N stays the same as the camera or the objects in the scene move and what was projected at pixel x in frame N is projected at point $x+u$ in frame $N+1$, modulo the noise $n[x]$, which in math lingo is plainly

$$I_N[x] - I_{N+1}[x+u] = n[x]$$

and to make life easier, the image is 1-D. We have a surfeit of assumptions here and we state the most important. The intensity is constant, the so called *Image Constancy Assumption* and the noise is assumed independent, zero mean, gaussian with variance equal to $+_n^2$. We also assume that u is small enough to make first order Taylor approximation reasonable. And since we are going to use Lucas and Kanade we have to assume that the flow u does not vary much in the neighborhood of x. The noise model then for a single pixel is

$$L(u) = p(n[x]\,|u) = const\; e^{-\frac{n^2[x]}{2+_n^2}} =$$

$$const\; e^{-\frac{(I_N[x]-I_{N+1}[x+u])^2}{2+_n^2}}$$

and we can multiply the noise models for all the pixels in one small patch to get the likelihood for this patch. Knowing better than that, we work with the log-likelihood

$$\ln L(u) = -\sum_{i=-3..3} \frac{(I_N[x+i] - I_{N+1}[x+i+u])^2}{2+_n^2}$$

omitting the constants that will not survive the differentiation. If we take derivatives right away we get

$$\frac{\partial}{\partial u} \ln L(u) = \sum_{i=-3..3} \frac{I_{N+1,x}[x+i+u](I_N[x+i] - I_{N+1}[x+i+u])}{+_n^2}$$

which is a non-linear equation since $I_{N+1}[x+i+u]$ and its derivative are not linear functions of u. We can use the standard Newton-Raphson approach and take the second derivative and then omit the second derivarives that appear. We opt for a more direct and much simpler approach, where we approximate the image by a linear function and then take the derivative. This results in the same mathematical expression in the end but eliminates the need for arguing about the omission of the second derivatives

$$\ln L(u) = -\sum_{i=-3..3} \frac{(I_N[x+i] - I_{N+1}[x+i] - I_{N+1,x}[x+i]u)^2}{2+_n^2}$$

from which the derivative (or score) becomes

$$\frac{\partial}{\partial u} \ln L(u) = \sum_{i=-3..3} \frac{I_{N+1,x}[x+i](I_N[x+i] - I_{N+1}[x+i] - I_{N+1,x}[x+i]u)}{+_n^2}$$

and after setting

$$E_{xx} = \sum_{i=-3..3} I_{N+1,x}^2[x+i]$$
$$E_{xt} = \sum_{i=-3..3} I_{N+1,x}[x+i](I_{N+1}[x+i] - I_N[x+i])$$

we get

$$S(u) = -\frac{E_{xt} + E_{xx}u}{+_n^2} = 0$$

that looks equivalent to the well known Lucas and Kanade equation. But we should not forget our original target, to compute the variance:

$$\mathbf{I}_u = -\frac{\partial}{\partial u} S(u) = \frac{E_{xx}}{+_n^2}$$

10.6. Minimum Variance

Variance is an important measure of the quality of an estimator but by itself is not enough. One can design an estimator that has zero variance easily: the estimator defined as $\hat{\theta} = 3$ or any other constant of your preference. You cannot beat that! But it will be a useless estimator, nevertheless. On the other hand one can define other quality measures for estimators, like the mean square error, but this is less convenient to use and in most interesting situations it does not provide us with more usefull information than the variance. So we use the variance.

But before we invest our precious time in it we have to make sure that any lower bounds, upper bounds or whatever bounds are not fooled by estimators as silly as the constant. So we first establish the concept of *bias* as the following difference

$$B(\theta) = E\{\hat{\theta}\} - \theta$$

where $\hat{\theta}$ is our estimator, i.e. a function of our data and nothing else, and θ is the ground truth of the estimated quantity. Notice that B is a function of θ only, and does not depend on the value of $\hat{\theta}$ or the data. Clearly we prefer estimators whose bias is zero, that is unbiased, if for no other reason because bias is a bad thing. Unfortunately, like many other things in life we have to settle for less. Most often it is very hard to even measure bias, let alone find an estimator that is unbiased. And almost equally often everybody is happy using an estimator that is universally proclaimed biased, so there is little point bothering. A less stringent requirement is for the estimator to be *asymptotically unbiased*, which it is

iff

$$\lim_{N\to\infty} E\{\hat{\theta}\} = \theta$$

where N is the number of data we have. This is a far less powerfull requirement and any estimator that does not satisfy it is almost useless. One very closely related concept is that of *consistency*. An estimator is *consistent* iff the estimator converges to the ground truth. We can distinguish between *strongly consistent* and *weakly consistent* depending on whether the convergence is *in probability* or *almost sure*. The understanding of the subtle difference between estimators that are *asymptotically unbiased*, *strongly consistent* or *weakly consistent* is what makes mathematicians so attractive to the opposite sex, and as such is beyond the scope of this modest text.

We have already seen that maximum likelihood can produce biased estimators. We proved it by example, since need only one biased estimator to establish that there is at least one. There are other theorems that indicate that maximum likelihood is in the habit of producing biased estimators. And in practice it does it all the time. If one wants desperately to rationalize on why, one can argue that maximum likelihood tends to produce parameter estimators that are associated with smaller variances so that the likelihood is increased just a bit more. And given that the poster child of estimators is addicted to bias, statisticians engineers and whoever else makes a living out of statistics have learned to live with biased estimators as long as the error is small.

Going back to mean square error

$$MSE(\hat{\theta}) = E\{(\hat{\theta}-\theta)^2\}$$

is closely linked to the variance because

$$MSE(\hat{\theta}) = E\{(\hat{\theta}-\theta)^2\} = E\left\lceil\left(\left(\hat{\theta}-E\{\hat{\theta}\}\right)+\left(E\{\hat{\theta}\}-\theta\right)\right)^2\right.$$

and taking into account that $E\{\hat{\theta}\}-\theta = B(\theta)$ is not a random variable and that the mean of $\hat{\theta}-E\{\hat{\theta}\}$ is zero we get

$$MSE(\hat{\theta}) = E\left\lceil\left(\hat{\theta}-E\{\hat{\theta}\}\right)^2\right. + B^2(\theta) = var\{\hat{\theta}\} + B^2(\theta)$$

and this means that if we already know the bias of an estimator the mean square error gives no additional information.

It should be obvious that one cannot meaningfully talk about the variance of any estimator, because many things fit the definition of an estimator and they are not meaningful, like the constant estimator for example. So to make the discussion meaningful we restrict ourselves to meaningful estimators. But despite the scarcity of unbiasedness, the discussion on minimum variance centers around unbiased estimators, rather than asymptotically unbiased or consistent ones, mainly because two nice fellows, Harald Cram ? and Calyampudi Radhakrishna Rao, presented us with the *Cramer-Rao Lower Bound*,

that in its simplest form deals only with unbiazed estimators.

What is quite a pleasant surprise is that the Cramer Rao Lower Bound applies to a wide variety of estimators and is by no means tied to the the maximum likelihood estimator, despite that many of the concepts we talked about in the context of maximum likelihood are employed in the proof and derivation of this bound, most notably the Fisher information matrix. Not only that, the simpler version of the bound rarely applies to the shamelessly biased maximum likelihood.

10.7. Cramer Rao Lower Bound

Theorem. The covariance of an estimator $\hat{\theta}$ whose expected value is μ_θ and its Fisher information matrix is $\mathbf{I}(\theta)$ satifies the following inequality

$$Var\{\hat{\theta}\} - (\nabla\mu_\theta(\theta))^T \mathbf{I}^{-1}(\theta)\nabla\mu_\theta(\theta) \quad 0$$

where $\quad 0$ means simply positive definite

Obviously if the estimator is unbiased then $\mu_{\hat{\theta}} = \theta$ and $\nabla\mu_{\hat{\theta}} = \mathbf{1}$ the identity matrix and the inequality gets its more popular form

$$Var\{\hat{\theta}\} - \mathbf{I}^{-1}(\theta) \quad 0$$

Before we prove the theorem we consider two lemmas that will come really handy.

Lemma 1. The expected value of $S(\theta)$ is

$$E\{S(\theta)\} = 0$$

Proof: Just apply the rules

$$E\{S(\theta)\} = \quad S(\theta)p(x|\theta)\mathrm{d}x = \quad \nabla \ln L(\theta)L(\theta)\mathrm{d}x = \quad \frac{\nabla L(\theta)}{L(\theta)} L(\theta)\mathrm{d}x =$$

$$\nabla L(\theta)\mathrm{d}x = \nabla \quad L(\theta)\mathrm{d}x = \nabla 1 = 0$$

QED. I wish all proofs were like this.

Lemma 2. If the following compound matrix

$$M = \begin{bmatrix} A & . & C \\ \dots & \dots & \dots \\ C^T & . & B \end{bmatrix}$$

is positive definite, then

$$K = A - CB^{-1}C^T$$

is also positive definite.

Proof: The proof for this is pure mechanics. The interested reader can easily verify that the inverse of matrix M is

$$M^{-1} = \begin{bmatrix} (A - CB^{-1}C^T)^{-1} & . & -(A - CB^{-1}C^T)^{-1}CB^{-1} \\ \cdots & \cdots & \cdots \\ -B^{-1}C^T(A - CB^{-1}C^T)^{-1} & . & (B - C^T A^{-1}C)^{-1} \end{bmatrix}$$

and since the inverse of a positive definite matrix is also positive definite then M^{-1} is positive definite. This implies that $K^{-1} = (A - CB^{-1}C^T)^{-1}$ is positive and consequently K is too. QED.

Proof of Theorem: It is easy to prove the theorem now. We form the compound random vector

$$V = \begin{bmatrix} S(\theta) \\ \hat{\theta} \end{bmatrix}$$

and compute its covariance

$$Cov\{V\} = E\{(V - E\{V\})(V - E\{V\})^T\} =$$
$$\begin{bmatrix} E\{(S(\theta) - E\{S(\theta)\})(S(\theta) - E\{S(\theta)\})^T\} & E\{(S(\theta) - E\{S(\theta)\})(\hat{\theta} - \mu_{\hat{\theta}})^T\} \\ E\{(\hat{\theta} - \mu_{\hat{\theta}})(S(\theta) - E\{S(\theta)\})^T\} & E\{(\hat{\theta} - \mu_{\hat{\theta}})(\hat{\theta} - \mu_{\hat{\theta}})^T\} \end{bmatrix}$$

and since $E\{S(\theta)\} = 0$

$$Cov\{V\} = \begin{bmatrix} E\{S(\theta)S(\theta)^T\} & E\{S(\theta)\hat{\theta}^T\} \\ E\{\hat{\theta}S(\theta)^T\} & Var\{\hat{\theta}\} \end{bmatrix} = \begin{bmatrix} \mathbf{I}(\theta) & E\{S(\theta)\hat{\theta}^T\} \\ E\{\hat{\theta}S(\theta)^T\} & Var\{\hat{\theta}\} \end{bmatrix}$$

and we now only need to compute the expected value of the product of $S(\theta)$ and $\hat{\theta}$

$$E\{S(\theta)\hat{\theta}^T\}$$

which can be manipulated to death as follows

$$E\{S(\theta)\hat{\theta}^T\} = \int (\nabla \ln L(\theta))\hat{\theta}^T p(x|\theta)\mathrm{d}x = \int (\nabla \ln p(x|\theta))\hat{\theta}^T p(x|\theta)\mathrm{d}x =$$
$$\int \frac{\nabla p(x|\theta)}{p(x|\theta)} \hat{\theta}^T p(x|\theta)\mathrm{d}x =$$
$$\nabla \int p(x|\theta)\hat{\theta}^T \mathrm{d}x = \nabla \int p(x|\theta)\hat{\theta}^T \mathrm{d}x = \nabla \mu_{\hat{\theta}}$$

from where a simple application of the second lemma brings us to the desired result. *QED*.

This is one of the most celebrated theorems in statistics and one revered by both applied and theoretical statisticians. The former because they hope to apply it some day somewhere, the latter because they know it cannot be applied. Nevertheless, it gives us a nice warm feeling and some reassurance when we use the inverse of the information matrix as an estimate of the covariance matrix.

11. Hypothesis Testing

It loves me; It loves me not. This is the question that is always in the back of the mind of every cat owner. You might be surprised to learn that similar dilemmas plague computer vision. Sometimes the question is asked explicitly and directly, some times it is just implied. Does this pixel belong to this nicely moving segment, or not? Is this pixel part of this independently moving object or not? Are we still tracking the right thing? Is this pixel visible in the next frame? Is this the picture of Aunt Rhodie? Just about any question that can be answered with a "Yes" or "No" will be our concern in the section.

Statisticians have invented a very strict methodology to answer this type of questions in a very rigorous manner called *Hypothesis Testing*. Using this methodology the answer would be a "Yes" or "No" accompanied by a margin of error or level of significance. Which is really good news. The bad news is that before we get an answer we have to provide a statistical model for the problem. Building models can be hard work and can involve tricky and hard to analyze approximations or assumptions, but this is the price to pay.

And this price can be steeper for problems that are more levels of abstraction away from the fundamental physics of the problem. As a result these statistical methodologies can be applied mainly to low level vision problems where mathematical models are more directly applicable and have been traditionally more cooperative. There are exceptions of course like face recognition where a seemingly high level vision problem is solved respectably with low level vision techniques like PCA and jets.

11.1. The Mechanics of Hypothesis Testing

But before we start thinking about integrals and convolutions we should understand what hypothesis testing is about. First of all, the yes-no dilemma is not symmetric, we do not have the same criteria for the yes and the no sides. Pretty much like a judge in a court that does not pronounce the defendant guilty or not guilty with symmetric arguments, but requires a very high level of confidence for the guilty verdict, we have the *null hypothesis* and the *alternative hypothesis*, which are treated differently.

The *null hypothesis*, H_o for short, is called "null" to uphold the stereotype of statistics as a difficult subject by confusing the newcomers. But, due to a fortunate congruence of events, it also provides a common name for this type of hypotheses, given that in different contexts these hypotheses should have different names: default, status quo, safe, not guilty, disprovable etc. And as this long parade of names indicates, the null hypothesis is the status quo, the hypothesis that we do not need or we can not prove. It is the one that has to be disproved before we accept the alternative hypothesis and has to be disproved beyond reasonable doubt. Even if the alternative hypothesis looks better, we have to rule out chance since we can not upset status quo without extremely high confidence level.

The alternative hypothesis H_1, is the guilty, favorite or novel hypothesis. We do not accept it lightheartedly and we accept it only if the H_o is truly condemned by the data, if the plausibility of H_o has slipped below some level and this level is normally very favorable towards H_o. In most situations where the hypothesis testing is applied the

experimenter wants the alternative hypothesis H_1 to be accepted since this is a new theory and most likely the experimenter's own. Given the well documented weaknesses of the human nature, it is wise to set the confidence bar really high. While in most low level vision problems, there is no favourite hypothesis, and there is no human against whose instincts we are fighting, the H_o is just the one we *can* disprove and the reason that we set the bar very high against H_1 is to guard against approximate models.

The most common scenario for the application of hypothesis testing is for the evaluation of new medicines. Suppose that for a certain disease the standard medication occasionally causes death from bleeding and a new medicine appears that claims to treat the disease without messy deaths. To test it the developers of the medication treat 100 patients with the new medicine, the target group, and 100 patients with the old medication, the control group. The null hypothesis is that the old medicine works at least as well. The alternative hypothesis is that the new medicine works better. They run the experiment for 10 years and in the end they count 3 deaths from bleeding in the control group, as expected, but no deaths from bleeding in the target group that was taking the new medicine. Can we infer that the new medicine is working? Can this happen by chance? It turns out that these results could be obtained even if the new medicine was identical in behavior to the old one with 5.9% chance or to put it in more concrete terms, if we repeated the same experiment 100 times we would get results that favorable about 6 times. So, is the 5.9% chance too high? It is matter of ethics, economics and history (or self-righteousness, greed and sloth). The answer to this particular question is of no importance here, but we will encounter similar questions in vision, and the choice of the cutoff there depends more on plain old practicality and less on cardinal virtues like sloth.

Enough with philosophy, let's do some work. Assume that we have two pixels from two different images and we want to see if they are projections of the same 3-D point. We know from instinct that we cannot give a definite answer with just one pixel but we try our best. We have to follow a few clear steps that are common to practically all such problems.

(1) Specify the H_0 and H_1, based on the nature of the problem and the statistical models available.

(2) Define a statistic[§] on the random variables $Q(I_1, I_2)$, where I_1 and I_2 are the intensities of the two pixels. The statistic should be easy to compute numerically and satisfy our instincts about the problem.

(3) Derive $p(Q(I_1, I_2)|H_0)$ the probability density of Q given the null hypothesis. It helps if you have a good model and a statistic for which you can compute probability distributions.

(4) Decide on a significance level p_t and define a range r (or a set of ranges) for Q such that Q is outside this range with probability equal to the significance level p_t. The range r is called *confidence interval*, e.g. if Q is within this range we have

[§] A statistic is just a function of the random variables. The average of the random variables and the sum of squared differences of pairs of random variables are two common statistics.

confidence in H_o. The significance level is typically 5%, 1%, 0.1% or thereabouts. If Q is outside r then the evidence against H_o is significant.

After we set up all this statistical infrastructure we flesh out the actual algorithm. The algorithm is astonishingly simple.

(1) Compute the statistic $Q(I_1, I_2)$.

(2) Determine if Q is within the range r.

(3) If it is within the range accept the null hypothesis. If not, accept the alternative hypothesis.

Other than the first step, the computation of Q, the rest of the steps are computationally trivial. A really simple procedure.

Now that we know what to do, we have to do it. For the first step we have exactly two choices. Either

- H_0 is that the pixels are the same and H_1 is that the pixels are different.

or

- H_0 is that the pixels are different and H_1 is that the pixels are the same.

We might be tempted to define H_0 as "pixels are different" but a couple of steps down the road we have to derive the probability of Q given the null hypothesis which is far from easy in this case since there is little help in the literature and it is very hard without oversimplifying assumptions. So we opt for the other definition that H_0 is "pixels are the same". But we have to elaborate a bit on that and define "pixels are the same" in a way even a mathematician can understand. We know from intuition and our experience as computer vision practitioners that if the two pixels are projections of the same thing, their intensities (or colors) will be the same modulo some noise. We develop a model for this noise and then the null hypothesis is defined as "the difference in the intensities of the two pixels is just noise that follows our noise model".

The second step is to define a statistic, which we want to be easy to compute and easy to find its probability density. For our purposes, this has to do with some kind of weighted sum of squared differences. Here we deal with one of the simplest forms of the problem where we have just one squared difference. Nevertheless we weight this squared difference just to keep everything on the same footing. So

$$Q(I_1, I_2) = \frac{(I_1 - I_2)^2}{+_n^2}$$

where $+_n^2$ is the variance of the noise.

The next step is to get the probability distribution of the statistic Q, so we examine the above formula carefully. Given that the two pixels are projections of the same 3-D point, the quantity

$$\frac{-I}{+_n} = \frac{I_1 - I_2}{+_n}$$

is the noise, so it has zero mean. It is also scaled by $+_n$ so it has unit variance and as a result the mean of Q

$$E\{Q\} = E\left\lceil\left(\frac{-I}{+_n}\right)^2\right. = \frac{E\left\lceil(-I)^2\right.}{+_n^2} = \frac{+_n^2}{+_n^2} = 1.$$

If the noise is Gaussian then we also know that Q follows the ^{2} (a.k.a. as *Chi Square*) distribution with one degree of freedom that has probability density $f_{\ 2}(Q;1)$, where the second argument indicates the degrees of freedom. We select a confidence level, say 0.5%, open the probability tables and find the confidence interval. If the noise is not Gaussian we have to decrease the confidence level to say 0.05% and run a few extra tests. So with confidence level 0.5% the tables give us 7.9. This means that if the weighted squared difference is greater than 7.9 we reject the null hypothesis and infer that the pixels are different. If it is less we concede the null hypothesis and infer that the pixels are equal.

There are a few observations to be made on this really simple test.

(1) We need the variance of the noise. It is not that hard to estimate especially off-line and can be done. The standard tool is *Maximum Likelihood*, but other tools like the *Method of the Moments*[§] can be used as well but the choice of good methods is rather limited. More complex situations require of course heavier investment for model building.

(2) We do not need a model for the alternative hypothesis, although it is really the hypothesis on which we are working.

(3) The justification behind the selection of the confidence interval is based largely on the intuition regarding the alternative hypothesis. Had the alternative hypothesis been that we have overfitted the model (i.e. fitted a model with too many parameters on too few independent data) the interval would have been something like $[0 \cdots 0.02]$.

(4) By examining just one pixel we cannot get the discrimination power we need for most applications. We can use more pixels but then we need a more sophisticated model and it is a bit harder to get a good set of parameters for this model.

(5) This statistic looks very much like least squares but then most of the statistics used in vision are similar to a weighted sum of squared differences. The design of statistics for our kind of problems, is the port of entry for ingenuity and the chance to shoot your foot or your palate.

Hypothesis testing is an excellent tool for deciding for or against a hypothesis, a tool that requires minimum investment in assumptions, models etc.

[§] It is the Method of the Moments, not the Method of the Moment, so it has nothing to do with the soup of the day.

11.2. Bayesian Ratios

Hypothesis Testing though is not the only game in town. There are other tools, most notable being the ones that are related to the Bayesian formula. Let's explore then one of the alternatives for doing more or less the same thing.

We have two distinct hypotheses, namely either H_o that the pixels are projections of the same point in space, or H_1 that the pixels are projections of the different points in space. We will use the same statistic and try to evaluate the relative merit of the two hypotheses. The discerning reader must have already noticed a difference. We do not discriminate against H_1.

In the H_o hypothesis things are as before and we know the probability density of Q is $p(Q|H_o) = f_{\ 2}(Q;1)$. In the other hypothesis things are different. In hypothesis testing we did not need a model or an explicit probability for H_1, but we do now. It is just a scaled squared difference, so it cannot be hard. Well, it is not hard if it is assumed Gaussian. We can find fairly easily the variance $+_d^2$, but since we still use the statistic Q where we divide by $+_n^2$ we need to do a couple of simple changes. If we divided by $+_d^2$ then $p(Q|H_1)$ would be a ^{2} distribution. Instead $p(Q|H_1)$ is a stretched version of the ^{2} namely

$$p(Q|H_1) = \frac{+_n^2}{+_d^2} f_{\ 2}(\frac{+_n^2}{+_d^2} Q, 1)$$

where the 1 as the second argument indicates one degree of freedom as before. So we know $P(Q|H_1)$ as well.

Now we drop the question. What is the probability of H_0 or H_1 given the value of the statistic. It seems that we need the help of Rev. Thomas Bayes, the 18^{th} century British mathematician and Presbyterian minister, who conceived the first version of the theorem that bears his name. The paper that started all this was published after his death by Richard Price, a friend of his, under the title *Essay Towards Solving a Problem in the Doctrine of Chances*. The theorem (applied on the problem at hand) says that

$$P(H_0|Q) = \frac{p(Q|H_0)P(H_0)}{p(Q|H_0)P(H_0) + p(Q|H_1)P(H_1)}$$

and similarly for

$$P(H_1|Q) = \frac{p(Q|H_1)P(H_1)}{p(Q|H_0)P(H_0) + p(Q|H_1)P(H_1)}$$

where the upper case P is the probability of an event like H_0 and lower case p is the probability density of a random variable. It is easy now to see that whenever the Bayesian ratio

$$\frac{p(Q|H_0)P(H_0)}{p(Q|H_1)P(H_1)}$$

is greater than 1 we should choose H_o, otherwise choose H_1. But we need to know the two numbers $P(H_0)$ and $P(H_1)$ before we compute the Bayesian ratio. We can examine three cases.

(1) We can select a neutral prior. Set $P(H_0) = P(H_1) = 0.5$. This just informs our mathematical machinery that we know nothing about the hypothesis and bet it is fifty fifty. Quite prudent but rather unrewarding.

(2) If our application is motion segmentation with static background, where most of the scene is background that stays the same then $P(H_0) = .95$ and $P(H_1) = 0.05$.

(3) If our application is point correspondence and we search the whole second image, then most of the images are different so $P(H_0) = 1.0e-6$ and $P(H_1) = 1-1.0e-6$.

One thing that is sure with this bayesian approach is that we have a flexibility rivaling that of the elastigirl. But this flexibility comes at a price. We need a model for the alternative hypothesis. And we also need the priors $P(H_0)$ and $P(H_1)$ which can be both good and bad. Good if we know something from another source, bad if we have to guess.

11.3. Hypothesis Testing vs Bayesian Ratios

Apart from these practical differences, there is a deep philosophical difference between Hypothesis Testing and Bayesian Ratios. When we compute the probability of H_0, essentially we compute the probability of an unknown but fixed binary constant, since the two pixels are either the same or different. This unknown binary constant is not a random variable, because we cannot conceive a realistic experiment with different outcomes for this fixed unknown. So in applying the Bayes theorem the way we did here, we assigned probability to a constant and there is no elegant and intuitive way to do it without excessive handwaving. The natural way of defining probability, what today is called *frequentist* definition, is that probability is the relative frequency of an event and only truly stochastic (random) events have such a probability associated with them. So the definition has been expanded to mean the degree of plausibility of the truth of a statement so that even fixed (non-random) unknowns can have probabilities associated with them.

Statistical conservatives of the 18^{th} century might not have noticed the licentious expansion of the definition of probability, but in the early 19^{th} century, some guy named Laplace treated the mass of Saturn as a random variable and computed it, in effect spilling the beans. "That's it" the conservatives said. "Saturn is not a TV celebrity on a yo-yo diet! Its mass is a constant". The controversy at this point seemed to be over and the conservatives, now known as frequentists, the winners. The Bayes theorem was forgotten, to be rediscovered again in the 20^{th} century. But the situation started changing in the fifties. Pattern Recognition became an established field and it was using a lot of things bayesian, the mass of Saturn turned out to be where Laplace said it probably was, mathematicians removed some rough spots in probability theory and yo-yo diets became popular among TV personalities.

Meanwhile the frequentists had carved an ecological niche of their own and had developed a very elaborate theory of probability. So the end result is that today we have

two camps, splitting the community of statisticians right in the middle. But on the positive side we have two classes of techniques to use. The frequentist sponsored ones like classic hypothesis testing, require minimum assumptions and leave no lingering doubts when our problem fits their strict requirements. And the omnipotent Bayesian techniques that give us immense freedom but require a heavier investment in assumptions and models. Choices, choices, choices.

Last but not least, the two methods differ in that hypothesis testing needs a model for the null hypothesis only, whereas bayesian ratios need a model for both the null and the alternative. This might seem like twice the work, but in many cases it is much more than that. The underlying statistic is chosen so that we can easily apply the model and compute the required thresholds. If we have one model only, then the job of selecting a good statistic is easy. If we have two models then there might not exist a single statistic on which we can apply the models easily. This can be a show stopper or the point of entry for approximations and simplifications. Nothing is perfect in this life.

11.4. Chi Square with more Degrees of Freedom

We return now to the hypothesis testing approach and examine what happens if we use more pixels in testing our hypothesis. In the example of the previous sections we used the squared difference between a pair of pixels as our statistic but this left us with a slightly unsalted taste in the mouth. There is only so much one can do with just a single pixel difference. So we use the *sum of squared differences* as our statistic. If we assume that the difference between two images of the same object is just the noise and that this noise is i.i.d. (independent, identically distributed also called white noise), zero mean Gaussian, then

$$Q(I_1, I_2) = \sum_i \left(\frac{I_1[i] - I_2[i]}{+_n} \right)^2 \tag{11.1}$$

where the summation is understood to cover a small patch R. This follows the ^{2} distribution as before but with N degrees of freedom, where N in our case is the number of pixels in the summation. We then decide on the level of significance, and look up the tables to decide on the interval of confidence, e.g. the threshold for Q above which we reject the null hypothesis.

This all sounds fine but there are two details. First detail is that we need to know the variance $+_n$, with some accuracy. As in the single pixel case above, if we underestimate the value of $+_n$ by a wide margin, then we will have too many false negatives and if we overestimate, by a wide margin too many false positives. The good news is that if the underlying model is appropriate, then we can have good results even with approximate $+_n$, at least better than dealing with a single pixel.

The second detail is the bad news. The noise is hardly ever i.i.d. in real life and there are many reasons for this other than Murphy's Law. The two patches are not always perfectly aligned, so their difference is not zero even without any other noise. But what makes the situation complex is that the difference is much higher around edges because the misalignment can make us subtract pixels from the opposite sides of the edge. The

probability distribution of the noise then depends on the existence of an edge in the immediate neighborhood, so the noise is not identically distributed. Furthermore, if there is misalignment affecting the whole patch, then all pixels in the patch will be affected by it and one cannot claim that the image differences are independent. The same thing happens if there is change in illumination. If a pixel becomes 1% brighter, then the noise is about 1 percent of the intensity, in other words proportional to the intensity, and thus again not identically distributed. And if a pixel is affected by a change of illumination, so does every pixel in its neighborhood, and the noise is again not independent.

Luckily the solution exists. We just have to have a model, then open a book and see what tools the mathematicians have invented before us for us. The method of choice is the *Mahalanobis Distance* which was invented by Prasanta Chandra Mahalanobis [1893-1972], one of the greatest statisticians who was immortalized by the above mentioned distance. This distance follows the ^{2} distribution (under Gaussian assumption) and it is suitable for correlated data.

To use it we first have to put our data into a vector so all the pixels $I_1[i]$ and $I_2[i]$ in the patch R are arranged on one dimensional vectors $\vec{I}_1$ and $\vec{I}_2$. The difference of the two vectors is

$$-\vec{I} = \vec{I}_2 - \vec{I}_1$$

and it is easy to see that Eq. (11.1) can be written in a more compact form

$$\sum_i \left(\frac{I_1[i] - I_2[i]}{+_n} \right)^2 = \frac{-\vec{I}^T - \vec{I}}{+_n{}^2} \tag{11.2}$$

which gives us hope that we will not have extremely large equations. The Mahalanobis Distance between image patches $\vec{I}_1$ and $\vec{I}_2$ is defined as

$$D_m^2 = -\vec{I}^T C_{-I}^{-1} - \vec{I}$$

where C_{-I} is the variance covariance matrix of the difference $-\vec{I}$. The Mahalanobis distance is a generalization of Eq. (11.2) since for $C_{-I} = +_n^2 \mathbf{1}$ the two of them are identical. We will now show that the Mahalanobis distance has the statistical behavior of the sum of the squares of a set of uncorrelated zero mean unit variance random variables. To do this we define

$$S = \sqrt{C_{-I}}$$

or to be less inexact

$$C_{-I} = SS^T.$$

Matrix S is not uniquely defined, but this is hardly an impediment for a determined statistician. We just compute the most convenient matrix S, which is in practice done with the *Choleski* decomposition. Then the vector

$$V = S^{-1} - \vec{I}$$

has mean equal to the zero vector

$$\mu_V = E\{V\} = E\left\lceil S^{-1}\vec{-I} \right\rceil = S^{-1}E\left\lceil \vec{-I} \right\rceil = \vec{0}$$

because the difference of the two patches has zero mean. And variance equal to the identity matrix

$$C_V = E\left\lceil VV^T \right\rceil = E\left\lceil S^{-1}\vec{-I}\vec{-I}^T S^{-T} \right\rceil =$$

$$S^{-1}E\left\lceil \vec{-I}\vec{-I}^T \right\rceil S^{-T} = S^{-1}C_{-I}S^{-T} =$$

$$S^{-1}SS^T S^{-T} = \mathbf{1}$$

in other words the elements of the vector V are uncorrelated with zero mean and unit variance. Then the Mahalanobis distance

$$D_m = V^T V = \sum_i V[i]^2$$

is equivalent to a sum of squares of zero mean unit variance uncorrelated random variables. Under Gaussian assumption D_m^2 then follows the ^{2} distribution with N degrees of freedom.

Are we there yet? Not quite. We have two more problems to solve. The first is estimate the matrix C_{-I} and second compute its inverse. To estimate the covariance matrix C_{-I} we decide on a parametric form for this matrix and then we estimate the parameters. The second problem, the inversion of the covariance matrix is mainly a matter of computational efficiency. If the patch R is a modest 5×5 square region, then the covariance matrix is 25×25 and it requires about 15,000 operations to be inverted. And a slightly less modest 11×11 region would require about 2 million operations. Applying the Mahalanobis distance on many regions would be prohibitive unless we speed up the operations, which we will show how it can be done.

So let's assume that the difference between the two images is just noise but this noise has some structure. While in practice there are several significant components in this noise, we consider only two and leave the rest as an exercise. One component is just an i.i.d noise, like the one that is due to the random arrival of photons at each pixel of the CCD, or due to the electron shot noise. Since we are dealing with a whole patch $\vec{I}_1$ which we still handle as a vector, we use a vector of independent random variables $n_{\vec{s}}$ with zero mean and variance $+_s^2$ to represent this noise. We also have random fluctuations of illumination which is a single scalar random variable with zero mean and variance $+_l^2$. Since these fluctuations increase or decrease the intensity proportionally then the noise is $I_{\vec{1}} n_l$ so the total noise is

$$-\vec{I} = n_{\vec{s}} + \vec{I}_1 n_l. \tag{11.3}$$

Since all the components of the noise have zero mean, $-\vec{I}$ has zero mean as well. The covariance matrix C_{-I} is

$$\begin{aligned} C_{-I} &= E\left[-\vec{I}-\vec{I}^T\right] = E\left[\vec{n}_s\vec{n}_s{}^T\right] + \vec{I}_1 E\left[n_l^2\right]\vec{I}_1^T \\ &+_s^2 \mathbf{1} +_l^2 \vec{I}_1\vec{I}_1^T. \end{aligned} \tag{11.4}$$

The next question is how to compute the parameters $+_s^2$ and $+_l^2$. We either compute them from first principles, like find how much shot noise we have in a particular model of the camera, under the current conditions (bliah!) or estimate them. If you choose to estimate them then there are not that many choices, either Maximum Likelihood, perhaps with a bit of Bayesian flavor, if we have a good guess what these parameters should look like, or Method of the Moments.

One might notice that it would be more accurate to write the second term of Eq. (11.3) as $(I_{\vec{1}} + n_{\vec{s}})n_l$ instead, to include in other words the cross-talk between the two noises. It would be more correct, indeed. This would change the first term of Eq. (11.4) to $(+_s^2 + +_s^2 +_l^2)\mathbf{1}$. Since the noise is usually small, the product of the two variances will be an even smaller number. But even then if the noise is not small, we can replace $(+_s^2 + +_s^2 +_l^2)$ with a new variable $+_{sl}^2$, and the form of the equation is preserved, so unless we attempt to compute these parameters from first principles, the inclusion of the of the cross-talk, does not change the process.