CHAPTER 3

Introduction to Statistical Estimation

1. The Simplest Kind of Statistics

Least Squares is the simplest and most intuitive kind of statistics and often the most useful. The most straightforward application is as follows. We have a set of quantities that we would like to be zero or as close to zero as possible and they all depend on a set of unknowns. We take the squares of all these quantities, sum them up and then minimize this sum with respect to the unknowns. There are many alternatives to least squares that sometimes have interesting properties (most notably robustness to outliers) but least squares is not only the simplest but is also the basis for most of the alternatives.

2. Point in the Middle

Consider the following very simple problem. We want to find a point P and all we have is a set of several approximations of P which we call P_i, $i = 1..N$. If of course all the P_is are identical the choice is easy. Otherwise we would like P to be as close to all of them as possible. We form the sum of the squared differences

$$Q(P) = \sum_{i=1}^{N} (P_i - P)^2.$$

The standard way to minimize Q is to take its derivatives with respect to the unknowns and equate them to zero. Solving these equations will give us P, the vector of the unknowns. In this very simple problem solving the equations is easy, but taking the derivatives is slightly more complex. We examine two ways to take these derivatives. One is scalar (element by element) derivatives and the other is vector derivatives.

2.1. Scalar Derivatives

Our unknowns are the elements of the vector P

$$P = \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_K \end{bmatrix}$$

and our data are the vectors P_i

$$P_i = \begin{bmatrix} p_{i1} \\ p_{i2} \\ \cdots \\ p_{iK} \end{bmatrix}$$

So

$$Q(P) = \sum_{i=1}^{N} \sum_{j=1}^{K} (p_{ij} - p_j)^2$$

and

$$\frac{\partial Q(P)}{\partial p_k} = \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{\partial}{\partial p_k} (p_{ij} - p_j)^2 =$$

$$-2 \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{\partial p_j}{\partial p_k} (p_{ij} - p_j) = -2 \sum_{i=1}^{N} \sum_{j=1}^{K} \delta_{jk} (p_{ij} - p_j)$$

where δ_{ij} is the Kronecker delta, e.g. $\delta_{ij} = 0$ iff $i \neq j$ and $\delta_{ii} = 1$ which of course makes perfect sense: the derivative of an unknown with respect to itself is equal to one and the derivative of an unknown with respect to a different unknown is zero. The delta affords us some simplifications, so

$$\frac{\partial Q(P)}{\partial p_k} = -2 \sum_{i=1}^{N} (p_{ik} - p_k)$$

which if we equate to zero we get

$$p_k = \frac{\sum_{i=1}^{N} p_{ik}}{N}$$

e.g. every element of the unknown vector is the average of the corresponding elements of the data.

2.2. Vector Derivatives

A more compact and mainly more elegant way of doing the same thing is taking vector derivatives. Most of the rules of scalar derivatives apply, some with a small quirk. Let's start.

The notation

$$\frac{\partial Q(P)}{\partial P}$$

indicates a vector whose elements are the scalar derivatives of $Q(P)$ with respect to the corresponding element of P (remember Q is a scalar). Sometimes the "grad" notation is used to indicate the same thing

$$\frac{\partial Q(P)}{\partial P} = \nabla_P Q(P)$$

where the subscript P is the vector with respect to which the derivatives are taken. If it is obvious what this vector is (in many physics problems it is always the position vector) it is omitted. So

$$\frac{\partial Q(P)}{\partial P} = \frac{\partial}{\partial P}\sum_{i=1}^{N}(P_i - P)^2 = \sum_{i=1}^{N}\frac{\partial}{\partial P}\left((P_i - P)^T(P_i - P)\right) =$$
$$2\sum_{i=1}^{N}\left(\frac{\partial}{\partial P}(P_i - P)^T\right)(P_i - P) = -2\sum_{i=1}^{N}\left(\frac{\partial}{\partial P}P^T\right)(P_i - P)$$

where the derivative of P_i is zero, because it is constant. The derivative of a row vector with with respect to a column vector is a matrix. Every row of this matrix is the derivative of the row vector with the corresponding element of the column vector. In our case the derivative of P with respect to itself is the identity matrix $\mathbf{1}$. So

$$\frac{\partial Q(P)}{\partial P} = 2\sum_{i=1}^{N}\mathbf{1}(P_i - P) = 2\sum_{i=1}^{N}(P_i - P)$$

which if we equate to zero we get

$$P = \frac{\sum_{i=1}^{N} P_i}{N} \tag{2.1}$$

which is essentially the same as before.

3. Line Fitting

In the problem above we had a collection of points P_i and we found a point P that is closest to all of them, in the least squares sense. We can try something slightly more complex now, like finding a line l that is closest to all points P_i. Let line l be represented by two vectors

$$l = (p, q)$$

where a point P_i belongs to l iff there is a λ such that $p + \lambda q = P_i$. Vector p is a point on the line and vector q is the direction of the line. There are other ways to represent a line but this one suits our purpose better.

Since we want to minimize the distance of the points P_i from the line l we first need to express this distance as a nice and easy to use expression. There are two equivalent ways to define the point to line distance. The one is to define a normal line that goes through the point and intersects the line l at a right angle and then measure the distance of the point from the line l along the normal line. The second way is to find the distance of the point P_i from a point P'_i that lies on the line l and then slide the point P'_i along the line l till this distance is minimized. This minimal distance is the one we want. Since we have the machinery to minimize things we opt for the second approach. If you have a

hammer everything looks like a nail.

So the (squared) distance D_i^2 of the point P_i from the line l is

$$D_i^2 = \min_{\lambda}(P_i - p - \lambda q)^2$$

but if we want to do anything useful with it we have to get rid of the min symbol by finding the minimizing with respect to λ. As before we take derivatives

$$\frac{\partial(P_i - p - \lambda q)^2}{\partial \lambda} = -2q^T(P_i - p - \lambda q)$$

and by equating it to zero we get

$$\lambda = \frac{q^T(P_i - p)}{q^T q}$$

which gives us the expression for distance

$$D_i^2 = \left((P_i - p) - \frac{qq^T}{q^T q}(P_i - p)\right)^2$$

which, striving for elegance we rewrite as

$$D_i^2 = \left((\mathbf{1} - \frac{qq^T}{q^T q})(P_i - p)\right)^2. \tag{3.1}$$

Now we have an expression for the distance of a point P_i from the line l and it is already squared. To proceed with our least squares we sum up all these squared distances and find the line parameters p and q that minimize this sum. We start by defining the sum

$$Q(p, q) = \sum_{i=1}^{N} D_i^2$$

and we take the derivatives first with respect to p

$$\frac{\partial Q(p,q)}{\partial p} = \sum_{i=1}^{N} \frac{\partial D_i^2}{\partial p} = -2\sum_{i=1}^{N} \frac{\partial p^T}{\partial p}(\mathbf{1} - \frac{qq^T}{q^T q})(\mathbf{1} - \frac{qq^T}{q^T q})(P_i - p) =$$

$$-2\sum_{i=1}^{N}(\mathbf{1} - \frac{qq^T}{q^T q})(\mathbf{1} - \frac{qq^T}{q^T q})(P_i - p) =$$

$$-2(\mathbf{1} - \frac{qq^T}{q^T q})(\mathbf{1} - \frac{qq^T}{q^T q})\sum_{i=1}^{N}(P_i - p)$$

which we can equate to zero. We can verify that

$$p = \frac{\sum_{i=1}^{N} P_i}{N} \tag{3.2}$$

satisfies the resulting equation, although it is not a unique solution. Since p is just a point on the line it is as good as any other point on the line. But we prefer the one given by Eq. (3.2) because it is identical to the one given by Eq. (2.1) and does not contain q.

On to q now. It appears that Eq. (3.1) is not elegant enough and we should improve it. Consider the following well known identities for a vector v

$$v^2 = v \cdot v = v^T v = tr(vv^T)$$

where $tr(\cdots)$ is the trace operator (sum of diagonal elements). We are mainly interested in the last version to apply it to Eq. (3.1) which becomes

$$Q(p,q) = \sum_{i=1}^{N} tr\left(\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)(P_i - p)(P_i - p)^T\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\right)$$

and by noticing that the trace operator is linear and that the first and last parenthesized quantities do not depend on the index i we can rewrite it as

$$Q(p,q) = tr\left(\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\left(\sum_{i=1}^{N}(P_i - p)(P_i - p)^T\right)\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\right) =$$
$$N\ tr\left(\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)C\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\right) = N\ tr\left(\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)C\right)$$

where

$$C = \frac{\sum_{i=1}^{N}(P_i - p)(P_i - p)^T}{N}$$

and we applied the following property of the trace for two matrices A and B of appropriate dimensions

$$tr(AB) = tr(BA)$$

Then noticing that the product

$$\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\left(\mathbf{1} - \frac{qq^T}{q^T q}\right) =$$
$$\mathbf{1} + \frac{qq^T qq^T}{q^T qq^T q} - \frac{qq^T}{q^T q} - \frac{qq^T}{q^T q} =$$
$$\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)$$

$$Q(p,q) = N\ tr\left(C\left(\mathbf{1} - \frac{qq^T}{q^T q}\right)\right)$$

and by invoking the above property of the trace we get

$$Q(p,q) = N\ tr(C) - N\ tr\left(\frac{q^T C q}{q^2}\right) \tag{3.3}$$

which is undoubtedly the most elegant of equations.

Matrix C is a constant so it does not affect the minimization which can be achieved by maximizing the scalar

$$\frac{q^T C q}{q^2}$$

which is just a Rayleigh Quotient and is maximized when q is the eigenvector that corresponds to the largest eigenvalue.

There are a few remarks that we can make on this result. First, we find the line direction q without taking derivatives, just by using a canned theorem (we, in other words, outsourced the derivatives to Dr. Rayleigh). Second the points P_i can be of any dimension: two dimensional points on the image plane, three dimensional points in the real world or ten dimensional characters in a Douglas Adams novel. Third we can extend the result to structures of higher dimensions than lines, like fitting a plane in a four dimensional space, as we might need if we fit an affine flow to a set of image displacement data. And finally, the same technique can be applied to find the principal direction of any elongated object, even if we are not particularly interested in line fitting.