SIO 230 Geophysical Inverse Theory 2009

Annex to Chapter 3 of GIT

Distribution of the Misfit Norm

Here is a short addendum to the discussion of misfit in Chapter 3 of GIT. We wish to use a norm to measure the misfit between the predictions of our model and the measured data vector $d \in \mathbb{R}^N$. Initially we will assume that the noise in the observations is normally distributed, the components of the noise vector being iid $\mathcal{N}(0, \sigma^2)$. The situation is illustrated below for $N = 3$, the dots representing a large number of independent measurements, each with its own random error attached to the true data vector. Given this assumption, how large should the misfit be when we use the ordinary 2-norm? The squared length of a random vector of iid normal components,

$$|| X ||^2 = X_1^2 + X_2^2 + \cdots + X_N^2 \tag{1}$$

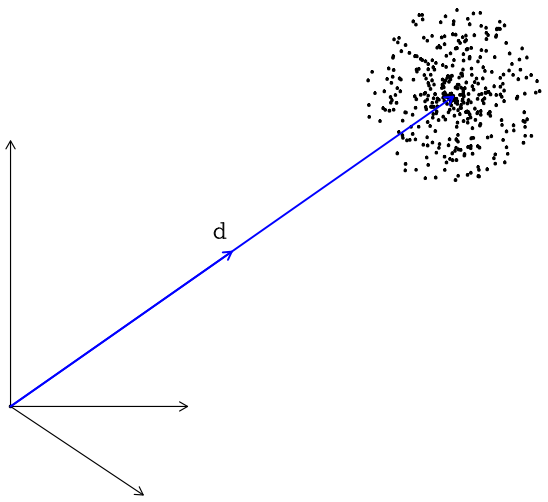
is distributed as the classical χ^2_N distribution scaled by the variance of the noise, σ^2. It is easy to see that the average, or expected value, of the squared norm is just $N\sigma^2$, just the sum of the variances in (1), since each component is independent of the the others. So

$$\mathcal{E}[\,|| X ||]^2] = N\sigma^2 \,. \tag{2}$$

Less obvious is the result that

$$\mathrm{var}\,[\,|| X ||^2] = 2N\sigma^4 \quad \text{or} \quad \sigma_{\chi^2} = \sigma^2\sqrt{2N} \,. \tag{3}$$

This means as N, the number data, grows large, the relative width of the distribution gets narrower.



We use the statistical theory in the following way: We ask for a tolerance T such that the actual error produced by random noise will be less than T with probability P, where P is a value like 0.5 or, if we are feeling cautious, 0.95 say. If we choose 0.5 we are saying that the tolerance will be met about half the time in hypothetical repeated experiments; with 0.95 the actual error will be larger than the assigned misfit only 5% of the time. Larger values of T allow bigger misfits between the observations and the model, and risk generation of over-simplified, smooth models. Conversely, T too small risks production of models that are too complex. In practice, the difference between solutions where P has been chosen to 0.5 and those with $P = 0.95$ is not very great.

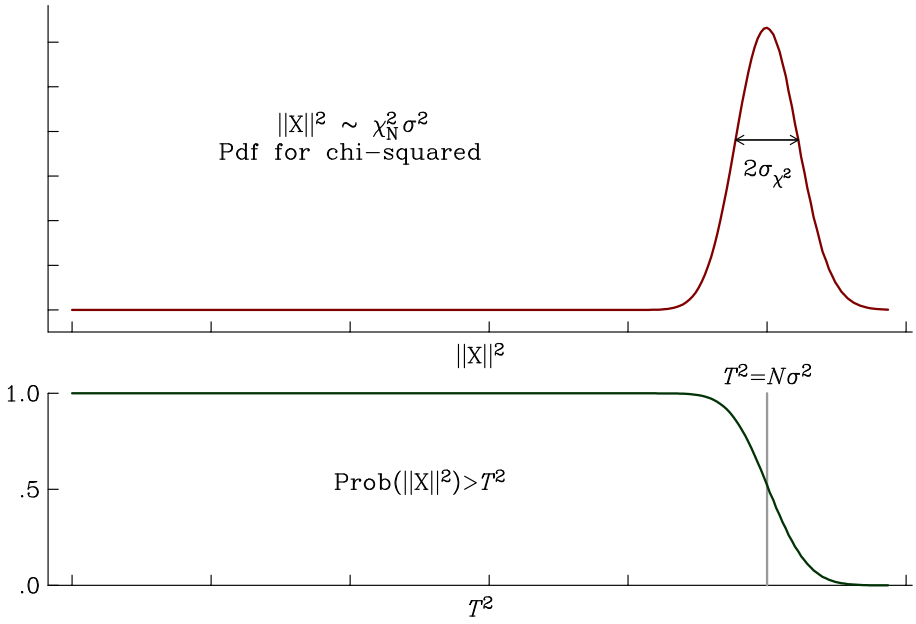
If we stick with zero-mean Gaussian noise, the first generalization beyond iid components in the noise is to allow different variances for each component. Then we can scale the measurements by dividing each observation by its standard error. The associated error is a *standardized* random variable $Y_j = X_j/\sigma_j$. This variable has zero mean and unit variance. We would write $Y_j \pm 1$. Then

$$|| \Sigma^{-1} X ||^2 = \frac{X_1^2}{\sigma_1^2} + \frac{X_2^2}{\sigma_2^2} + \cdots + \frac{X_N^2}{\sigma_N^2} \tag{4}$$

$$= Y_1^2 + Y_2^2 + \cdots + Y_N^2 \sim \chi_N^2 \tag{5}$$

where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots \sigma_N)$. So now we choose a tolerance for $|| \Sigma^{-1} X ||$ just as before.

The next generalization is to allow the noise to be correlated with known covariance matrix $C \in \mathbb{R}^{N \times N}$. This means that



$$\mathrm{cov}[X_j, X_k] = \mathcal{E}[X_j X_k] = C_{jk} \,. \tag{6}$$

We find that the following quadratic form has the chi-squared distribution:

$$X^T C^{-1} X \;\sim\; \chi^2_N \,. \tag{7}$$

Because C is always positive definite the quadratic form is the square of a valid norm on $\mathbb{R}^N$. In my experience the situation in which C is known, or even only roughly estimated, for a real geophysical data set is very rare.

The three types of Gaussian noise, iid noise, uncorrelated Gaussian noise, or correlated noise, can all be treated in the same general way by demanding that

$$\| A\,[\Theta(m) - d] \|^2 \le T^2 \tag{8}$$

where $\Theta(m)$ gives the prediction of the data vector d from the model m, and $A \in \mathbb{R}^{N\times N}$ is a square matrix: either I/σ, Σ^{-1}, or L^{-1} where $C = L\,L^T$ is the Cholesky factorization of a positive definite matrix, briefly described in Section 7 of the Supplementary Notes. In (8) T is determined by the using the chi-squared distribution to choose a probability level for the validity of the inequality. Roughly speaking for large N, we find $T^2 = N$, corresponds to $P = 0.5$, and $T^2 = N + 2\sqrt{2N}$, gives $P = 0.977$. In MATLAB you can use the function `chi2inv` which gives

$$T^2 = \mathrm{chi2inv}(P, N)$$