2.7 Minimum Variance Estimation & Simultaneous Equations

The fundamental objective for least-squares is minimization of the noise norm (2.90), although we complicated the discussion somewhat by introducing trade-offs against $\|\tilde{\mathbf{x}}\|$, various weights in the norms, and even the restriction that $\tilde{\mathbf{x}}$ should satisfy certain equations exactly. Least-squares methods, whether used directly as in (2.96) or indirectly through the vector representations of the SVD, are fundamentally deterministic. Statistics were used only to understand the sensitivity of the solutions to noise, and to obtain measures of the expected deviation of the solution from some supposed truth.

But there is another, very different, approach to obtaining estimates of the solution to equation sets like (2.88), directed more clearly toward the physical goal: to find an estimate $\tilde{\mathbf{x}}$ which deviates as little as possible in the *mean-square* from the true solution. That is, we wish to minimize the statistical quantities $\langle(\tilde{\mathbf{x}}_i - x_i)^2\rangle$ for all i. The next section is devoted to understanding how to find such an $\tilde{\mathbf{x}}$ (and the corresponding $\tilde{\mathbf{n}}$), through an excursion into statistical estimation theory. It is far from obvious that this $\tilde{\mathbf{x}}$ should bear any resemblance to one of the least-squares estimates; but as will be seen, under some circumstances the two are identical. Their possible identity is extremely useful, but has apparently led many investigators to seriously confuse the methodologies, and therefore the interpretation of the result.

2.7.1 The Fundamental Result

Suppose we are interested in making an estimate of a physical variable, $\mathbf{x}$, which might be a vector or a scalar, and is either constant or varying with space and time. To be definite, let $\mathbf{x}$ be a function of an independent variable $\mathbf{r}$, written discretely as $\mathbf{r}_j$ (it might be a vector of space coordinates, or a scalar time, or an accountant's label). Let us make some suppositions

about what is usually called "prior information." In particular, suppose we have an estimate of the low-order statistics describing $\mathbf{x}$, that is, specifying its mean and second moments:

$$\langle \mathbf{x} \rangle = \mathbf{0}\,, \qquad \langle \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j)^T \rangle = \mathbf{R}_{xx}(\mathbf{r}_i,\, \mathbf{r}_j)\,. \tag{2.386}$$ {36001}

To make this problem concrete, one might think of $\mathbf{x}$ as being the temperature anomaly (about the mean) at a fixed depth in the ocean (a scalar) and $\mathbf{r}_j$ a vector of horizontal positions; or conductivity in a well, where $\mathbf{r}_j$ would be the depth coordinate, and $\mathbf{x}$ is the vector of scalars at any location, $\mathbf{r}_p$, $x_p = x\,(\mathbf{r}_p)$. Alternatively, $\mathbf{x}$ might be the temperature at a fixed point, with r_j being the scalar of time. But if the field of interest is the velocity vector, then each element of $\mathbf{x}$ is itself a vector, and one can extend the notation in a straightforward fashion. To keep the notation a little cleaner, however, we will treat the elements of $\mathbf{x}$ as scalars.

Now suppose that we have some observations, y_i, as a function of the same coordinate $\mathbf{r}_i$, with a known, zero mean, and second moments

$$\mathbf{R}_{yy}\,(\mathbf{r}_i, \mathbf{r}_j) = \langle \mathbf{y}\,(\mathbf{r}_i)\,\mathbf{y}\,(\mathbf{r}_j)^T \rangle\,, \qquad \mathbf{R}_{xy}(\mathbf{r}_i,\, \mathbf{r}_j) = \langle \mathbf{x}(\mathbf{r}_i)\mathbf{y}(\mathbf{r}_j)^T \rangle\,, \quad 1 \le i,j \le M \tag{2.387}$$ {36002}

(the individual observation elements can also be vectors—for example, two or three components of velocity and a temperature at a point—but as with $\mathbf{x}$, the modifications required to treat this case are straightforward, and we here assume scalar observations). Could the measurements be used to make an estimate of $\mathbf{x}$ at a point $\tilde{\mathbf{r}}_\alpha$ where no measurement is available? Or could many measurements be useto obtain a better estimate even at points where there exists a measurement? The idea is to exploit the concept that finite covariances carry predictive capabilities from known variables to unknown ones. A specific example would be to suppose the measurements are of temperature $y(\mathbf{r}_j) = y_0(\mathbf{r}_j) + n(\mathbf{r}_j)$, where n is the noise and we wish to estimate the temperature at different locations, perhaps on a regular grid $\tilde{\mathbf{r}}_\alpha$, $1 \le \alpha \le N$. This special problem is one of gridding or mapmaking (the tilde is placed on $\mathbf{r}_\alpha$ as a device to emphasize that this is a location where an estimate is sought; the numerical values of these places or labels are assumed known). Alternatively, and somewhat more interesting, perhaps the measurements are more indirect, with $y(r_i)$ representing a velocity field component at depth in a fluid and believed connected through a differential equation to the temperature field. We might want to estimate the temperature from measurements of the velocity.

Given the previous statistical discussion (P. 30), it is reasonable to ask for an estimate $\tilde{x}(\tilde{\mathbf{r}}_\alpha)$, whose dispersion about its true value, $x(\tilde{\mathbf{r}}_\alpha)$ is as small as possible, that is,

$$P\,(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\alpha) = \langle (\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta)) \rangle|_{\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta}$$

is to be minimized. If we would like to answer the question for more than one point, and if we would like to understand the covariance of the errors of our estimates at various points $\tilde{\mathbf{r}}_\alpha$, then

we can form a vector of values to be estimated, $\{\tilde{x}(\mathbf{r}_\alpha)\} \equiv \tilde{\mathbf{x}}$, and the uncertainty among them,

{36003}

$$\begin{aligned}\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) &= \langle(\tilde{x}(\tilde{\mathbf{r}}_\alpha) - x(\tilde{\mathbf{r}}_\alpha))(\tilde{x}(\tilde{\mathbf{r}}_\beta) - x(\tilde{\mathbf{r}}_\beta))\rangle \\ &= \langle(\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T\rangle\,, \quad 1 \le \alpha \le N\,,\ 1 \le \beta \le N\,,\end{aligned} \tag{2.388}$$

where the *diagonal* elements, $\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\alpha)$, are to be *individually* minimized (not in the sum of squares). Thus we seek the solution with *minimum variance about the correct value.*

What should the relationship be between data and estimate? At least initially, one might try a linear combination of data,

$$\tilde{x}(\tilde{\mathbf{r}}_\alpha) = \sum_{j=1}^{M} B(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j) y(\mathbf{r}_j)\,, \tag{2.389}$$

{36004}

for all α, which makes the diagonal elements of $\mathbf{P}$ in (2.388) as small as possible. By letting $\mathbf{B}$ be an $N \times M$ matrix all the points can be handled at once,

$$\tilde{\mathbf{x}}\,(\tilde{\mathbf{r}}_\alpha) = \mathbf{B}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)\mathbf{y}\,(\mathbf{r}_j)\ . \tag{2.390}$$

{36005}

(This notation is redundant. Eq. (2.390) is a shorthand for (2.389), in which the argument has been put into $\mathbf{B}$ explicitly as a reminder that there is a summation over all the data locations $\mathbf{r}_j$ for all mapping locations $\tilde{\mathbf{r}}_\alpha$, but it is automatically accounted for by the usual matrix multiplication convention. It suffices to write $\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}$.)

An important result, often called the "Gauss-Markov theorem," produces the values of $\mathbf{B}$ that will minimize the diagonal elements of $\mathbf{P}$.[54] Substituting (2.390) into (2.388) and expanding,

$$\begin{aligned}&\mathbf{P}(\tilde{\mathbf{r}}_\alpha, \tilde{\mathbf{r}}_\beta) = \langle(\mathbf{B}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)\mathbf{y}\,(\mathbf{r}_j) - \mathbf{x}(\tilde{\mathbf{r}}_\alpha))(\mathbf{B}(\tilde{\mathbf{r}}_\beta, \mathbf{r}_l)\mathbf{y}\,(\mathbf{r}_l) - \mathbf{x}(\tilde{\mathbf{r}}_\beta))^T\rangle \\ &\equiv \langle(\mathbf{B}\mathbf{y} - \mathbf{x})(\mathbf{B}\mathbf{y} - \mathbf{x})^T\rangle \\ &= \mathbf{B}\left\langle\mathbf{y}\mathbf{y}^T\right\rangle - \left\langle\mathbf{x}\mathbf{y}^T\right\rangle\mathbf{B}^T - \mathbf{B}\left\langle\mathbf{y}\mathbf{x}^T\right\rangle + \left\langle\mathbf{x}\mathbf{x}^T\right\rangle\end{aligned} \tag{2.391}$$

{36006}

Using $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, Eq. (2.391) is,

$$\mathbf{P} = \mathbf{B}\mathbf{R}_{yy}\mathbf{B}^T - \mathbf{R}_{xy}\mathbf{B}^T - \mathbf{B}\mathbf{R}_{xy}^T + \mathbf{R}_{xx}\,. \tag{2.392}$$

{36007}

Notice that because $\mathbf{R}_{xx}$ represents the moments of $\mathbf{x}$ evaluated at the estimation positions, it is a function of $\tilde{\mathbf{r}}_\alpha$, $\tilde{\mathbf{r}}_\beta$, whereas $\mathbf{R}_{xy}$ involves covariances of $\mathbf{y}$ at the data positions with $\mathbf{x}$ at the estimation positions, and is consequently a function $\mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_j)$.

Now, using the matrix identity (2.38)—that is, completing the square (adding and subtracting $\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T$), (2.392) becomes,

$$\mathbf{P} = (\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^T + \mathbf{R}_{xx}\,. \tag{2.393}$$

{36009}

Setting $\tilde{\mathbf{r}}_\alpha = \tilde{\mathbf{r}}_\beta$ so that (2.393) is the variance of the estimate at point $\tilde{\mathbf{r}}_\alpha$ about its true value, and noting that all three terms in Eq. (2.393) are positive definite, minimization of any diagonal element of $\mathbf{P}$ is obtained by choosing $\mathbf{B}$ so that the first term vanishes or,

$$\mathbf{B}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j) = \mathbf{R}_{xy}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_i\right) \mathbf{R}_{yy}\left(\mathbf{r}_i, \mathbf{r}_j\right)^{-1} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\,. \tag{2.394}$$

{36010}

(The diagonal elements of $(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^T$ need to be written out explicitly to see that Eq. (2.394) is necessary. Consider the 2×2 case: The first term of Eq. (2.393) is of the form,

$$\left\{ \begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array} \right\} \left\{ \begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right\} \left\{ \begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array} \right\}^T ,$$

where $\mathbf{C} = \mathbf{B} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$. Then, one has the diagonal of,

$$\left\{ \begin{array}{cc} C_{11}^2 R_{11} + C_{12}C_{11}\left(R_{21} + R_{12}\right) + C_{12}^2 R_{22} & \cdot \\ \cdot & C_{21}^2 R_{11} + C_{21}C_{22}\left(R_{21} + R_{12}\right) + C_{22}^2 R_{22} \end{array} \right\},$$

and these diagonals vanish (with $R_{11}, R_{22} > 0$, only if $C_{11} = C_{12} = C_{21} = C_{22} = 0$ or, $\mathbf{B} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$). Thus the minimum variance estimate is,

$$\tilde{\mathbf{x}}(\tilde{\mathbf{r}}_\alpha) = \mathbf{R}_{xy}\left(\tilde{\mathbf{r}}_\alpha, \mathbf{r}_i\right) \mathbf{R}_{yy}^{-1}\left(\mathbf{r}_i, \mathbf{r}_j\right) \mathbf{y}\left(\mathbf{r}_j\right)\,, \tag{2.395}$$

{36011}

and the actual minimum value of the diagonal elements of $\mathbf{P}$ is found by substituting back into (2.392) producing,

$$\mathbf{P}(\tilde{\mathbf{r}}_\alpha,\, \tilde{\mathbf{r}}_\beta) = \mathbf{R}_{xx}(\tilde{\mathbf{r}}_\alpha,\, \tilde{\mathbf{r}}_\beta) - \mathbf{R}_{xy}(\tilde{\mathbf{r}}_\alpha,\, \mathbf{r}_j)\mathbf{R}_{yy}^{-1}(\mathbf{r}_j,\, \mathbf{r}_k)\mathbf{R}_{xy}^T(\tilde{\mathbf{r}}_\beta,\, \mathbf{r}_k)\,. \tag{2.396}$$

{36012}

{pagemap1}

The bias of (2.396) is

$$\langle \tilde{\mathbf{x}} - \ \mathbf{x}\rangle = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\langle \mathbf{y}\rangle - \mathbf{x}\,. \tag{2.397}$$

{36013}

If $\langle \mathbf{y}\rangle = \mathbf{x} = 0$, the estimator is unbiased , and called a "best linear unbiased estimator," or "BLUE"; otherwise it is biassed. The whole development here began with the assumption that $\langle \mathbf{x}\rangle = \langle \mathbf{y}\rangle = 0$; what is usually done is to remove the *sample* mean from the observations $\mathbf{y}$, and to ignore the difference between the true and sample means. An example of using this machinery for mapping purposes will be seen in Ch. 3. Under some circumstances, this approximation is unacceptable, and one must account for the mapping error introduced by the use of the sample mean. A general approach falls under the label of "kriging", which is also briefly discussed in Chapter 3.

2.7.2 Linear Algebraic Equations

The result (2.394)–(2.396) is the abstract general case and is deceptively simple. Invocation of the physical problem of interpolating temperatures etc., is not necessary: the only information actually used is that there are finite covariances between $\mathbf{x}, \mathbf{y}, \mathbf{n}$. Although we will explicitly explore its use for mapping in Chapter 3, suppose instead that the observations are related to the unknown vector $\mathbf{x}$ as in our canonical problem, that is, through a set of linear equations: $\mathbf{Ex} + \mathbf{n} = \mathbf{y}$. The measurement covariance, $\mathbf{R}_{yy}$, can then be computed directly as:

$$\mathbf{R}_{yy} = \langle (\mathbf{Ex} + \mathbf{n})(\mathbf{Ex} + \mathbf{n})^T \rangle = \mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}. \tag{2.398}$$ {36014}

The unnecessary, but simplifying and often excellent, assumption was made that the cross-terms of form,

$$\mathbf{R}_{xn} = \mathbf{R}_{nx}^T = \mathbf{0}, \tag{2.399}$$ {36015}

so that

$$\mathbf{R}_{xy} = \langle \mathbf{x}(\mathbf{Ex} + \mathbf{n})^T \rangle = \mathbf{R}_{xx}\mathbf{E}^T, \tag{2.400}$$ {36016}

that is, there is no correlation between the measurement noise and the actual state vector (e.g., that the noise in a temperature measurement does not depend upon whether the true value is 10° or 25°).

Under these circumstances, Eqs. (2.395), (2.396) take on the form:

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{y} \tag{2.401}$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \tag{2.402}$$

$$\mathbf{P} = \mathbf{R}_{xx} - \mathbf{R}_{xx}\mathbf{E}^T \left(\mathbf{E}\mathbf{R}_{xx}\mathbf{E}^T + \mathbf{R}_{nn}\right)^{-1} \mathbf{E}\mathbf{R}_{xx} \tag{2.403}$$

These latter expressions are extremely important; they permit discussion of the solution to a set of linear algebraic equations in the presence of noise using information concerning the statistics of both the noise and the solution. Notice that they are *identical to the least-squares expression* (2.136) *if* $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$, except that there the uncertainty was estimated about the mean solution; here it is taken about the true one. As is generally true of all linear methods, the uncertainty, $\mathbf{P}$, is independent of the actual data, and can be computed in advance should one wish.

From the matrix inversion lemma, Eqs. (2.401, 2.403) can be rewritten

$$\tilde{\mathbf{x}} = \left(\mathbf{R}_{xx}^{-1}+\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{y} \tag{2.404}$$

$$\tilde{\mathbf{n}} = \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}} \tag{2.405}$$

$$\mathbf{P} = \left(\mathbf{R}_{xx}^{-1}+\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1} \tag{2.406}$$

Although these alternate forms are algebraically and numerically identical to Eqs. (2.401-2.403), the size of the matrices to be inverted changes from $M \times M$ matrices to $N \times N$, where $\mathbf{E}$ is $M \times N$ (but note that $\mathbf{R}_{nn}$ is $M \times M$; the efficacy of this alternate form may depend upon whether the *inverse* of $\mathbf{R}_{nn}$ is known). Depending upon the relative magnitudes of M, N, one form may be much preferable to the other. Finally, (2.406) has an important interpretation we will discuss when we come to recursive methods. Recall, too, the options we had with the SVD of solving $M \times M$ or $N \times N$ problems. Note that in the limit of complete *a priori* ignorance of the solution, $\left\|\mathbf{R}_{xx}^{-1}\right\| \to 0$, Eqs. (2.404, 2.406) reduce to,

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{y},$$
$$\mathbf{P} = \left(\mathbf{E}^T\mathbf{R}_{nn}^{-1}\mathbf{E}\right)^{-1},$$

the conventional weighted least-squares solution, now with $\mathbf{P} = \mathbf{C}_{xx}$. More generally, the presence of finite $\mathbf{R}_{xx}^{-1}$ introduces a bias into the solution so that $\langle\tilde{\mathbf{x}}\rangle \neq \mathbf{x}$, which, however, produces a smaller solution variance than in the unbiased solution.

The solution (2.401-2.403, 2.404-2.406) is an "estimator"; it was found by demanding a solution with the minimum dispersion about the true solution and is found, surprisingly, to be identical with the tapered, weighted least-squares solution when $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$, the least-squares objective function weights are chosen, as is commonly done. This correspondence of the two solutions often leads them to be seriously confused. It is essential to recognize that the logic of the derivations are quite distinct: We were free in the least-squares derivation to use weight matrices which were anything we wished—as long as appropriate inverses existed.

The correspondence of least-squares with what is usually known as minimum variance estimation can be understood by recognizing that the Gauss-Markov estimator was derived by minimizing a quadratic objective function. The least-squares estimate was obtained from minimizing a summation which is a sample *estimate* of the Gauss-Markov objective function when $\mathbf{S} = \mathbf{R}_{xx}$, $\mathbf{W} = \mathbf{R}_{nn}$.

2.7.3 Testing After the Fact

As with any statistical estimator, an essential step is the testing after an apparent solution has been found, that the behavior of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ is consistent with the assumed prior statistics reflected in $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$, and any assumptions about their means or other properties. Such *a posteriori* checks are both necessary and very demanding. One sometimes hears it said that estimation using Gauss-Markov and related methods is "pulling solutions out of the air" because the prior covariance matrices $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$ often are only poorly known. But producing solutions which pass the test of consistency with the prior covariances can be very difficult. It is also true that the solutions tend to be somewhat insensitive to the details of the prior covariances and it is easy to become overly concerned with the detailed structure of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$.

As stated previously, it is also rare to be faced with a situation in which one is truly ignorant of the covariances, true ignorance meaning that arbitrarily large or small numerical values of x_i, n_i would be acceptable. In the box inversions of Chapter 1 (to be revisited in Chapter 5), solution velocities of order 1000 cm/s might be regarded as absurd, and their absurdity is readily asserted by choosing $\mathbf{R}_{xx} = \text{diag}(10\text{cm/s})^2$, which reflects a mild belief that velocities are 0(10cm/s) with no known correlations with each other. Testing of statistical estimates against prior hypotheses is a highly developed field in applied statistics, and we leave it to the references already listed for their discussion. Should such tests be failed, one must reject the solutions $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and ask why they failed—as it usually implies an incorrect model, ($\mathbf{E}$, and the assumed statistics of solution and/or noise).

Example

The underdetermined system

$$\left\{\begin{matrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{matrix}\right\} \mathbf{x} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

with noise variance $\langle \mathbf{n}\mathbf{n}^T \rangle = .01\mathbf{I}$, *has a solution, if* $\mathbf{R}_{xx} = \mathbf{I}$, *of*

$$\tilde{\mathbf{x}} = \mathbf{E}^T(\mathbf{E}\mathbf{E}^T + .01\mathbf{I})^{-1}\mathbf{y} = \begin{bmatrix} 0 & .4988 & .4988 & 0 \end{bmatrix}^T, \qquad \tilde{\mathbf{n}} = \begin{bmatrix} .0025 & -.0025 \end{bmatrix}^T.$$

If the solution was thought to be large scale and smooth, one might use the covariance

$$\mathbf{R}_{xx} = \begin{Bmatrix} 1 & .999 & .998 & .997 \\ .999 & 1 & .999 & .998 \\ .998 & .999 & 1 & .999 \\ .997 & .998 & .999 & 1 \end{Bmatrix},$$

which produces a solution,

$$\tilde{\mathbf{x}} = \begin{bmatrix} 0.2402 \pm 0.028 & 0.2595 \pm 0.0264 & 0.2595 \pm 0.0264 & 0.2402 \pm 0.0283 \end{bmatrix}^T,$$

$$\tilde{\mathbf{n}} = \begin{bmatrix} 0.0006 & -0.9615 \end{bmatrix}^T,$$

which has the desired large-scale property. (One might worry a bit about the structure of the residuals; but two equations are wholly inadequate to draw any conclusions.)

2.7.4 Use of Basis Functions

{sisfunctions}

A superficially different way of dealing with prior statistical information is often commonly used. Suppose that the indices of x_i refer to a spatial or temporal position, call it r_i, so that $x_i = x(r_i)$. Then it is often sensible to consider expanding the unknown $\mathbf{x}$ in a set of basis functions, F_j, for example, sines and cosines, Chebyschev polynomials, ordinary polynomials, etc. One might write

$$x(r_i) = \sum_{j=1}^{L} \alpha_j F_j(r_i)$$

or

$$\mathbf{x} = \mathbf{F}\boldsymbol{\alpha}\,, \quad \mathbf{F} = \begin{Bmatrix} F_1(r_1) & F_2(r_1) & \cdots & F_L(r_1) \\ F_1(r_2) & F_2(r_2) & \cdots & F_L(r_2) \\ . & . & . & . \\ F_1(r_N) & F_2(r_N) & \cdots & F_L(r_N) \end{Bmatrix}, \quad \boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_L]^T$$

which, when substituted into (2.88), produces

{36019}

$$\mathbf{L}\boldsymbol{\alpha} + \mathbf{n} = \mathbf{y}\,, \quad \mathbf{L} = \mathbf{EF}\,. \tag{2.407}$$

If $L < M < N$, one can convert an underdetermined system into one which is formally overdetermined and, of course, the reverse is possible as well. It should be apparent, however, that the

solution to (2.407) will have a covariance structure dictated in large part by that contained in the basis functions chosen, and thus there is no fundamental gain in employing basis functions although they may be convenient, numerically or otherwise. If $\mathbf{P}_{\alpha\alpha}$ denotes the uncertainty of $\boldsymbol{\alpha}$ then,

$$\mathbf{P} = \mathbf{F}\mathbf{P}_{\alpha\alpha}\mathbf{F}^T, \tag{2.408}$$

is the uncertainty of $\tilde{\mathbf{x}}$. If there are special conditions applying to $\mathbf{x}$, such as boundary conditions at certain positions, r_B, a choice of basis function satisfying those conditions could be more convenient than appending them as additional equations.

Example

If, in the last example, one attempts a solution as a first order polynomial,

$$x_i = a + br_i, \quad r_1 = 0,\ r_2 = 1,\ r_3 = 2, \ldots$$

the system will become two equations in the two unknowns a, b:

$$\mathbf{EF}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{Bmatrix} 4 & 6 \\ 0 & 0 \end{Bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{n} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

and if no prior information about the covariance of a, b is provided,

$$[\tilde{a},\ \tilde{b}] = [0.0769,\ 0.1154]\,,$$

$$\tilde{\mathbf{x}} = \Big[0.0769 \pm 0.0077 \;\; 0.1923 \pm 0.0192 \;\; 0.3076 \pm 0.0308 \;\; 0.4230 \pm 0.0423\Big]^{\,T},$$

$$\tilde{\mathbf{n}} = [0.0002,\ -1.00]^T\,,$$

which is also large scale and smooth, but clearly different than that from the Gauss-Markov estimator. Although this latter solution has been obtained from a just-determined system, it is not clearly "better." If a linear trend is expected in the solution, then the polynomial expansion is certainly convenient—although such a structure can be imposed through use of $\mathbf{R}_{xx}$ *by specifying a growing variance with* r_i.

2.7.5 Determining a Mean Value

Let the measurements of the physical quantity continue to be denoted y_i and suppose that each is made up of an unknown large scale mean, m, plus a deviation from that mean of θ_i. Then,

$$m + \theta_i = y_i\,, \quad 1 \le i \le M \tag{2.409}$$ {36020a}

or

$$\mathbf{D}m + \boldsymbol{\theta} = \mathbf{y}\,, \qquad \mathbf{D} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \end{bmatrix}^T\,, \tag{2.410}$$ {36020b}

and we seek a best estimate, $\tilde{m}$, of m. In (2.409) or (2.410) the unknown $\mathbf{x}$ has become the scalar m, and the deviation of the field from its mean is the noise, that is, $\boldsymbol{\theta} \equiv \mathbf{n}$, whose true mean is zero. The problem is evidently a special case of the use of basis functions, in which only one function—a zeroth–order polynomial, m, is retained.

Set $\mathbf{R}_{nn} = \langle \boldsymbol{\theta}\boldsymbol{\theta}^T \rangle$. If, for example, we were looking for a large-scale mean temperature in a fluid flow filled with eddies, then $\mathbf{R}_{nn}$ is the sum of the covariance of the eddy field plus that of observational errors and any other fields contributing to the difference between y_i and the true mean m. To be general, suppose $\mathbf{R}_{xx} = \langle m^2 \rangle = m_0^2$, and from (2.404),

$$\begin{aligned} \tilde{m} &= \left\{ \frac{1}{m_0^2} + \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{D} \right\}^{-1} \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{y} \\ &= \frac{1}{1/m_0^2 + \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{D}}\, \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{y}. \end{aligned} \tag{2.411}$$ {36021b}

($\mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{D}$ is a scalar).[55] The expected uncertainty of this estimate is (2.406),

$$P = \left\{ \frac{1}{m_0^2} + \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{D} \right\}^{-1} = \frac{1}{1/m_0^2 + \mathbf{D}^T\mathbf{R}_{nn}^{-1}\mathbf{D}}\,, \tag{2.412}$$ {36021c}

(also a scalar).

The estimates may appear somewhat unfamiliar; they reduce to more common expressions in certain limits. Let the θ_i be uncorrelated, with uniform variance σ^2; $\mathbf{R}_{nn}$ is then diagonal and (2.411) reduces to,

$$\tilde{m} = \frac{1}{(1/m_0^2 + M/\sigma^2)\sigma^2} \sum_{i=1}^{M} y_i = \frac{m_0^2}{\sigma^2 + Mm_0^2} \sum_{i=1}^{M} y_i\,, \tag{2.413}$$ {36022a}

where the relations $\mathbf{D}^T\mathbf{D} = M$, $\mathbf{D}^T\mathbf{y} = \sum_{i=1}^{M} y_i$ were used. The expected value of the estimate is

$$\langle \tilde{m} \rangle = \frac{m_0^2}{\sigma^2 + Mm_0^2} \sum_{i}^{M} \langle y_i \rangle = \frac{m_0^2}{\sigma^2 + Mm_0^2} Mm \neq m\,, \tag{2.414}$$ {36022b}

that is, it is biassed, as inferred above, unless $\langle y_i \rangle = 0$, implying $m = 0$. $\mathbf{P}$ becomes,

$$P = \frac{1}{1/m_0^2 + M/\sigma^2} = \frac{\sigma^2 m_0^2}{\sigma^2 + Mm_0^2}\,. \tag{2.415}$$ {36022c}

Under the further assumption that $m_0^2 \to \infty$,

$$\tilde{m} = \frac{1}{M} \sum_{i=1}^{M} y_i\,, \tag{2.416}$$ {36022d}

$$P = \sigma^2/M\,, \tag{2.417}$$ {36022e}

which are the ordinary average and its variance (the latter expression is the well-known "square root of M rule" for the standard deviation of an average; recall Eq. (2.43)); $\langle \tilde{m} \rangle$ in (2.416) is readily seen to be the true mean—this estimate has become unbiased. But the magnitude of (2.417) always exceeds that of (2.415)—acceptance of bias in the estimate (2.413) reduces the uncertainty of the result—a common trade-off in estimation problems.

{emeanvalue2}

Eqs. (2.411)–(2.412) are the more general estimation rule—accounting through $\mathbf{R}_{nn}$ for correlations in the observations and their irregular distribution. Because many samples are not independent, (2.417) may be extremely optimistic. Eq. (2.412) gives one the appropriate expression for the variance when the data are correlated (that is, when there are fewer degrees of freedom than the number of sample points).

Example

The mean is needed for the M =500 values of the measured time series, shown in Fig. 2.14. If one calculates the ordinary average, $\tilde{m} = -20.0$, and the standard error, treating the measurements as uncorrelated, is by Eq. (2.417) is ± 0.31. *If on the other hand, one uses the covariance function displayed in Fig. 2.14, and (Eqs. 2.411, 2.412) with $m_0^2 \to \infty$, one obtains $\tilde{m} = -23.7$, with a standard error of* ± 20. *The true mean of the time series is actually zero (it was generated that way), and one sees the dire effects of assuming uncorrelated measurement noise, when the correlation is actually very strong. Within 2 standard deviations (a so-called 95% confidence interval for the mean), one finds, correctly, that the sample mean is indistinguishable from zero, whereas the mean assuming uncorrelated noise would appear to be very well determined and markedly different from zero.*[56] *(One might be tempted to apply a transformation to render the observations uncorrelated before averaging, and so treat the result as having M degrees-of-freedom. But recall, e.g. that for Gaussian variables (P. 37), the resulting numbers will have different variances, and one would be averaging apples and oranges.*

The use of the prior estimate, m_0^2, is interesting. Letting m_0^2 go to infinity does not mean that an infinite mean is expected ((2.416) is finite). It is is merely a statement that there is no information whatever, before we start, as to the magnitude of the true average—it could be arbitrarily large (or small and of either sign) and if it came out that way, would be acceptable. Such a situation is, of course, unlikely and even though we might choose not to use information concerning the probable size of the solution, we should remain aware that we could do so (the importance of the prior estimate diminishes as M grows—so that with an infinite amount of data it has no effect at all on the estimate). If a prior estimate of m itself is available, rather than just its mean square, the problem should be reformulated as one for the estimate of the perturbation about this value.

It is very important not to be tempted into making a first estimate of m_0^2 by using (2.416),

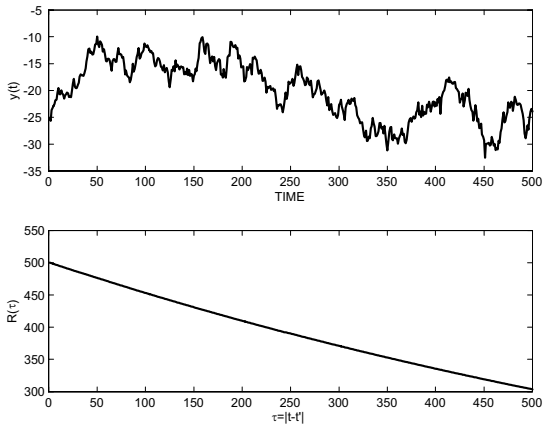

Figure 2.14: Time series y_t (upper panel) whose mean is required. Lower panel displays the autocovariance $\langle y_t y_{t'} \rangle$ as a function of $|t - t'|$ (in this special case, it does not depend upon t, t' separately.) True mean of y_t is zero by construction.

corrmean.eps}

substituting into (2.413), thinking to reduce the error variance. For the Gauss-Markov theorem to be valid, the prior information must be truly independent of the data being used.