2.4 Least-Squares

Much of what follows in this book can be described using very elegant and powerful mathematical tools. On the other hand, by restricting ourselves to discrete models and finite numbers of measurements (all that ever goes into a digital computer), almost everything can also be viewed as a form of ordinary least-squares, providing a much more intuitive approach than one through functional analysis. It is thus useful to go back and review what "everyone knows" about this most-familiar of all approximation methods.

2.4.1 Basic Formulation

{pagestraightl

Consider the elementary problem motivated by the "data" shown in figure 2.2. t is supposed to be an independent variable, which could be time, or a spatial coordinate or just an index. Some physical variable, call it $\theta(t)$, perhaps temperature at a point in a laboratory tank, has been measured at coordinates $t = t_i$, $1 \leq i \leq M$, as depicted in the figure.

We have reason to believe that there is a linear relationship between $\theta(t)$ and t in the form

Continued on next page...

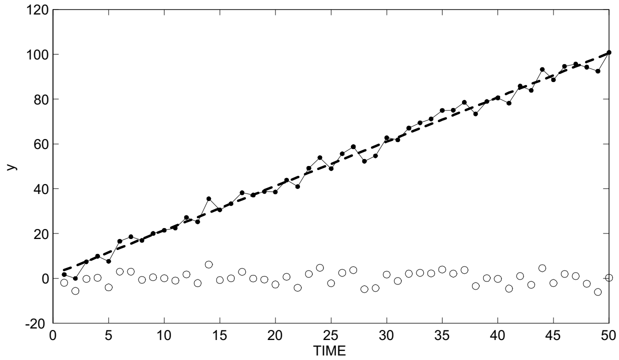

Figure 2.2: "Data" generated through the rule $y = 1 + 2t + n_t$, where $\langle n_t \rangle = 0$, $\langle n_i n_j \rangle = 9\delta_{ij}$ shown as + connected by the solid line. Dashed line is the simple least-squares fit, $\tilde{y} = 1.69 \pm 0.83 + (1.98 \pm 0.03)\, t$. Residuals are plotted as open circles, and at least visually, show no obvious structure. Note that the fit is correct within its estimated standard errors. The sample variance of the estimated noise was used for calculating the uncertainty, not the theoretical value. {fig3_2.eps}

$\theta(t) = a + bt$, so that the measurements are,

$$y(t_i) = \theta(t_i) + n(t_i) = a + bt_i + n(t_i), \tag{2.87}$$ {33001}

where $n(t)$ is the inevitable measurement noise. The straight-line relationship might as well be referred to as a "model," as it represents our present conception of the data structure. We want to determine a, b.

The set of observations can be written in the general standard form,

{33002} $$\mathbf{Ex} + \mathbf{n} = \mathbf{y} \tag{2.88}$$

where,

{33003} $$\mathbf{E} = \left\{ \begin{matrix} 1 & t_1 \\ 1 & t_2 \\ . & . \\ . & . \\ 1 & t_M \end{matrix} \right\}, \quad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y(t_1) \\ y(t_2) \\ . \\ . \\ y(t_M) \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} n(t_1) \\ n(t_2) \\ . \\ . \\ n(t_M) \end{bmatrix}. \tag{2.89}$$

Equation sets like (2.88) are seen in many practical situations, including the ones described in Chapter 1. The matrix $\mathbf{E}$ in general represents arbitrarily complicated linear relations between the parameters $\mathbf{x}$, and the observations $\mathbf{y}$. In some real cases, it has many thousands of rows and columns. Its construction involves specifying what those relations are, and in a very general sense, it requires a "model" of the data set. Unfortunately, the term "model" is used in a variety of other ways in this context, including statistical assumptions, and often for auxiliary relationships among the elements of $\mathbf{x}$ which are independent of those contained in $\mathbf{E}$. To separate these difference usages, we will sometimes append various adjectives to the use ("statistical model", "exact relationships" etc.).

One sometimes sees (2.88) written as

$$\mathbf{Ex} \sim \mathbf{y}$$

or even

$$\mathbf{Ex} = \mathbf{y}\,.$$

But Eq. (2.88) is preferable, because it explicitly recognizes that $\mathbf{n} = \mathbf{0}$ is exceptional. Sometimes, by happenstance or arrangement, one finds $M = N$ and that $\mathbf{E}$ has an inverse. But the obvious solution, $\mathbf{x} = \mathbf{E}^{-1}\mathbf{y}$, leads to the conclusion, $\mathbf{n} = \mathbf{0}$, which should be unacceptable if the $\mathbf{y}$ are the result of measurements. We will need to return to this case, but for now, let us consider the commonplace problem where $M > N$.

Then, one often sees a "best possible" solution—defined as producing the smallest possible value of $\mathbf{n}^T\mathbf{n}$, that is the minimum of

$$J = \sum_{i=1}^{M} n_i^2 = \mathbf{n}^T\mathbf{n} = (\mathbf{y} - \mathbf{Ex})^T(\mathbf{y} - \mathbf{Ex})\,. \qquad (2.90)$$ {33004}

(Whether the smallest noise solution really is the best one is considered later.) In the special case of the straight-line model,

$$J = \sum_{i=1}^{M} (y_i - a - bt_i)^2\,. \qquad (2.91)$$ {33004a}

J is an example of what is called an "objective" or "cost" function.[30]

Taking the differential of (2.91) with respect to a, b or $\mathbf{x}$ (using (2.32)) and setting it to zero produces,

$$\begin{aligned} dJ &= \sum_i \frac{\partial J}{\partial x_i} dx_i = \left(\frac{\partial J}{\partial \mathbf{x}}\right)^T d\mathbf{x} \\ &= 2\left(\mathbf{E}^T\mathbf{y} - \mathbf{E}^T\mathbf{Ex}\right)^T d\mathbf{x} = 0. \qquad (2.92) \end{aligned}$$

This equation is of the form

$$dJ = \sum a_i dx_i = 0. \qquad (2.93)$$

It is an elementary result of multivariable calculus that an extreme value (here a minimum) of J is found where $dJ = 0$. Because the x_i are free to vary independently, dJ will vanish only if the coefficients of the dx_i are separately zero or,

{normal1} $$\mathbf{E}^T\mathbf{y} - \mathbf{E}^T\mathbf{E}\mathbf{x} = \mathbf{0}. \tag{2.94}$$

That is,

{33005} $$\mathbf{E}^T\mathbf{E}\mathbf{x} = \mathbf{E}^T\mathbf{y}, \tag{2.95}$$

called the "normal equations." Note that Eq. (2.94) asserts that the columns of $\mathbf{E}$ are orthogonal (that is "normal") to $\mathbf{n} = \mathbf{y} - \mathbf{E}\mathbf{x}$. Making the sometimes-valid-assumption that $(\mathbf{E}^T\mathbf{E})^{-1}$ exists,

{33006} $$\tilde{\mathbf{x}} = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}\,. \tag{2.96}$$

By looking at the second derivatives of J with respect to $\mathbf{x}$, we could show what is intuitively clear—that we have a minimum and not a maximum. The solution is written as $\tilde{\mathbf{x}}$ rather than as $\mathbf{x}$ because the relationship between (2.96) and the "correct" value is obscure. Fig. 2.2, displays the fit along with the residuals,

{33007} $$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} = \left[\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\right]\mathbf{y}\,. \tag{2.97}$$

That is, the M equations have been used to estimate N values, $\tilde{\mathbf{x}}_i$, and M values $\tilde{\mathbf{n}}_i$, or $M + N$ altogether. The combination

{H1} $$\mathbf{H} = \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T \tag{2.98}$$

occurs sufficiently often that it is worth a special symbol. Note the "idempotent" property $\mathbf{H}^2 = \mathbf{H}$. If the solution $\tilde{\mathbf{x}}$ is substituted into the original equations, the result is,

$$\mathbf{E}\tilde{\mathbf{x}} = \mathbf{H}\mathbf{y} = \tilde{\mathbf{y}}, \tag{2.99}$$

and

{orthog1} $$\tilde{\mathbf{n}}^T\tilde{\mathbf{y}} = \left[(\mathbf{I} - \mathbf{H})\,\mathbf{y}\right]^T\mathbf{H}\mathbf{y} = \mathbf{0}. \tag{2.100}$$

The residuals are orthogonal (normal) to the inferred noise-free "data" $\tilde{\mathbf{y}}$.

All of this is easy and familiar and applies to any set of simultaneous linear equations, not just the straight-line example. Before proceeding, let us apply some of the statistical machinery to understanding (2.96). Notice that no statistics were used in obtaining (2.96), but we can nonetheless ask the extent to which this value for $\tilde{\mathbf{x}}$ is affected by the random elements: the noise in $\mathbf{y}$. Let $\mathbf{y}_0$ be the value of $\mathbf{y}$ that would be obtained in the hypothetical situation for which $\mathbf{n} = \mathbf{0}$. Assume further that $\langle\mathbf{n}\rangle = \mathbf{0}$ and that $\mathbf{R}_{nn} = \mathbf{C}_{nn} = \langle\mathbf{n}\mathbf{n}^T\rangle$ is known. Then the expected value of $\tilde{\mathbf{x}}$ is,

{33008} $$\langle\tilde{\mathbf{x}}\rangle = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{y}_0\,. \tag{2.101}$$

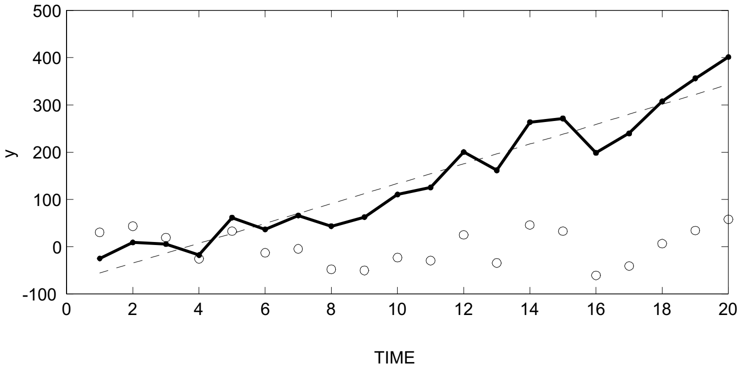

Figure 2.3: Here the "data" were generated from a quadratic rule, $y = 1 + t^2 + n(t)$, $\langle n^2 \rangle = 900$. Note that only the first $1 \le t \le 20$ data points are used. An incorrect straight line fit was used resulting in $\tilde{y} = (-76.3 \pm 17.3) + (20.98 \pm 1.4)\,t$, which is incorrect, but the residuals at least visually, do not appear unacceptable. At this point some might be inclined to claim the model has been "verified," or "validated." {fig3_4.eps}

If the matrix inverse exists, then in many situations, including the problem of fitting a straight-line to data, perfect observations would produce the correct answer, and Eq. (2.96) provides an unbiased estimate of the true solution, $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$. A more transparent demonstration of this result will be given on P. 105. eunbiassed1}

On the other hand, if the data were actually produced from physics governed for example, by a quadratic rule, $\theta(t) = a + ct^2$, then fitting the linear rule to such observations, even if they are perfect, could never produce the right answer and the solution would be biassed. An example of such a fit is shown in figures 2.3, 2.4. Such errors are conceptually distinguishable from the noise of observation, and are properly labeled "model errors."

Assume however, that the correct model is being used, and therefore that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$. Then the uncertainty of the solution is,

$$\begin{aligned}\mathbf{P} = \mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} &= \langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T \rangle \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T \langle \mathbf{n}\mathbf{n}^T \rangle\, \mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1} \\ &= (\mathbf{E}^T\mathbf{E})^{-1}\,\mathbf{E}^T\,\mathbf{R}_{nn}\,\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\,.\end{aligned} \tag{2.102}$$ {33009}

In the special case, $\mathbf{R} = \sigma_n^2\mathbf{I}$, that is, no correlation between the noise in different equations (white noise), Eq. (2.102) simplifies to,

$$\mathbf{P} = \sigma_n^2(\mathbf{E}^T\mathbf{E})^{-1}\,. \tag{2.103}$$ {33010}

If we are not confident that $\langle \tilde{\mathbf{x}} \rangle = \mathbf{x}$, perhaps because of doubts about the straight-line model,

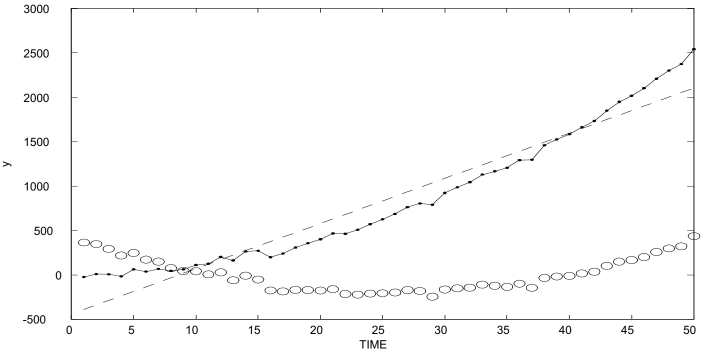

Figure 2.4: The same situation as in Fig. 2.3, except the series was extended to 50 points. Now the residuals ('o') are visually structured, and one would have a powerful suggestion that some hypothesis (something about the model or data) is not correct. This straightline fit should be rejected as being inconsistent with the assumption that the residuals ar unstructured: the model has been "invalidated." {fig3_5.eps}

Eqs. (2.102)–(2.103) are still interpretable, but as $C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = D^2(\tilde{\mathbf{x}} - \langle\tilde{\mathbf{x}}\rangle)-$ the covariance of $\tilde{\mathbf{x}}$. The "standard error" of $\tilde{x}_i$ is usually defined to be $\pm\sqrt{C_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}ii}}$ and is used to understand the adequacy of data for distinguishing different possible estimates of $\tilde{\mathbf{x}}$. If applied to the straight-line fit of fig. 2.2, we obtain an estimate, $\tilde{\mathbf{x}}^T = [\tilde{a}, \tilde{b}] = [1.69 \pm 0.83, 1.98 \pm 0.03]$, which are within one standard deviation of the true values, $[a, b] = [1, 2]$. If the noise in $\mathbf{y}$ is Gaussian, it follows that the probability density of $\tilde{\mathbf{x}}$ is also Gaussian, with mean $\langle\tilde{\mathbf{x}}\rangle$ and covariance $\mathbf{C}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$. Of course, if $\mathbf{n}$ is not Gaussian, then the estimate won't be either, and one must be wary of the utility of the standard errors. A Gaussian, or other, assumption should be regarded as part of the model definition. The uncertainty of the residuals as,

{cnn1}
$$\begin{aligned}\mathbf{C}_{nn} &= \left\langle (\tilde{\mathbf{n}} - \langle\tilde{\mathbf{n}}\rangle)(\tilde{\mathbf{n}} - \langle\tilde{\mathbf{n}}\rangle)^T \right\rangle = (\mathbf{I} - \mathbf{H})\,\mathbf{R}_{nn}\,(\mathbf{I} - \mathbf{H})^T \\ &= \sigma_n^2\,(\mathbf{I} - \mathbf{H})^2 = \sigma_n^2\,(\mathbf{I} - \mathbf{H})\,,\end{aligned} \tag{2.104}$$

where zero-mean white noise was assumed, and $\mathbf{H}$ was defined in Eq. (2.98). Notice that the true noise, $\mathbf{n}$, was assumed to be white, but that the estimated noise, $\tilde{\mathbf{n}}$, has a non-diagonal covariance and so in a formal sense does not have the expected covariance. We return to this point below.

The fit of a straight-line to observations demonstrates many of the issues involved in making inferences from real, noisy data that appear in more complex situations. In figure 2.5, the correct model used to generate the data was the same as in Fig. 2.2, but the noise level is very high. The parameters $[\tilde{a}, \tilde{b}]$ are numerically inexact, but consistent within one standard error with the

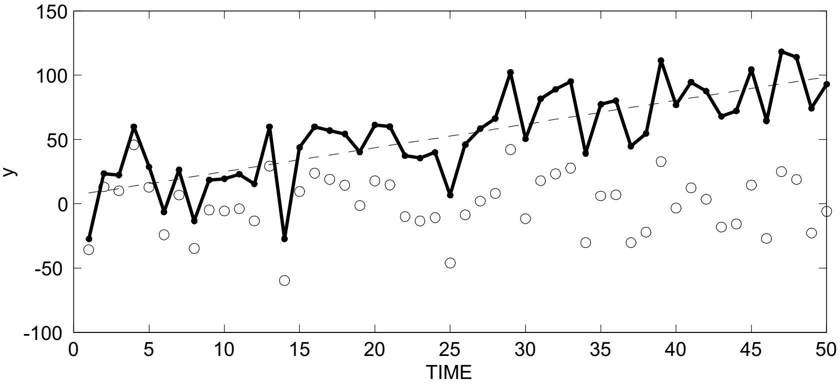

Figure 2.5: The same situation as in Fig. 2.2, $y = 1 + 2t$, except $\langle n^2 \rangle = 900$ to give very noisy data. Now the best fitting straight line is $y = (6.62 \pm 6.50) + (1.85 \pm 0.22)\, t$ which includes the correct answer within one standard error. Note that the intercept value is indistinguishable from zero.

{fig3_3.eps}

correct values, which is all one can hope for.

In figure 2.3, a quadratic model $y = 1 + t^2 + n\,(t)$ was used to generate the numbers, with $\langle n^2 \rangle = 900$. Using only the first 20 points, and fitting an incorrect model produces a reasonable straight-line fit to the data as shown. Modeling a quadratic field with a linear model produces a systematic or "model" error, which is not easy to detect here. One sometimes hears it said that "least-squares failed" in situations such as this one. But this conclusion shows a fundamental misunderstanding: least-squares did exactly what it was asked to do—to produce the best-fitting straight-line to the data. Here, one might conclude that "the straight-line fit *is* consistent with the data." Such a conclusion is completely different from asserting that one has proven a straight-line fit correctly "explains" the data or, in modeler's jargon, that the model has been "verified" or "validated." If the outcome of the fit were sufficiently important, one might try more powerful tests on the $\tilde{n}_i$ than a mere visual comparison. Such tests might lead to rejection of the straight-line hypothesis; but even if the tests are passed, the model has *never* been verified: it has only been shown to be consistent with the available data.

If the situation remains unsatisfactory (perhaps one suspects the model is inadequate, but there are not enough data to produce sufficiently powerful tests), it can be very frustrating. But sometimes the only remedy is to obtain more data. So in Fig. 2.4, the number of observations was extended to 50 points. Now, even visually, the $\tilde{n}_i$ are obviously structured, and one would almost surely reject any hypothesis that a straight-line was an adequate representation of the data. *The model has been invalidated.* If one fits a quadratic rule, $y = a + bt + ct^2$, a perfectly

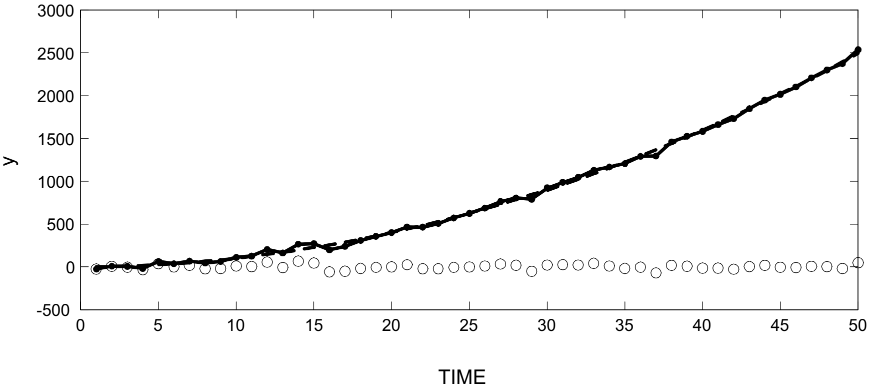

Figure 2.6: Same as Fig. 2.4, except a more complete model, $y = a + bt + ct^2$ was used, and which gives acceptable residuals. {fig3_6.eps}

acceptable solution is found; see Fig. 2.6.

One must always confirm, after the fact, that J, which is a direct function of the residuals, behaves as expected when the solution is substituted. In particular, its expected value,

{expJ1}
$$\langle J\rangle = \sum_{i}^{M} \left\langle n_i^2 \right\rangle = M - N, \tag{2.105}$$

assuming that the n_i have been scaled so that each has an expected value $\left\langle n_i^2 \right\rangle = 1$. That there are only $M-N$ independent terms in (2.105) follows from the N supposed-independent constraints linking the variables. For any particular solution, $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$, J will be a random variable, whose expectation is (2.105). Assuming the n_i are at least approximately Gaussian, J itself is the sum of $M-N$ independent χ_1^2 variables, and is therefore distributed in χ_{M-N}^2. One can and should make histograms of the individual n_i^2 to check them against the expected χ_1^2 probability density. This type of argument leads to the large literature on hypothesis testing.

As an illustration of the random behavior of residuals, 30 equations, $\mathbf{Ex}+\mathbf{n}=\mathbf{y}$ in 15 unknowns were constructed, such that $\mathbf{E}^T\mathbf{E}$ was non-singular. Fifty different values of $\mathbf{y}$ were then constructed by generating 50 separate $\mathbf{n}$ using a pseudo-random number generator. An ensemble of 50 different solutions were calculated using (2.96), producing $50\times 30 = 1500$ separate values of $\tilde{n}_i^2$. These are plotted in Fig. 2.7 and compared to χ_1^2. The corresponding value, $\tilde{J}^{(p)} = \sum_1^{30} \tilde{n}_i^2$, was found for each set of equations, and also plotted. A corresponding frequency function for $\tilde{J}^{(p)}$ is compared in Fig. 2.7 to χ_{15}^2, with reasonably good results. The empirical mean value of all $\tilde{J}_i$ is 14.3. Any particular solution may, completely correctly, produce individual residuals $\tilde{n}_i^2$ differing considerably from the mean of $\left\langle \chi_1^2 \right\rangle = 1$, and similarly, their sums, the

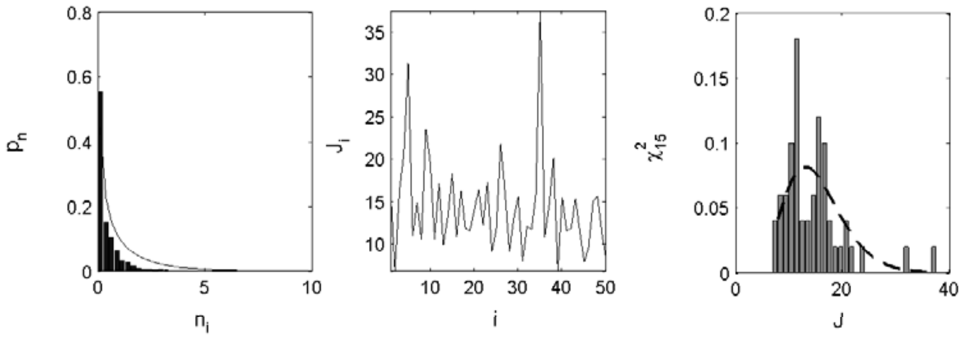

Figure 2.7: χ_1^2 probability density (left panel), and the empirical frequency function of *all* residuals, $\tilde{n}_i^2$ from 50 separate experiments for simple least-squares solution of $\mathbf{Ex}+\mathbf{n}=\mathbf{y}$. There is at least rough agreement between the theoretical and calculated frequency functions. Middle panel displays the 50 values of J_i computed from the same experiments in the left panel. Right panel displays the empirical frequency function for the J_i as compared to the theoretical value of χ_{15}^2, (dashed line). Tests exist, not discussed here, of the hypothesis that the calculated J_i are consistent with the theoretical distribution. {fig3_8.eps}

$J^{(p)}$ may differ greatly from $\left\langle \chi_{15}^2 \right\rangle = 15$. But one can readily calculate the probability of finding a much larger or smaller value, and employ it to help evaluate the possibility that one has used an incorrect model.

Visual tests for randomness of residuals have obvious limitations, and elaborate statistical tests in addition to the comparison with χ^2 exist to help determine objectively whether one should accept or reject the hypothesis that no significant structure remains in a sequence of numbers. Books on regression analysis[31] should be consulted for general methodologies. As an indication of what can be done, figure 2.8 shows the “sample autocorrelation,”

$$\tilde{\phi}_{nn}(\tau)=\frac{\frac{1}{M}\sum_{i=1}^{M-|\tau|}\tilde{n}_i\tilde{n}_{i+\tau}}{\frac{1}{M}\sum_{i=1}^{M}\tilde{n}_i^2}\,, \tag{2.106}$$ {autocorr1}

for the residuals of the fits shown in figs. 2.4, 2.6 is displayed. For white noise,

$$\left\langle \tilde{\phi}\left(\tau\right)\right\rangle=\delta_{0\tau}, \tag{2.107}$$ {whitecovar}

and deviations of the estimated $\tilde{\phi}\left(t\right)$ from Eq. (2.107) can be used in simple tests. The adequate fit (Fig. 2.6) produces an autocorrelation of the residuals indistinguishable from a delta function at the origin, while the inadequate fit, shows a great deal of structure which would lead to the

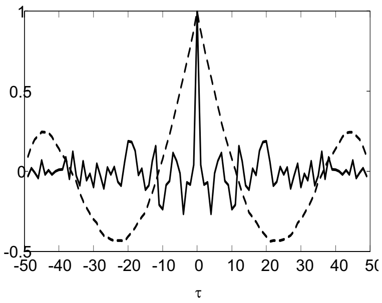

Figure 2.8: Autocorrelations of the estimated residuals in Figs. 2.4 (dashed line), and 2.6 (solid). The latter is indistinguishable, by statistical test, from a delta function at the origin, and so with this test, the residuals are not distinguishable from white noise. {fig3_7.eps}

conclusion that the residuals are too different from white noise to be acceptable. (Not all cases are this obvious.).

As already pointed out, the residuals of the least-squares fit cannot be expected to be precisely white noise. Because there are M-relationships among the parameters of the problem (M-equations), and the number of $\tilde{\mathbf{x}}$ elements determined is N, there are $M-N$-degrees of freedom in the determination of $\tilde{\mathbf{n}}$ and structures are imposed upon them. The failure, for this reason, of $\tilde{\mathbf{n}}$ strictly to be white noise, is generally only an issue in practice when $M-N$ becomes small compared to M.[32]

2.4.2 Weighted and Tapered Least-Squares

The least-squares solution (2.96)–(2.97) was derived by minimizing the objective function (2.90), in which each residual element is given equal weight. An important feature of least-squares is that we can give whatever emphasis we please to minimizing individual equation residuals, for example, by introducing an objective function,

{33011}
$$J=\sum_i W_{ii}^{-1}n_i^2, \tag{2.108}$$

where W_{ii} are any numbers desired. The choice $W_{ii}=1$, as used above, might be reasonable, but it is clearly an arbitrary one which without further justification does not produce a solution with any special claim to significance. In the least-squares context, we are free to make any other reasonable choice, including demanding that some residuals should be much smaller than others—perhaps just to see if it is possible.

A general formalism is obtained by defining a diagonal weight matrix, $W=\mathrm{diag}(W_{ii})$. Divide each equation by $\sqrt{W_{ii}}$,

{33012} $$W_{ii}^{-T/2} \sum_i E_{ij} x_j + W_{ii}^{-T/2} n_i = W_{ii}^{-T/2} y_i, \tag{2.109}$$

or

$$\begin{gathered} \mathbf{E}'\mathbf{x} + \mathbf{n}' = \mathbf{y}' \\ \mathbf{E}' = \mathbf{W}^{-T/2}\mathbf{E}, \quad \mathbf{n}' = \mathbf{W}^{-T/2}\mathbf{n}, \quad \mathbf{y}' = \mathbf{W}^{-T/2}\mathbf{y} \end{gathered} \tag{2.110}$$ {33013}

where we used the fact that the square root of a diagonal matrix is the diagonal matrix of element-by-element square roots. Such a matrix is its own transpose. The operation in (2.109) or (2.110) is usually called "row-scaling" because it operates on the rows of $\mathbf{E}$ (as well as on $\mathbf{n}$, $\mathbf{y}$).

For the new equations (2.110), the objective function,

$$\begin{aligned} J &= \mathbf{n}'^T\mathbf{n}' = (\mathbf{y}' - \mathbf{E}'\mathbf{x})^T(\mathbf{y}' - \mathbf{E}'\mathbf{x}) \\ &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} = (\mathbf{y} - \mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{E}\mathbf{x}), \end{aligned} \tag{2.111}$$

weights the residuals as desired. If, for some reason, $\mathbf{W}$ is non-diagonal, but symmetric and positive-definite, then it has a Cholesky decomposition, (see P. 38) and,

$$\mathbf{W} = \mathbf{W}^{T/2}\mathbf{W}^{1/2},$$

and (2.110) remains valid more generally.

The values $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, minimizing (2.111) are,

$$\begin{aligned} \tilde{\mathbf{x}} &= (\mathbf{E}'^T\mathbf{E}')^{-1}\mathbf{E}'^T\mathbf{y}' = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y}, \\ \tilde{\mathbf{n}} &= \mathbf{W}^{T/2}\mathbf{n}' = \left[\mathbf{I} - \mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\right]\mathbf{y}, \end{aligned} \tag{2.112}$$ {33015}

$$\mathbf{C}_{xx} = (\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E})^{-1}. \tag{2.113}$$ {33016}

Uniform diagonal weights are clearly a special case. The rationale for choosing differing diagonal weights, or a non-diagonal $\mathbf{W}$, is probably not very obvious to the reader. Often one chooses $\mathbf{W} = \mathbf{R}_{nn} = \{\langle n_i n_j \rangle\}$, that is, the weight matrix is chosen to be the expected second moment matrix of the residuals. Then

$$\left\langle \mathbf{n}'\mathbf{n}'^T \right\rangle = \mathbf{I},$$

and Eq. (2.113) simplifies to

$$\mathbf{C}_{xx} = (\mathbf{E}^T \mathbf{R}_{nn}^{-1} \mathbf{E})^{-1} . \tag{2.114}$$ {33017}

In this special case, the weighting (2.110) has a ready interpretation: The equations (and hence the residuals) are rotated and stretched so that in the new coordinate system of n_i', the covariances are all diagonal and the variances are all unity. Under these circumstances, an objective function

$$J = \sum_i n_i'^2$$

as used in the original form of least-squares (Eq. (2.90)) is a reasonable choice.

Consider the system

$$\begin{aligned} x_1 + x_2 + n_1 &= 1 \\ x_1 - x_2 + n_2 &= 2 \\ x_1 - 2x_2 + n_3 &= 4. \end{aligned}$$

Then if $\langle n_i \rangle = 0$, $\left\langle n_i^2 \right\rangle = \sigma^2$, the least-squares solution is $\tilde{\mathbf{x}} = [2.0, 0.5]^T$. Now suppose that

$$\langle n_i n_j \rangle = \left\{ \begin{matrix} 1 & 0.99 & 0.98 \\ 0.99 & 1 & 0.99 \\ 0.98 & 0.99 & 4 \end{matrix} \right\} .$$

Then from Eq. (2.112), $\tilde{\mathbf{x}} = [1.51, -0.48]^T$. Calculation of the two different solution uncertainties is left to the reader.

But we emphasize that this choice of $\mathbf{W}$ is a very special one and has confused many users of inverse methods. To emphasize again: Least-squares is an approximation procedure in which $\mathbf{W}$ is a set of weights wholly at the disposal of the investigator; setting $\mathbf{W} = \mathbf{R}_{nn}$ is a special case whose significance is best understood after we examine a different, statistical, estimation procedure.

Whether the equations are scaled or not, the previous limitations of the simple least-squares solutions remain. In particular, we still have the problem that the solution may produce elements in $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, whose relative values are not in accord with expected or reasonable behavior and the solution uncertainty or variances could be unusably large, as the solution is determined, mechanically, and automatically, from combinations such as $(\mathbf{E}^T \mathbf{W}^{-1} \mathbf{E})^{-1}$. Operators like these are neither controllable nor very easy to understand; if any of the the matrices is singular, they will not even exist.

It was long ago recognized that some control over the magnitudes of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, $\mathbf{C}_{xx}$ could be obtained in the simple least-squares context by modifying the objective function (2.108) to have an additional term:

$$\begin{aligned} J' &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n} + \gamma^2\mathbf{x}^T\mathbf{x} && (2.115) \\ &= (\mathbf{y}-\mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y}-\mathbf{E}\mathbf{x}) + \gamma^2\mathbf{x}^T\mathbf{x}, && (2.116) \end{aligned}$$

in which γ^2 is a positive constant.

If the minimum of (2.115) is sought by setting the derivatives with respect to $\mathbf{x}$ to zero, we obtain,

$$\tilde{\mathbf{x}} = \left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y} \qquad (2.117)$$

$$\tilde{\mathbf{n}} = \mathbf{y} - \mathbf{E}\tilde{\mathbf{x}} \qquad (2.118)$$

$$\begin{aligned} &\mathbf{C}_{xx} = && (2.119) \\ &\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}. \end{aligned}$$

By letting $\gamma^2 \to 0$, the solution 2.112, 2.113 is recovered, and if $\gamma^2 \to \infty$, $\|\tilde{\mathbf{x}}\|_2 \to 0$, $\tilde{\mathbf{n}} \to \mathbf{y}$; γ^2 is called a "trade-off parameter," because it trades the magnitude of $\tilde{\mathbf{x}}$ against that of $\tilde{\mathbf{n}}$. By varying the size of γ^2 we gain some influence over the norm of the residuals relative to that of $\tilde{\mathbf{x}}$. The expected value of $\tilde{\mathbf{x}}$ is now,

$$\langle\tilde{\mathbf{x}}\rangle = \left[\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\gamma^2\mathbf{I}\right]^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y}_0\,. \qquad (2.120)$$ {33020}

If the true solution is believed to be (2.101), then this new solution is biassed. But the variance of $\tilde{\mathbf{x}}$ has been reduced, (2.119), by introduction of $\gamma^2 > 0$—that is, the acceptance of a bias reduces the variance, possibly very greatly. Eqs. (2.117-2.118) are sometimes known as the "tapered least-squares" solution, a label whose implication becomes clear later. $\mathbf{C}_{nn}$, which is not displayed, is readily found by direct computation as in Eq. (2.104).

The most basic, and commonly seen form of this solution assumes $\mathbf{W} = \mathbf{R}_{nn} = \mathbf{I}$, and then,

$$\begin{aligned} \tilde{\mathbf{x}} &= \left(\mathbf{E}^T\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{y} && (2.121) \\ \mathbf{C}_{xx} &= \left(\mathbf{E}^T\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}\mathbf{E}^T\mathbf{E}\left(\mathbf{E}^T\mathbf{E}+\gamma^2\mathbf{I}\right)^{-1}, && (2.122) \end{aligned}$$

a special case.

A physical motivation for the modified objective function (2.115) is obtained by noticing that a preference for a bounded $\|\mathbf{x}\|$ is easily produced by adding an equation set, $\mathbf{x}+\mathbf{n}_1=\mathbf{0}$, so that the combined set is,

{combined1}
$$\mathbf{E}\mathbf{x}+\mathbf{n}=\mathbf{y} \tag{2.123}$$

{combined2}
$$\mathbf{x}+\mathbf{n}_1=\mathbf{0} \tag{2.124}$$

or

$$\mathbf{E}_1\mathbf{x}+\mathbf{n}_2=\mathbf{y}_2$$

{33021}
$$\mathbf{E}_1=\begin{Bmatrix}\mathbf{E}\\ \\ \gamma^2\mathbf{I}\end{Bmatrix}, \quad \mathbf{n}_2^T=\left[\mathbf{n}^T\ \ \gamma^2\mathbf{n}_1^T\right], \quad \mathbf{y}_2^T=\left[\mathbf{y}^T\ \ \mathbf{0}^T\right], \tag{2.125}$$

and in which γ^2 expresses a preference for fitting the first or second sets more closely. Then J in Eq. (2.115) becomes the natural objective function to use. A preference that $\mathbf{x}\approx\mathbf{x}_0$ is readily imposed instead, with an obvious change in (2.115) or (2.124).

Note the important points, to be shown later, that the matrix inverses in Eqs. (2.117-2.118) will *always* exist, as long as $\gamma^2>0$, and that the expressions remain valid even if $M<N$. Tapered least-squares produces some control over the sum of squares of the relative norms of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, but still does not produce control over the individual elements $\tilde{x}_i$.

To gain some of that control, we can further generalize the objective function by introducing another non-singular $N\times N$ weight matrix, $\mathbf{S}$ (which is usually symmetric) and,

$$\begin{aligned} J &= \mathbf{n}^T\mathbf{W}^{-1}\mathbf{n}+\mathbf{x}^T\mathbf{S}^{-1}\mathbf{x} && (2.126)\\ &= (\mathbf{y}-\mathbf{E}\mathbf{x})^T\mathbf{W}^{-1}(\mathbf{y}-\mathbf{E}\mathbf{x})+\mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}\,, && (2.127)\end{aligned}$$

for which Eq. (2.115) is a special case. Setting the derivatives with respect to $\mathbf{x}$ to zero results in,

$$\tilde{\mathbf{x}}=\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\mathbf{S}^{-1}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{y}, \tag{2.128}$$

$$\tilde{\mathbf{n}}=\mathbf{y}-\mathbf{E}\tilde{\mathbf{x}}, \tag{2.129}$$

$$\mathbf{C}_{xx}= \tag{2.130}$$
$$\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\mathbf{S}^{-1}\right)^{-1}\mathbf{E}^T\mathbf{W}^{-1}\mathbf{R}_{nn}\mathbf{W}^{-1}\mathbf{E}\left(\mathbf{E}^T\mathbf{W}^{-1}\mathbf{E}+\mathbf{S}^{-1}\right)^{-1},$$

and Eqs. (2.117-2.119) are a special case, with $\mathbf{S}^{-1}=\gamma^2\mathbf{I}$. $\mathbf{C}_{xx}$ simplifies if $\mathbf{R}_{nn}$= $\mathbf{W}$.

Suppose $\mathbf{S}$, $\mathbf{W}$ are positive definite and symmetric and thus have Cholesky decompositions. Then we can employ both matrices directly on the equations, $\mathbf{Ex}+\mathbf{n}=\mathbf{y}$,

$$\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{-T/2}\mathbf{S}^{T/2}\mathbf{x}+\mathbf{W}^{-T/2}\mathbf{n}{=}\mathbf{W}^{-T/2}\mathbf{y} \tag{2.131}$$

$$\mathbf{E}'\mathbf{x}'{+}\mathbf{n}'{=}\mathbf{y}' \tag{2.132}$$

$$\mathbf{E}'{=}\mathbf{W}^{-T/2}\mathbf{E}\mathbf{S}^{T/2},\ \mathbf{x}'=\mathbf{S}^{-T/2}\mathbf{x},\ \mathbf{n}'=\mathbf{W}^{-T/2}\mathbf{n},\ \mathbf{y}'{=}\ \mathbf{W}^{-T/2}\mathbf{y} \tag{2.133}$$

The use of $\mathbf{S}$ in this way is “column scaling” because it weights the columns of $\mathbf{E}$. With Eqs. (2.132) the obvious objective function is,

$$J=\mathbf{n}'^T\mathbf{n}'+\mathbf{x}'^T\mathbf{x}', \tag{2.134}$$ {33027}

which is identical to Eq. (2.126) in the original variables, and the solution must be that in Eqs. (2.128-2.130).

Like $\mathbf{W}$, one is completely free to choose $\mathbf{S}$ as one pleases. A common example is to write, where $\mathbf{F}$ is $N\times N$,

$$\begin{aligned}\mathbf{S}&=\mathbf{F}^T\mathbf{F}\\ \mathbf{F}&=\gamma^2\left\{\begin{matrix}1 & -1 & 0 & \cdots & 0\\ 0 & 1 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & \cdots & 0 & 1\end{matrix}\right\},\end{aligned} \tag{2.135}$$ {33025}

whose effect is to minimize a term $\gamma^2\sum_i(x_i-x_{i+1})^2$, which can be regarded as a “smoothest” solution, and using γ^2 to trade smoothness against the size of $\|\tilde{\mathbf{n}}\|_2$, $\alpha\mathbf{F}$ is obtained from the Cholesky decomposition of $\mathbf{S}$.

By invoking the matrix inversion lemma, an alternative form for Eqs. $(2.128-2.130)$ is found,

$$\tilde{\mathbf{x}}=\mathbf{S}\mathbf{E}^T\left(\mathbf{E}\mathbf{S}\mathbf{E}^T{+}\mathbf{W}\right)^{-1}\mathbf{y}, \tag{2.136}$$

$$\tilde{\mathbf{n}}=\mathbf{y}-\mathbf{E}\tilde{\mathbf{x}}, \tag{2.137}$$

$$\mathbf{C}_{xx}=\mathbf{S}\mathbf{E}^T\left(\mathbf{E}\mathbf{S}\mathbf{E}^T{+}\mathbf{W}\right)^{-1}\mathbf{R}_{nn}\left(\mathbf{E}\mathbf{S}\mathbf{E}^T{+}\mathbf{W}\right)^{-1}\mathbf{E}\mathbf{S}. \tag{2.138}$$

A choice of which form to use is sometimes made on the basis of the dimensions of the matrices being inverted. Note again that $\mathbf{W} = \mathbf{R}_{nn}$ is a special case.

So far, all of this is conventional. But we have made a special point of displaying explicitly not only the elements $\tilde{\mathbf{x}}$, but those of the residuals, $\tilde{\mathbf{n}}$. Notice that although we have considered only the formally over-determined system, $M > N$, we *always* determine not only the $N-$elements of $\tilde{\mathbf{x}}$, but also the M-elements of $\tilde{\mathbf{n}}$, for a total of $M + N$ values—extracted from the M-equations. It is apparent that any change in any element $\tilde{n}_i$ forces changes in $\tilde{\mathbf{x}}$. In this view, to which we adhere, systems of equations involving observations *always* contain more unknowns than equations. Another way to make the point is to re-write Eqs. (2.88) without distinction between $\mathbf{x}, \mathbf{n}$ as,

$$\mathbf{E}_1\boldsymbol{\xi}= \mathbf{y}, \tag{2.139}$$

$$\mathbf{E}_1 = \{\mathbf{E}, \mathbf{I}_M\}, \ \boldsymbol{\xi}^T= [\mathbf{x}, \mathbf{n}]^T. \tag{2.140}$$

A combined weight matrix,

$$\mathbf{S}_1 = \left\{ \begin{matrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{matrix} \right\}, \tag{2.141}$$

would be used, and any distinction between the $\mathbf{x}, \mathbf{n}$ solution elements is suppressed. Eqs. (2.139) are a formally underdetermined system, derived from the formally over-determined observed one. This identity leads us to the problem of formal underdetermination in the next Section.

In general with least-squares problems, the solution we seek can be regarded as any of the following equivalents:

1. The $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ satisfying

$$\mathbf{Ex} + \mathbf{n} = \mathbf{y}. \tag{2.142}$$

2. $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ satisfying the normal equations arising from J (Eq. 2.126).

3. $\tilde{\mathbf{x}}, \tilde{\mathbf{n}}$ producing the minimum of J in Eq. (2.126)

The point of this list lies with item 3: algorithms exist to find minima of functions by deterministic methods ("go downhill" from an initial guess)[33], or stochastic search methods (Monte Carlo) or even, conceivably, through a shrewd guess by the investigator. If an acceptable minimum of J is found, by whatever means, it is an acceptable solution (subject to further testing, and the possibility that there is more than one such solution). Search methods become essential for the nonlinear problems taken up later.

2.4.3 Underdetermined Systems and Lagrange Multipliers

What does one do when the number, M, of equations is less than the number, N, of unknowns and no more observations are possible? We have seen that the claim that a problem involving observations is ever overdetermined is misleading—because each equation or observation always has a noise unknown, but to motivate some of what follows, it is helpful to first pursue a conventional approach.

One often attempts when $M < N$ to reduce the number of unknowns so that the formal overdeterminism is restored. Such a parameter reduction procedure may be sensible; but there are pitfalls. Let $p_i\left(t\right)$, $0 \leq i$ be a set of polynomials, e.g. Chebyschev or Laguerre, etc. Consider data produced from the formula,

$$y\left(t\right) = 1 + a_M p_M\left(t\right) + n(t)\,, \qquad (2.143)$$ {33030}

which might be deduced by fitting a parameter set $[a_0, \ldots, a_M]$ and finding $\tilde{a}_M$. If there are fewer than M observations, an attempt to fit with fewer parameters,

$$y = \sum_{j=0}^{Q} a_j p_j\left(t\right), \quad Q < M \qquad (2.144)$$ {33031}

may give a good, even perfect fit; but it would be incorrect. The reduction in model parameters in such a case biases the result, perhaps hopelessly so. One is better off retaining the underdetermined system and making inferences concerning the possible values of a_i rather than using the form (2.144), in which any possibility of learning something about a_M has been eliminated.

Example *Consider a tracer problem, not unlike those encountered in medicine, hydrology, oceanography, etc. A box (Fig. 1.2) is observed to contain a steady tracer concentration C_0, and is believed fed at the rates J_1, J_2 from two reservoirs each with tracer concentration of C_1, C_2 respectively. One seeks to determine J_1, J_2. Tracer balance is,*

$$J_1C_1 + J_2C_2 - J_0C_0,$$

where J_0 is rate at which fluid is removed. Mass balance then requires

$$J_1 + J_2 = J_0.$$

Evidently, there are but two equations in three unknowns (and a perfectly good solution would be $J_1 = J_2 = J_3 = 0$); but as many have noticed, we can nonetheless, determine the relative fraction of the fluid coming from each reservoir. Divide both equations through by J_0,

$$\frac{J_1}{J_0}C_1 + \frac{J_2}{J_0}C_2 = C_0$$

$$\frac{J_1}{J_0} + \frac{J_2}{J_0} = 1$$

producing two equations in two unknowns,, J_1/J_0, J_2/J_0, which has a unique stable solution (noise is being ignored). Many examples can be given of such calculations in the literature—determining the flux ratios—apparently definitively. But suppose the investigator is suspicious :reservoir1} *that there might be a third reservoir with tracer concentration C_3. Then the equations become*

$$\frac{J_1}{J_0}C_1 + \frac{J_2}{J_0}C_2 + \frac{J_3}{J_0}C_3 = C_0$$

$$\frac{J_1}{J_0} + \frac{J_2}{J_0} + \frac{J_3}{J_0} = 1,$$

now underdetermined with two equations in three unknowns. If it is obvious that no such third reservoir exists, then the reduction to two equations in two unknowns is the right thing to do. But if there is even a suspicion of a third (or more) reservoir, one should solve these equations with one of the methods we will develop—permitting construction and understanding of all possible solutions.

In general terms, parameter reduction can lead to model errors, that is, bias errors, which can produce wholly illusory results.[34] A common situation particularly in problems involving tracer movements in groundwater, ocean, or atmosphere, fitting a one or two-dimensional model to data which represent a fully three-dimensional field. The result may be apparently pleasing, but possibly completely erroneous. (See Chapter 4.)

A general approach to solving underdetermined problems is to render the answer apparently unique by minimizing an objective function, subject to satisfaction of the linear constraints. To see how this can work, suppose that $\mathbf{Ax} = \mathbf{b}$, exactly and formally underdetermined, $M < N$, and seek the solution which exactly satisfies the equations and simultaneously renders an objective function, $J = \mathbf{x}^T\mathbf{x}$, as small as possible. Direct minimization of J leads to,

{33033} $$dJ = \frac{\partial J^T}{\partial \mathbf{x}} d\mathbf{x} = 2\mathbf{x}^T d\mathbf{x} = 0, \tag{2.145}$$

but unlike the case in Eq. (2.92), the coefficients of the individual dx_i can no longer be separately set to zero (i.e., $\mathbf{x} = 0$ is an incorrect solution) because the dx_i no longer vary independently, but are restricted to values satisfying $\mathbf{Ax} = \mathbf{b}$. One approach is to use the known dependencies to reduce the problem to a new one in which the differentials are independent. For example, suppose that there are general functional relationships

$$\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} = \begin{bmatrix} \xi_1(x_{M+1}, \ldots, x_N) \\ \vdots \\ \xi_M(x_{M+1}, \ldots, x_N) \end{bmatrix}.$$

Then the first M elements of x_i may be eliminated, and the objective function becomes,

$$J = \left[\xi_1(x_{M+1}, \ldots, x_N)^2 + \cdots + \xi_M(x_{M+1}, \ldots, x_N)^2\right] + \left[x_{M+1}^2 + \cdots + x_N^2\right],$$

in which the remaining x_i, $M+1 \le i \le N$ are independently varying. In the present case, one can choose (arbitrarily) the first M unknowns, $\mathbf{q} = [x_i]$, and define the last $N-M$ unknowns $\mathbf{r} = [x_i]$, $N-M+1 \le i \le N$, and rewrite the equations as

$$\left\{ \mathbf{A}_1 \;\; \mathbf{A}_2 \right\} \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \end{bmatrix} = \mathbf{b} \tag{2.146}$$

where $\mathbf{A}_1$ is $M \times M$, $\mathbf{A}_2$ is $M \times (N-M)$. Then solving the first set for $\mathbf{q}$,

$$\mathbf{q} = \mathbf{b} - \mathbf{A}_2\mathbf{r} \tag{2.147}$$

$\mathbf{q}$ can be eliminated from J leaving and unconstrained minimzation problem in the independent variables, $\mathbf{r}$. If $\mathbf{A}_1^{-1}$ does not exist, one can try any other subset of the x_i to eliminate until a suitable group is found. This approach is completely correct, but finding an explicit solution for L elements of $\mathbf{x}$ in terms of the remaining ones may be difficult or inconvenient.

Example Solve

$$x_1 - x_2 + x_3 = 1,$$

for the solution of minimum norm. The objective function is $J = x_1^2 + x_2^2 + x_3^2$. *With one equation, one variable can be eliminated. Choosing, arbitrarily,* $x_1 = 1 + x_2 - x_3$, $J = (1 + x_2 - x_3)^2 + x_2^2 + x_3^2$. x_2, x_3 *are now independent variables, and the corresponding derivatives of* J *can be independently set to zero.*

Example

A somewhat more interesting example involves two equations in three unknowns:

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1 - x_2 + x_3 &= 2, \end{aligned}$$

and we choose to find a solution minimizing,

$$J = x_1^2 + x_2^2 + x_3^2.$$

Solving for two unknowns x_1, x_2 *from*

$$\begin{aligned} x_1 + x_2 &= 1 - x_3, \\ x_1 - x_2 &= 2 - x_3, \end{aligned}$$

produces $x_2 = -1/2, x_1 = 3/2 - x_3$ *and then,*

$$J = (3/2 - x_3)^2 + 1/4 + x_3^2.$$

whose minimum with respect to x_3 *(the only remaining variable) is ,* $x_3 = \frac{3}{4}$*, and the full solution is*

$$x_1 = \frac{3}{4}, x_2 = -\frac{1}{2}, x_3 = \frac{3}{4}.$$

Lagrange Multipliers and Adjoints

When it is inconvenient to find such an explicit representation by eliminating some variables in favor of others, a standard procedure for finding the constrained minimum is to introduce a new vector "Lagrange multiplier," $\boldsymbol{\mu}$, of M-unknown elements, to make a new objective function,

$$\begin{aligned} J' &= J - 2\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{b}) \qquad (2.148) \\ &= \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{b}), \end{aligned}$$

and ask for its stationary point—treating both $\boldsymbol{\mu}$ and $\mathbf{x}$ as independently varying unknowns. The numerical 2 is introduced solely for notational tidiness.

The rationale for this procedure is straightforward.[35] Consider first, a very simple example, of one equation in two unknowns,

{lagrange1}
$$x_1 - x_2 = 1, \tag{2.149}$$

and we seek the minimum norm solution,

$$J = x_1^2 + x_2^2, \tag{2.150}$$

subject to Eq. (2.149). The differential,

$$dJ = 2x_1dx_1 + 2x_2dx_2 = 0, \tag{2.151}$$

leads to the unacceptable solution $x_1 = x_2 = 0$, if we should incorrectly set the coefficients of dx_1, dx_2 to zero. Consider instead a modified objective function

$$J' = J - 2\mu\left(x_1 - x_2 - 1\right), \tag{2.152}$$

where μ is unknown. The differential of J' is

{lagrange2}
$$dJ' = 2x_1 dx_1 + 2x_2 dx_2 - 2\mu\,(dx_1 - dx_2) - 2\,(x_1 - x_2 - 1)\,d\mu = 0, \tag{2.153}$$

or

$$dJ'/2 = dx_1\,(x_1 - \mu) + dx_2\,(x_2 + \mu) - d\mu\,(x_1 - x_2 - 1) = 0. \tag{2.154}$$

We are free to choose, $x_1 = \mu$ which kills off the differential involving dx_1. But then only the differentials $dx_2, d\mu$ remain; as they can vary independently, their coefficients must vanish separately, and we have,

$$x_2 = -\mu \tag{2.155}$$

$$x_1 - x_2 = 1. \tag{2.156}$$

Note that the second of these recovers the original equation. Substituting $x_1 = \mu$, we have $2\mu = 1$, or $\mu = 1/2$, and $x_1 = 1/2, x_2 = -1/2$, $J = 0.5$, and one can confirm that this is indeed the "constrained" minimum. (A "stationary" value of J' was found, not an absolute minimum value, because J' is no longer necessarily positive; it has a saddle point, which we have found.)

Before writing out the general case, note the following question: Suppose the constraint equation was changed to,

$$x_1 - x_2 = \Delta. \tag{2.157}$$
{lineqs2}

How much would J change as Δ is varied? With variable Δ, (2.153) becomes,

$$dJ' = 2dx_1\,(x_1 - \mu) + 2dx_2\,(x_2 + \mu) - 2d\mu\,(x_1 - x_2 - \Delta) + 2\mu d\Delta. \tag{2.158}$$

But the first three terms on the right vanish, and hence,

$$\frac{\partial J'}{\partial \Delta} = 2\mu = \frac{\partial J}{\partial \Delta}, \tag{2.159}$$
{sensiv1}

because $J = J'$ at the stationary point (from (2.157). *Thus 2μ is the sensitivity of the objective function J to perturbations in the right-hand side of the constraint equation.* If Δ is changed from 1, to 1.2, it can be confirmed that the approximate change in the value of J is 0.2 as one deduces immediately from Eq. (2.159).

We now develop this method generally. Reverting to Eq. (2.148),

$$\begin{aligned} dJ' &= dJ - 2\boldsymbol{\mu}^T\mathbf{A}d\mathbf{x} - 2\left(\mathbf{A}\mathbf{x}-\mathbf{b}\right)^T d\boldsymbol{\mu} \\ &= \left(\frac{\partial J}{\partial x_1} - 2\boldsymbol{\mu}^T\mathbf{a}_1\right)dx_1 + \left(\frac{\partial J}{\partial x_2} - 2\boldsymbol{\mu}^T\mathbf{a}_2\right)dx_2 + \cdots + \left(\frac{\partial J}{\partial x_N} - 2\boldsymbol{\mu}^T\mathbf{a}_N\right)dx_N \qquad (2.160) \\ &\quad -2\left(\mathbf{A}\mathbf{x}-\mathbf{b}\right)^T d\boldsymbol{\mu} \\ &= \left(2x_1 - 2\boldsymbol{\mu}^T\mathbf{a}_1\right)dx_1 + \left(2x_2 - 2\boldsymbol{\mu}^T\mathbf{a}_2\right)dx_2 + \ldots + \left(2x_N - 2\boldsymbol{\mu}^T\mathbf{a}_N\right)dx_N \qquad (2.161) \\ &\quad -2\left(\mathbf{A}\mathbf{x}-\mathbf{b}\right)^T d\boldsymbol{\mu} = 0 \end{aligned}$$

{lagrange3}

Here the $\mathbf{a}_i$ are the corresponding columns of $\mathbf{A}$. The coefficients of the first $M-$differentials dx_i can be set to zero by assigning, $x_i = \boldsymbol{\mu}^T\mathbf{a}_i$, leaving $N-M$ differentials dx_i whose coefficients must separately vanish (hence they *all* vanish, but for two separate reasons), plus the coefficient of the $M - d\mu_i$ which must also vanish separately. This recipe produces, from Eq. (2.161),

{33036a}
$$\tfrac{1}{2}\frac{\partial J'}{\partial \mathbf{x}} = \mathbf{x} - \mathbf{A}^T\boldsymbol{\mu} = 0 \qquad (2.162)$$

{33036b}
$$\tfrac{1}{2}\frac{\partial J'}{\partial \boldsymbol{\mu}} = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}\,, \qquad (2.163)$$

where the first equation set is the result of the vanishing of the coefficients of dx_i and the second, which is the original set of equations, arises from the vanishing of the coefficients of the $d\mu_i$. The convenience of being able to treat all the x_i as independently varying is offset by the increase in problem dimensions by the introduction of the $M-$unknown μ_i. The first set is $N-$equations for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$, and the second set is $M-$equations in $\mathbf{x}$ in terms of $\mathbf{y}$. Taken together, these are $M+N$ equations in $M+N$ unknowns, and hence just-determined no matter what the ratio of M to N.

Eq. (2.162) is,

{33037}
$$\mathbf{A}^T\boldsymbol{\mu} = \mathbf{x} \qquad (2.164)$$

and substituting for $\mathbf{x}$ into (2.163),

{33038}
$$\begin{aligned} \mathbf{A}\mathbf{A}^T\boldsymbol{\mu} &= \mathbf{b}\,, \\ \tilde{\boldsymbol{\mu}} &= (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}\,, \qquad (2.165) \end{aligned}$$

assuming the inverse exists, and

{33039a}
$$\tilde{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b} \qquad (2.166)$$

{33039b}
$$\tilde{\mathbf{n}} = \mathbf{0} \qquad (2.167)$$

{33039c}
$$\mathbf{C}_{xx} = 0. \qquad (2.168)$$

($\mathbf{C}_{xx} = 0$ because formally we estimate $\tilde{\mathbf{n}} = \mathbf{0}$).

Eqs.(2.166-2.168) are the classical solution of minimum norm of $\mathbf{x}$, satisfying the constraints exactly while minimizing the solution length. That a minimum is achieved can be verified by evaluating the second derivatives of J' at the solution point. The minimum occurs at a saddle point in $\mathbf{x}$, $\boldsymbol{\mu}$ space[36] and where the term proportional to $\boldsymbol{\mu}$ necessarily vanishes. The operator $\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ is sometimes called a "Moore-Penrose inverse."

Eqs. (2.164) for $\boldsymbol{\mu}$ in terms of $\mathbf{x}$ involves the coefficient matrix $\mathbf{A}^T$. An intimate connection exists between matrix transposes and adjoints of differential equations (see the Appendix to this Chapter), and thus $\boldsymbol{\mu}$ is sometimes called the "adjoint solution," with $\mathbf{A}^T$ defining the "adjoint model"[37] in Eq.(2.164), and $\mathbf{x}$ acting as a forcing term. The original Eqs. $\mathbf{Ax} = \mathbf{b}$, were assumed formally underdetermined, and thus the adjoint model equations in (2.164) are necessarily formally overdetermined.

Example

Now do the last example using matrix vector notation defining,

$$\mathbf{A} = \left\{ \begin{matrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{matrix} \right\}, \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$J = \mathbf{x}^T\mathbf{x} - \mathbf{2}\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{b})$$

$$\frac{d}{d\mathbf{x}}\left(\mathbf{x}^T\mathbf{x} - \mathbf{2}\boldsymbol{\mu}^T(\mathbf{Ax} - \mathbf{b})\right) = 2\mathbf{x} - \mathbf{2}\mathbf{A}^T\boldsymbol{\mu} = \mathbf{0}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{b}$$

$$\mathbf{x} = [3/4, -1/2, 3/4]^T$$

Example

Write out J' :

$$J' = x_1^2 + x_2^2 + x_3^2 - 2\mu_1(x_1 + x_2 + x_3 - 1) - 2\mu_2(x_1 - x_2 + x_3 - 2)$$

$$\begin{aligned} dJ' &= (2x_1 - 2\mu_1 - 2\mu_2)\,dx_1 + (2x_2 - 2\mu_1 + 2\mu_2)\,dx_2 + (2x_3 - 2\mu_1 - 2\mu_2)\,dx_3 \\ &+ (-2x_1 - 2x_2 + 2 - 2x_3)\,d\mu_1 + (-2x_1 + 2x_2 - 2x_3 + 4)\,d\mu_2 \\ &= 0 \end{aligned}$$

Set $x_1 = \mu_1 + \mu_{2,} x_2 = \mu_1 - \mu_2$ *so that the first two terms vanish, and set the coefficients of the differentials of the remaining, independent terms to zero,*

$$\begin{aligned}
\frac{dJ'}{dx_1} &= 2x_1 - 2\mu_1 - 2\mu_2 = 0, \\
\frac{dJ'}{dx_2} &= 2x_2 - 2\mu_1 + 2\mu_2 = 0, \\
\frac{dJ'}{dx_3} &= 2x_3 - 2\mu_1 - 2\mu_2 = 0, \\
\frac{dJ'}{d\mu_1} &= -2x_1 - 2x_2 + 2 - 2x_3 = 0, \\
\frac{dJ'}{d\mu_2} &= -2x_1 + 2x_2 - 2x_3 + 4 = 0.
\end{aligned}$$

Then,

$$dJ' = (2x_3 - 2\mu_1 - 2\mu_2)\, dx_3 + (-2x_1 - 2x_2 + 2 - 2x_3)\, d\mu_1 + (-2x_1 + 2x_2 - 2x_3 + 4)\, d\mu_2 = 0,$$

or

$$\begin{aligned}
x_1 &= \mu_1 + \mu_{2,} \\
x_2 &= \mu_1 - \mu_2 \\
x_3 - \mu_1 - \mu_2 &= 0 \\
-x_1 - x_2 + 1 - x_3 &= 0 \\
-x_1 + x_2 - x_3 + 2 &= 0
\end{aligned}$$

That is,

$$\begin{aligned}
\mathbf{x} &= \mathbf{A}^T \boldsymbol{\mu} \\
\mathbf{A}\mathbf{x} &= \mathbf{b}
\end{aligned}$$

or,

$$\left\{ \begin{array}{cc} \mathbf{I} & -\mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{array} \right\} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

But in this particular case, the first set can be solved for $\mathbf{x} = \mathbf{A}^T \boldsymbol{\mu}$,

$$\begin{aligned}
\boldsymbol{\mu} &= \left(\mathbf{A}\mathbf{A}^T\right)^{-1} \mathbf{b} = \begin{bmatrix} 1/8 & 5/8 \end{bmatrix}^T, \\
\mathbf{x} &= \mathbf{A}^T \begin{bmatrix} \frac{1}{8} \\ \frac{5}{8} \end{bmatrix} = \begin{bmatrix} 3/4 & -1/2 & 3/4 \end{bmatrix}^T
\end{aligned}$$

Suppose instead we wanted to minimize,

$$J = (x_1 - x_2)^2 + (x_2 - x_3)^2 = \mathbf{x}^T \mathbf{F}^T \mathbf{F} \mathbf{x}$$

where

$$\mathbf{F} = \left\{ \begin{matrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{matrix} \right\}$$

$$\mathbf{F}^T \mathbf{F} = \left\{ \begin{matrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{matrix} \right\}^T \left\{ \begin{matrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{matrix} \right\} = \left\{ \begin{matrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{matrix} \right\}.$$

S*uch an objective function might be used to find a "smooth" solution. One confirms,*

$$\begin{aligned} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \left\{ \begin{matrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{matrix} \right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + x_3^2 \\ &= (x_1 - x_2)^2 + (x_2 - x_3)^2 . \end{aligned}$$

The stationary point of,

$$J' = \mathbf{x}^T \mathbf{F}^T \mathbf{F} \mathbf{x} - \mathbf{2}\boldsymbol{\mu}^T (\mathbf{A}\mathbf{x} - \mathbf{b}),$$

leads to

$$\begin{aligned} \mathbf{F}^T \mathbf{F} \mathbf{x} &= \mathbf{A}^T \boldsymbol{\mu} \\ \mathbf{A}\mathbf{x} &= \mathbf{b} \end{aligned}$$

But,

$$\mathbf{x} \neq \left(\mathbf{F}^T \mathbf{F} \right)^{-1} \mathbf{A}^T \boldsymbol{\mu}$$

because there is no inverse (guaranteed). But the coupled set

$$\left\{ \begin{matrix} \mathbf{F}^T \mathbf{F} & -\mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{matrix} \right\} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

does have a solution.

The physical interpretation of $\boldsymbol{\mu}$ can be obtained as above by considering the way in which J would vary with infinitesimal changes in $\mathbf{b}$. As in the special case done above, $J = J'$ at the stationary point. Hence,

$$dJ' = dJ - 2\boldsymbol{\mu}^T \mathbf{A} d\mathbf{x} - 2\left(\mathbf{A}\mathbf{x} - \mathbf{b}\right)^T d\boldsymbol{\mu} + 2\boldsymbol{\mu}^T d\mathbf{b} = 0, \tag{2.169}$$

or, since the first three terms on the right vanish at the stationary point,

$$\frac{\partial J'}{\partial \mathbf{b}} = \frac{\partial J}{\partial \mathbf{b}} = 2\boldsymbol{\mu}. \tag{2.170}$$

Thus, as inferred above, the Lagrange multipliers are the sensitivity of J, at the stationary point, to perturbations in the parameters $\mathbf{y}$. This conclusion leads, in Chapter 4, to the scrutiny of the Lagrange multipliers as a means of understanding the sensitivity of models and the flow of information within them.

Now revert to $\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}$, that is, equations containing noise. If these are first column scaled using $\mathbf{S}^{-T/2}$, Eqs. (2.166)–(2.168) are in the primed variables, and the solution in the original variables is,

{33041a}
$$\tilde{\mathbf{x}} = \mathbf{S}\mathbf{E}^T (\mathbf{E}\mathbf{S}\mathbf{E}^T)^{-1} \mathbf{y} \tag{2.171}$$

{33041b}
$$\tilde{\mathbf{n}} = \mathbf{0} \tag{2.172}$$

{33041c}
$$\mathbf{C}_{xx} = \mathbf{0}, \tag{2.173}$$

and the result depends directly upon $\mathbf{S}$. If a row scaling with $\mathbf{W}^{-T/2}$ is used, it is readily shown that $\mathbf{W}$ disappears from the solution and has no effect on it (see page 111, below). gerowscale1}

Eqs. (2.171)–(2.173) are a solution, but there is the same fatal defect as in Eq. (2.172)— $\tilde{\mathbf{n}} = \mathbf{0}$ is usually unacceptable when $\mathbf{y}$ are observations. Furthermore, $\|\tilde{\mathbf{x}}\|$ is again uncontrolled, and $\mathbf{E}\mathbf{S}\mathbf{E}^T$ may not have an inverse.

$\mathbf{n}$ must be regarded as fully an element of the solution, as much as $\mathbf{x}$. Equations representing observations can always be written as in (2.139), and can be solved exactly. Therefore, we now use a modified objective function, allowing for general $\mathbf{S}, \mathbf{W}$,

{33042}
$$J = \mathbf{x}^T \mathbf{S}^{-1} \mathbf{x} + \mathbf{n}^T \mathbf{W}^{-1} \mathbf{n} - 2\boldsymbol{\mu}^T (\mathbf{E}\mathbf{x} + \mathbf{n} - \mathbf{y}), \tag{2.174}$$

with both $\mathbf{x}$, $\mathbf{n}$ appearing in the objective function. Setting the derivatives of (2.174) with respect to $\mathbf{x}$, $\mathbf{n}$, $\boldsymbol{\mu}$ to zero, and solving the resulting normal equations produces,

$$\begin{aligned}
\tilde{\mathbf{x}} &= \mathbf{SE}^T\left(\mathbf{ESE}^T{+}\mathbf{W}\right)^{-1}\mathbf{y} && (2.175)\\
\tilde{\mathbf{n}} &= \mathbf{y}-\mathbf{E}\tilde{\mathbf{x}} && (2.176)\\
\mathbf{C}_{xx} &= \mathbf{SE}^T\left(\mathbf{ESE}^T{+}\mathbf{I}\right)^{-1}\mathbf{R}_{nn}\left(\mathbf{ESE}^T{+}\mathbf{I}\right)^{-1}\mathbf{ES} && (2.177)
\end{aligned}$$

$$\tilde{\boldsymbol{\mu}}{=}\,\mathbf{W}^{-1}\tilde{\mathbf{n}} \tag{2.178}$$

---

which are identical to Eqs. (2.136-2.138) or to the alternate from Eq.(2.128 − 2.130) derived from an objective function without Lagrange multipliers.

Eqs. (2.136-2.138) and (2.175-2.177) result from two very different appearing objective functions—one in which the equations are imposed in the mean-square, and one in which they are imposed exactly, using Lagrange multipliers. Constraints in the mean-square will be termed "soft", and those imposed exactly are "hard."[38] The distinction is, however, largely illusory: although (2.88) are being imposed exactly, it is only the presence of the error term, **n**, which permits the equations to be written as equalities and thus as hard constraints. The hard and soft constraints here produce an identical solution. In some (rare) circumstances, which we will discuss briefly below, one may wish to impose exact constraints upon the elements of $\tilde{x}_i$. The solution (2.166)–(2.168) was derived from the noise-free hard constraint, $\mathbf{Ax}=\mathbf{b}$, but we ended by rejecting it as generally inapplicable.

Once again, **n** is only by convention discussed separately from **x**, and is fully a part of the solution. The combined form (2.139), which literally treats **x**, **n** as the solution, are imposed through a hard constraint on the objective function,

$$J=\boldsymbol{\xi}^T\boldsymbol{\xi}-2\boldsymbol{\mu}^T\left(\mathbf{E}_1\boldsymbol{\xi}-\mathbf{y}\right), \tag{2.179}$$ {33045}

where $\boldsymbol{\xi}=[\mathbf{S}^{-T/2}\mathbf{x},\mathbf{W}^{-T/2}\mathbf{n}]^T$, which is Eq. (2.174). (There are numerical advantages, however, in working with objects in two spaces of dimensions M and N, rather than a single space of dimension $M+N$.)

2.4.4 Interpretation of Discrete Adjoints

When the operators are matrices, as they are in discrete formulations, then the adjoint is just the transposed matrix. Sometimes the adjoint has a simple physical interpretation. Suppose, e.g., that scalar y was calculated from a sum,

$$y=\mathbf{Ax},\ \mathbf{A}=\left\{\begin{matrix}1 & 1 & . & 1 & 1\end{matrix}\right\}. \tag{2.180}$$ {sumoper1}

Then the adjoint operator applied to y is evidently,

$$\mathbf{r} = \mathbf{A}^T y = \left\{ \begin{matrix} 1 & 1 & 1 & . & 1 \end{matrix} \right\}^T y = \mathbf{x} \tag{2.181}$$

Thus the adjoint operator "sprays" the average back out onto the originating vector, and might be thought of as an inverse operator.

A more interesting case is a first-difference forward operator,

$$\mathbf{A} = \left\{ \begin{matrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & . & . & . \\ & & & & -1 & 1 \\ & & & & & -1 \end{matrix} \right\}, \tag{2.182}$$

that is,

$$y_i = x_{i+1} - x_i, \tag{2.183}$$

(with the exception of the last element, $y_N = -x_N$).

Then its adjoint is,

$$\mathbf{A}^T = \left\{ \begin{matrix} -1 & & & & & \\ 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & . & . & & \\ & & & 1 & -1 & \\ & & & & 1 & -1 \end{matrix} \right\} \tag{2.184}$$

that is a first-difference *backward* operator with $\mathbf{z} = \mathbf{A}^T \mathbf{y}$, producing $z_i = y_{i-1} - y_i$ with again, the exception of the end point, now z_1.

In general, the transpose matrix, or adjoint operator is *not* simply interpretable as an inverse operation as in the summation/spray-out case might have suggested.[39] A more general understanding of the relationship between adjoints and inverses will be obtained in the next Section.