## 2.9 A Recapitulation

This chapter has not exhausted the possibilities for inverse methods, and the techniques will be extended in several directions in the next Chapters. Given the lengthy nature of the discussion so far, however, some summary of what has been accomplished may be helpful.

The focus is on making inferences about parameters or fields, $\mathbf{x}$, $\mathbf{n}$ satisfying linear relationships of the form

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y} \, .$$

Such equations arise as we have seen, from both "forward" and "inverse" problems, but the techniques for estimating $\mathbf{x}$, $\mathbf{n}$ and their uncertainty are useful whatever the physical origin of the equations. Two methods for estimating $\mathbf{x}$, $\mathbf{n}$ have been the focus of the chapter: least-squares (including the singular value decomposition) and the Gauss-Markov or minimum variance technique. Least-squares, in any of its many guises, is a very powerful method—but its power and

Continued on next page...

ease of use have (judging from the published literature) led many investigators into serious confusion about what they are doing. This confusion is compounded by the misunderstandings about the difference between an inverse problem and an inverse method.

An attempt is made therefore, to emphasize the two distinct roles of least-squares: as a method of *approximation*, and as a method of *estimation*. It is only in the second formulation that it can be regarded as an inverse method. A working definition of an inverse method is a technique able to estimate unknown parameters or fields of a model, while producing an estimate of the uncertainties of the results. Solution plus uncertainty are the fundamental requirements. There are many desirable additional features of inverse methods which can prove extremely important. Among these features are: (1) Separation of nullspace uncertainties from observational noise uncertainties; (2) the ability to rank the data in its importance to the solution; (3) the ability to use prior statistical knowledge; (4) understanding of solution structures in terms of data structure, (5) the ability to trade resolution against variance. (The list is not exhaustive. For example, we will briefly examine in Chapter 4 the use of inequality information.) As with all estimation methods, one also trades computational load against the need for information. (The SVD, for example, is a powerful form of least-squares, but requires more computation that do other forms). The Gauss-Markov approach has the strength of forcing explicit use of prior statistical information and is directed at the central goal of obtaining $\mathbf{x}$, $\mathbf{n}$ with the smallest mean-square error, and for this reason might well be regarded as the default methodology for linear inverse problems. It has the added advantage that we know we can obtain precisely the same result with appropriate versions of least-squares, including the SVD, permitting the use of least-squares algorithms, but at the risk of losing sight of what we are actually attempting. A limitation is that the underlying probability densities of solution and noise have to be unimodal (so that a minimum variance estimate makes sense). If unimodality fails, one must look to other methods.

The heavy emphasis here on noise and uncertainty may appear to be a tedious business. But readers of the scientific literature will come to recognize how qualitative much of the discussion is—the investigator telling a story about what he thinks is going on with no estimate of uncertainties, and no attempt to resolve quantitatively differences with previous competing estimates of the circulation in the particular region. In a quantitative subject, such vagueness is ultimately intolerable.

A number of different procedures for producing estimates of the solution to a set of noisy simultaneous equations of arbitrary dimension have been described here. The reader may wonder which of the variants makes the most sense to use in practice. Because, in the presence of noise one is dealing with a statistical estimation problem, there is no single "best" answer, and one

must be guided by model context and goals. A few general remarks might be helpful.

In any problem where data are to be used to make inferences about physical parameters, one typically needs some approximate idea of just how large the solution is likely to be and how large the residuals probably are. In this nearly agnostic case, where almost nothing else is known, and the problem is very large, the weighted, tapered least-squares solution is a good first choice—it is easily and efficiently computed and coincides with the Gauss-Markov and tapered SVD solutions, if the weight matrices are the appropriate covariances. Sparse matrix methods exist for its solution should that be necessary.[58] Coincidence with the Gauss-Markov solution means one can reinterpret it as a minimum-variance or maximum-likelihood solution (See the Chapter Appendix) should one wish.

It is a comparatively easy matter to vary the trade-off parameter, $\gamma^2$, to explore the consequences of any errors in specifying the noise and solution variances. Once a value for $\gamma^2$ is known, the tapered SVD can then be computed to understand the relationships between solution and data structures, their resolution and their variance. For problems of small to moderate size (the meaning of "moderate" is constantly shifting, but it is difficult to examine and interpret matrices of more than about $500 \times 500$), the SVD, whether in the truncated or tapered forms is probably the method of choice—because it provides the fullest information about data and its relationship to the solution. Its only disadvantages are that one can easily be overwhelmed by the available information, particularly if a range of solutions must be examined, and it cannot take advantage of sparsity in large problems. The SVD has a flexibility beyond even what we have discussed—one could, for example, change the degree of tapering in each of the terms of (2.336)–(2.337) should there be reason to repartition the variance between solution and noise, or some terms could be dropped out of the truncated form at will—should the investigator know enough to justify it.

To the extent that either or both of $\mathbf{x}, \mathbf{n}$ have expected structures expressible through covariance matrices, these structures can be removed from the problem through the various weight matrix and/or the Cholesky decomposition. The resulting problem is then one in completely unstructured (equivalent to white noise) elements $\mathbf{x}, \mathbf{n}$. In the resulting scaled and rotated systems, one can use the simplest of all objective functions. Covariance, resolution etc., in the original spaces of $\mathbf{x}, \mathbf{n}$ is readily recovered by appropriately applying the weight matrices to the results of the scaled/rotated space.

Both ordinary weighted least-squares and the SVD applied to row- and column-weighted equations are best thought of as approximation, rather than estimation, methods. In particular, the truncated SVD does not produce a minimum variance estimate the way the tapered version can. The tapered SVD (along with the Gauss-Markov estimate, or the tapered least-squares

solutions) produce the minimum variance property by tolerating a bias in the solution. Whether the bias is more desirable than a larger uncertainty is a decision the user must make. But the reader is warned against the belief that there is any single best method.