7. Gauss-Markov Theorem

It is often the case that one seeks a signal in a noise background. Examples would be the determination of a sinusoid in the presence of a background continuum stochastic process, and the determination of a trend in the presence of other processes with a different structure. A perhaps less obvious example, because it is too familiar, is the estimation of the mean value of a time series, in which the mean is the signal and the deviations from the mean are therefore arbitrarily defined as the "noise." The cases one usually encounters in elementary textbooks and classes are some simple variant of a signal in presence of

Continued on next page...

white noise. Unfortunately, while this case is easy to analyze, it usually does not describe the situation one is faced with in practice, when for example, one must try to determine a trend in the presence of red noise processes, ones which may exhibit local trends easily confused with a true secular (deterministic) trend. This problem plagues climate studies.

There are several approaches to finding some machinery which can help one to fall into statistical traps of confusing signal with noise. One widely applicable methodology is called variously "minimum variance estimation", the "Gauss-Markov Theorem", and in a different context, sometimes the "stochastic inverse." Although a more extended discussion is given in Wunsch (1996), or see Liebelt (1967) for a complete derivation, the following heuristic outline may help.

Let there be some set of N unknown parameters, written here as a vector, $\mathbf{s}$. We suppose that they have zero mean, $<\mathbf{s}>=\mathbf{0}$, and that there is a known covariance for $\mathbf{s}$, $\mathbf{R}_{ss} = \langle \mathbf{ss}^T \rangle$. Let there be a set of M measurements, written here also as another vector $\mathbf{y}$, also with zero mean, and known covariance $<\mathbf{yy}^T> = \mathbf{R}_{yy}$. Finally, we will assume that there is a known covariance between the measurements and the unknowns: $\mathbf{R}_{sy} = <\mathbf{sy}^T>$. Given these covariances (correlations), what can we say about $\mathbf{s}$, given $\mathbf{y}$, if we try to estimate any of the elements s_i, as a linear combination of the measurements:

$$\tilde{s}_i = \sum_{j=1}^{M} B_{ij} y_j. \tag{7.1}$$

The question is, how should the weights, B_{ij} be chosen? Arbitrarily, but reasonably, we seek B_{ij} such that the variance of the estimated s_i about the true value is as small as possible, that is,

$$\left\langle (\tilde{s}_i - s_i)^2 \right\rangle = \left\langle \left(\sum_{j=1}^{m} B_{ij} y_j - s_i \right)^2 \right\rangle, \ 1 \le i \le N \tag{7.2}$$

should be a minimum. Before proceeding, note that B_{ij} is really a matrix, and each row will be separately determinable for each element s_i. This simple observation permits us to re-write (7.2) in a matrix-vector form. Minimize the *diagonal elements* of:

$$< (\tilde{\mathbf{s}} - \mathbf{s}) (\tilde{\mathbf{s}} - \mathbf{s})^T > = < (\mathbf{By} - \mathbf{s}) (\mathbf{By} - \mathbf{s})^T > \equiv \mathbf{P}. \tag{7.3}$$

The important point here is that, in (7.3), we are meant to minimize the N separate diagonal elements, each separately determining a row of $\mathbf{B}$; but we can use the notation to solve for all rows simultaneously.

At this stage, one expands the matrix product in (7.3), and uses the fact that quantities such as $<\mathbf{ss}^T> = \mathbf{R}_{ss}$ are known. One can show without too much difficulty (it involves invoking the properties of positive definite matrices) that the minimum of the diagonals is given by the unique choice,

$$\mathbf{B} = \mathbf{R}_{sy} \mathbf{R}_{yy}^{-1}, \tag{7.4}$$

with the first row being the solution for $\tilde{s}_1$, etc.

The result (7.4) is general and abstract. Let us now consider a special case in which the measurements y_q are some linear combination of the parameters, corrupted by noise, that is, $y_q = \sum_{l=1}^{N} E_{ql} s_l + n_q$, which

can also be written generally as,

$$\mathbf{Es} + \mathbf{n} = \mathbf{y}. \tag{7.5}$$

With this assumption, we can evaluate

$$\mathbf{R}_{sy} = <\mathbf{s}\,(\mathbf{Es} + \mathbf{n})^T> = \mathbf{R}_{ss}\mathbf{E}^T, \tag{7.6}$$

assuming $<\mathbf{sn}^T> = 0$, and

$$\mathbf{R}_{yy} = \mathbf{ER}_{ss}\mathbf{E}^T + \mathbf{R}_{nn} \tag{7.7}$$

where $\mathbf{R}_{nn} = <\mathbf{nn}^T>$. Then one has immediately,

$$\mathbf{B} = \mathbf{R}_{ss}\left(\mathbf{ER}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1}, \tag{7.8}$$

and

$$\tilde{\mathbf{s}} = \mathbf{R}_{ss}\left(\mathbf{ER}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1}\mathbf{y}. \tag{7.9}$$

There is one further, extremely important step: how good is this estimate? This question is readily answered by substituting the value of $\mathbf{B}$ back into the expression (7.3) for the actual covariance about the true value. We obtain immediately,

$$\mathbf{P} = \mathbf{R}_{ss} - \mathbf{R}_{ss}\mathbf{E}^T\left(\mathbf{ER}_{ss}\mathbf{E}^T + \mathbf{R}_{nn}\right)\mathbf{ER}_{ss}. \tag{7.10}$$

One of the benefits of the general approach is that we have obtain the complete matrix $\mathbf{P}$, which gives us not only the variances (uncertainties) of each of the $\tilde{s}_i$ about the true value, but also the covariances of these errors or uncertainties in each, with all the others—they do after all, depend upon the same data—so that it is no surprise that they would have correlated errors.)

A special case, written out in Wunsch (1996), and which is particularly illuminating is the simple problem of determining a mean value (so that $\mathbf{s}$ is a scalar), in the presence of a noise field which has an arbitrary correlation. One finds there, that the uncertainty of the mean can be vastly greater than the conventional estimates based upon white noise, if the noise is correlated in time.

REMARK 2. *A common complaint among beginning users of Gauss-Markov and related estimation methods is: "I have no idea what the covariances are. This all becomes completely arbitrary if I just make up something." The answer to this worry is found by examining the statement "I have no idea what the covariances are." If this is really true, it means that an acceptable answer for any element, $\tilde{s}_i$ could have any value at all, including something infinitesimal, 10^{-40}, or astronomical, $\pm 10^{40}$ and one would say "that's acceptable, because I know nothing at all about the solution". The reader may say, "that's not what I really meant". In fact, it is extremely rare to be completely ignorant, and if one is completely ignorant, so that any values at all would be accepted, the investigator ought perhaps to stop and ask if the problem makes any sense? More commonly, one usually knows something, e.g., that the parameters are very unlikely to be bigger than about $\pm S_0$. If that is all one is willing to say, then one simply takes*

$$\mathbf{R}_{ss} = S_0^2\mathbf{I}_N \tag{7.11}$$

with something analogous perhaps, for $\mathbf{R}_{nn}$*, which becomes an estimate of the noise magnitude. Letting* $S_0^2 \to \infty$ *is the appropriate limit if one really knows nothing, and one might study 7.10) in that limit. The point is, that the estimation procedure can use whatever information one has, and one need not stipulate anything that is truly unknown. In particular, if one does not know the non-diagonal elements of the covariances, one need not state them. All that will happen is that a solution will be obtained that likely will have a larger uncertainty than one could have obtained had additional information been available. The more information one can provide, the better the estimate. A final comment of course, is that one* must *check that the solution and the errors left are actually consistent with what was postulated for the covariances. If the solution and noise residuals are clearly inconsistent with* $\mathbf{R}_{ss}$*, etc., one should trash the solution and try to understand what was wrong; this is a very powerful test, neglected at one's peril. If used, it can rescue even the naivest user from a truly silly result.*