Spectrum (and other data) fitting

We normally fit by minimizing a *cost function*, usually χ^2 :

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - F(x_i, a_1, \cdots, a_m)}{\sigma_i} \right]^2 ,$$

where the a_k are parameters (m of them in total), y_i are measurements (*e.g.*, the spectrum), x_i are values of independent variable (*e.g.* wavelength or wavenumber), and σ_i are the uncertainties ($1/\sigma_i^2$ = *weight*).

Linear case: $F(x_i, a_1, \cdots, a_m) = \sum_{k=1}^{m} a_k X_k(x_i)$. The $X_k(x_i)$ are *basis functions*. They can be wildly nonlinear in x_i (like a spectrum usually is): only the a_k dependence is linear. They might be cross sections for different molecules, for example, making up a spectrum that is *optically thin* or that can be linearized using the Beer-Lambert condition. Then,

$$\frac{\partial F(x_i, \bar{a})}{\partial a_k} = X_k(x_i), \quad (\bar{a} \text{ is the vector of the parameters})$$

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - \sum_{k=1}^{m} a_k X_k(x_i)}{\sigma_i} \right]^2$$

$b_i \equiv y_i / \sigma_i$ and $A_{ij} \equiv X_j(x_i) / \sigma_i \quad (m \times n$ matrix), then $\chi^2 = (\bar{b} - \underline{A}\bar{a})^2$.

at the minimum, $\dfrac{\partial \chi^2}{\partial a_j} = 0, j = 1, \cdots, m$

$\dfrac{\partial \chi^2}{\partial \bar{a}} = 0 = -2\underline{A}^T(\bar{b} - \underline{A}\bar{a}), \quad (\underline{A}^T \underline{A})\bar{a} = \underline{A}^T \bar{b}$. *T* denotes the *transpose* of the matrix.

Our vector of parameters is thus $\bar{a} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \bar{b}$.

$A^T A$ is usually called $\underline{\alpha}, \alpha_{kj} = \sum_i \dfrac{\partial F(x_i, \bar{a})}{\partial a_j} \dfrac{\partial F(x_i, \bar{a})}{\partial a_k}$

$\underline{\alpha}$ = ½ of the *Hessian* matrix (2nd derivatives of χ^2)

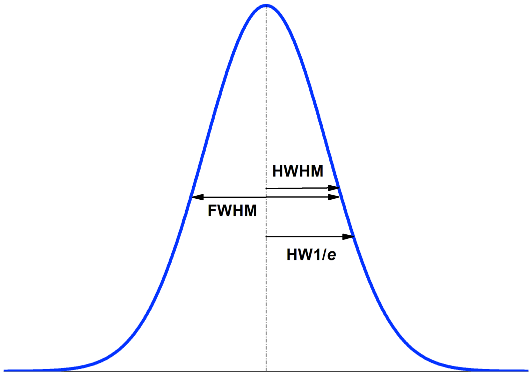
$\underline{\alpha}^{-1} = (A^T A)^{-1} \equiv \underline{C}$, the *covariance* matrix (of the standard errors). The uncertainty in each parameter is $\sigma(a_j) = \sqrt{c_{jj}}$. c_{jk} gives the covariance among parameters.

The *correlation* matrix $\equiv \dfrac{c_{ij}}{\sqrt{c_{ii} c_{jj}}}$.

Minimum χ^2 gives a goodness of fit indicator, $\Gamma\left(\frac{n-m}{2},\frac{\chi^2}{2}\right), 0 \le \Gamma \le 1$. Γ is the probability that χ^2 should exceed the fitted χ^2 by chance (see *Numerical Recipes* for details). Rule of thumb: $\chi^2 \sim n-m$ is good.

However, if the σ_i are not known or trusted *and* the model is known to be good, one may use $\sigma = RMS\sqrt{\frac{n}{n-m}}, \quad RMS = \left[\sum_{i=1}^{n}\frac{(y_i - F(x_i,\bar{a}))^2}{n}\right]^{1/2}$. If we do this, however, we cannot obtain an independent goodness of fit.

Forms of spectral noise, signal-to-noise ratio (S/N)



Horowitz and Hill, Chapter 7 has valuable discussions of noise types and sources.

A Gaussian line,

$$l_g(\sigma) = \frac{\pi^{-1/2}}{b_e}\exp-\left[\frac{(\sigma-\sigma_0)^2}{b_e^2}\right], \int_{-\infty}^{\infty} l_g(\sigma)\,d\sigma = 1.$$

where b_e is the *half-width at 1/e intensity* (hw1/e), to be compared later to the *half-width at half maximum (HWHM)* and the *full-width at half maximum (FWHM).* I prefer using hw1/e to describe Gaussians and HWHM for *Lorentzian* lineshapes (of which more later).

Gaussians widths add in *quadrature* (when convolving): $b_{total} = \sqrt{b_1^2 + b_2^2}$ (Easy to show with the convolution theorem – *try it!*)

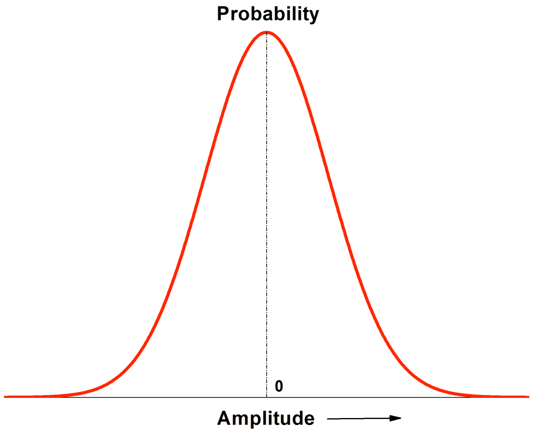
Signal-to-noise-ratio, S/N

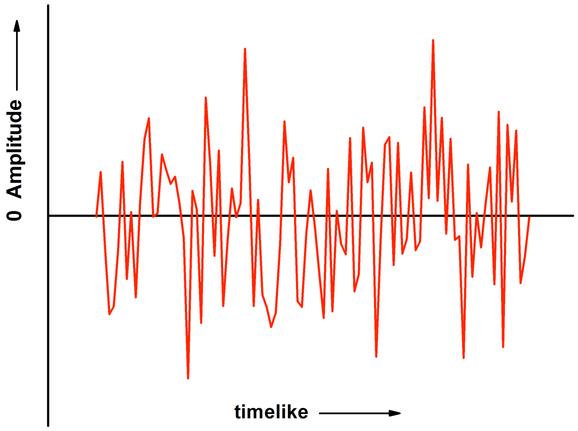
Signal: The signal of a system increases linearly with power. **Antenna temperature**, T_A: The signal (at a particular wavenumber, σ, or frequency, ν) is equivalent to the antenna being enclosed in a blackbody of temperature T. T_A of a line is usually defined for the line center.

Noise: We usually have (approximately) *band-limited white Gaussian noise*:

- Equal power per Hz (or cm^{-1}: a frequency unit)
- Gaussian distribution (±) of amplitudes

Measure at a given frequency:





Gaussian description of noise: For noise,
$\sigma_0 = 0, b_n = b_e / \sqrt{2}$
b_n = root-mean-square (RMS) noise = our noise for S/N purposes

Probability of amplitude $A, P_A = \frac{(2\pi)^{-1/2}}{b_n} \exp - \left[\frac{A^2}{2{b_n}^2} \right]$

Noise integrates up as $\sqrt{t}$ (because Gaussians add in quadrature), while signal integrates up as $t \Rightarrow$ S/N increases as $\sqrt{t}$.

Types of noise:

1. Noise components from the instruments (detector noise, readout noise, electronic noise) will generally be independent of the spectral intensity. They are generally (to a reasonable degree of fidelity) described as Gaussian white noise. In radio physics and astronomy noise, squares of noise sources are often described as *temperatures*, which add linearly to give a noise system *temperature*: Remember that, in the Rayleigh-Jeans limit, power is linearly proportional to temperature. Since noise increases as $\sqrt{\text{power}}$, again because sources add in quadrature, noise temperature sources add linearly.

2. A component to the whole system noise that is due to photon statistics, that is, to the fact that we are counting a discrete number of photons, *N*, is also proportional to $\sqrt{N}$ (proportional to $\sqrt{t}$ for linear integration). The S/N is thus proportional to $N / \sqrt{N} = \sqrt{N}.$ Where the spectrum is larger (say, at the peak of an emission line), the noise will be larger than at the trough, but the S/N signal will be lower. The margin of error (1 standard deviation, although almost never stated) usually given in political polls is $1/\sqrt{N}$, where *N* is the number of persons polled. This can result in less popular candidates having possibly negative approval ratings or likely voters! See where the problem arises?

The second noise source is described by **Poisson** statistics: Poisson statistics describes discrete events. From the ***Wikipedia*:**

In probability theory and statistics, the ***Poisson distribution*** *is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume ... The fluctuations about the mean value of events are denoted as* ***Poisson noise*** *or (particularly in electronics) as* ***shot noise****.*

A good noise generation program is often very useful. **noise.f90** is available at the class website. You may want to generate a noise spectrum with this program and test it to see how Gaussian the amplitude distribution is.

System temperature (T_{sys}) and noise temperature (T_N)

At low σ,

$$R(\sigma_0) = \frac{2\pi hc^2\sigma^3}{e^{c_2\sigma/T}-1};\ 2\pi kTc\sigma^2 \equiv \text{Rayleigh-Jeans } (RJ) \text{ limit.}$$

$$(2\pi hc^2 = 3.74177118\times10^{-5}; 2\pi kc = 2.6006643\times10^{-5})$$

T_{sys} and T_N are defined for 1 second integration time ($\propto T\times t^{-1/2}$).

Nonlinear fitting

In general, fitting is nonlinear: $\chi^2 = \sum_{i=1}^{n}[y_i - F(x_i,\bar{a})]^2$ (suppress σ_i for now; it can always be re-introduced).

$\frac{\partial\chi^2}{\partial a_j} = -2\sum_{i=1}^{n}(y_i - F(x_i,\bar{a}))\frac{\partial F(x_i,\bar{a})}{\partial a_j} = 0,\ j = 1,\cdots m.$ If linear, $\frac{\partial F(x_i,\bar{a})}{\partial a_j} = X_j$. Otherwise, $\frac{\partial F(x_i,\bar{a})}{\partial a_j}$ may be calculated analytically sometimes (*e.g.*, with some Hamiltonians in spectroscopic analysis, and note LIDORT radiative transfer model, where Jacobian of the intensity field is determined analytically), but usually not.

For convenience, $\beta_k \equiv -\frac{1}{2}\frac{\partial\chi^2}{\partial a_k} = \sum_{i=1}^{n}(y_i - F(x_i,\bar{a}))\frac{\partial F(x_i,\bar{a})}{\partial a_k},$ and also (for later use)

$$\frac{\partial^2\chi^2}{\partial a_j \partial a_k} = 2\sum_i\left[\frac{\partial F(x_i,\bar{a})}{\partial a_j}\frac{\partial F(x_i,\bar{a})}{\partial a_k} - \underset{\searrow\,\simeq 0}{(y_i - F(x_i,\bar{a}))}\underset{\searrow\,\text{unstable}}{\frac{\partial^2 F(x_i,\bar{a})}{\partial a_j \partial a_k}}\right].$$

Note that we are going to discard the 2nd order term before proceeding further. The justifications are that a solution involving first derivatives should be valid for fitting an arbitrary function near the χ^2 minimum and that $yi - F(x_i,\bar{a})$ should be near zero (or average to near zero) near the minimum, and should average out for a precise model, thus allowing us to avoid the necessity to calculate second derivatives (and also to avoid the instabilities they can generate if there are significant outliers or if the model does not precisely fit the data – see *Numerical Recipes* for details). Also, *Bevington and Robinson* note that it is “convenient to use a first order approximation for fitting nonlinear functions.”

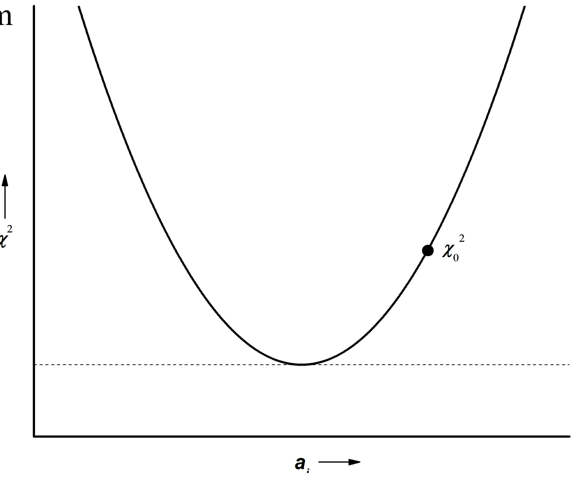
Then, $\frac{\partial^2\chi^2}{\partial a_j \partial a_k} \simeq 2\sum_i \frac{\partial F(x_i,\bar{a})}{\partial a_j}\frac{\partial F(x_i,\bar{a})}{\partial a_k} =$ the Hessian matrix

As before, $\alpha = \sum_i \frac{\partial F(x_i,\bar{a})}{\partial a_j}\frac{\partial F(x_i,\bar{a})}{\partial a_k} \simeq \frac{1}{2}\frac{\partial^2\chi^2}{\partial a_j \partial a_k}.$

Consider how χ^2 will vary near the minimum for one of the set of parameters by expanding in a Taylor series about a point near the minimum:

$$\chi^2 = \chi_0^2 + \frac{\partial\chi_0^2}{\partial a_i}\delta a_i + \frac{1}{2}\frac{\partial^2\chi_0^2}{\partial a_i^2}\delta a_i^2 + \cdots.$$

At the minimum, $\frac{\partial\chi^2}{\partial a_i} = 0$, and thus χ^2 is approximately a quadratic function in the parameter.



This shows one dimension of the *n*-dimensional minimization for *j* = 1, ⋯, *m*.

Now expand χ^2 for all parameters in a Taylor series about a starting point χ_0^2 (= χ^2 of the starting guess on parameters $\bar{a}$):

$$\chi^2 = \chi_0^2 + \sum_j \frac{\partial\chi_0^2}{\partial a_j}\delta a_j + \frac{1}{2}\sum_k\sum_j \frac{\partial^2\chi_0^2}{\partial a_k \partial a_j}\delta a_j \delta a_k + \cdots.$$

At the minimum, the first derivatives are zero:

$$\frac{\partial\chi^2}{\partial(\delta a_k)} = \frac{\partial\chi_0^2}{\partial a_k} + \sum_j \frac{\partial^2\chi_0^2}{\partial a_k \partial a_j}\delta a_j = 0, k = 1,\cdots m.$$

This is now analogous to our linear problem of before, now linearized in $\delta\overline{a}$:

$\overline{\beta} = \delta\overline{a}\underline{\alpha}, \quad \delta\overline{a} = \overline{\beta}\underline{\alpha}^{-1}$. (really β_0, α_0)

Again, $\underline{C} = \underline{\alpha}^{-1} = \left[\frac{\partial^2\chi}{\partial a_k \partial a_j}\right]^{-1}$ (covariance of the standard errors, inverse of the Hessian of χ^2), the correlation matrix is $c_{ij} / \sqrt{c_{ii}c_{jj}}$ as before and the uncertainties of the parameters are given as $\sigma(\delta a_j) = \sqrt{c_{jj}}$, but with some caveats (*cf. Numerical Recipes*), including the correlation among parameters, and departures from assumptions of normally-distributed (Gaussian) errors.

Aside on correlated parameters

The correlation matrix describes how *entangled* parameters are. Even with a "perfect" model, the fact that the measurements have noise will cause parameters to be correlated, particularly when they are physically related (*e.g.*, for O_3 versus height). Negative correlation (the most common type) means that an increase in parameter a_i will be partially offset by a decrease in parameter a_j.

Consider a case where atmospheric ozone measurements are fitted to a model with 11 layers, 3 in the troposphere (1-3) and 8 in the stratosphere (4-11). If a_i and a_j are adjacent parameters (or even if they are not adjacent) denoting ozone amounts, with uncertainties σ_i and σ_j, then $a_i + a_j$ has uncertainty $\sigma_{i+j} = \left[\sigma_i^2 + \sigma_j^2 + 2cor_{ij}\sigma_i\sigma_j\right]^{1/2}$, where cor_{ij} is the off-diagonal term of the (symmetric) correlation matrix. cor_{ij} would normally be negative, so that the uncertainty for the sum of the ozone in the two layers would be less than the *RSS* of the corresponding layer uncertainties. To put it more simply, in terms of the covariance matrix, $\sigma_{i+j} = [c_{ii} + c_{jj} + 2c_{ij}]^{1/2}$.

The uncertainty for the tropospheric ozone is

$$\sigma_{trop} = \left[\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} cor_{ij}\sigma_i\sigma_j\right]^{1/2}, n = 3.$$

The uncertainty for the stratospheric ozone is

$$\sigma_{strat} = \left[\sum_{i=4}^{n}\sigma_i^2 + \sum_{i=4}^{n-1}\sum_{j=i+1}^{n} cor_{ij}\sigma_i\sigma_j\right]^{1/2}, n = 11.$$

The uncertainty for the total ozone is

$$\sigma_{total} = \left[\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} cor_{ij}\sigma_i\sigma_j\right]^{1/2}, n = 11.$$

Back to fitting: We showed a case where we started close enough to the multi-dimensional minimum to solve by a linear expansion. We would be done except for the pesky problem of finding the minimum efficiently. Numerous methods, such as grid searches, *etc*., exist for doing so. For many nonlinear problems the *Levenberg-Marquardt* method is a standard and generally useful approach. It provides an elegant way to approach the solution (the minimum) quickly when the starting guess is far away and gently when the minimum is being approached.

The Levenberg-Marquardt method

If we are far from the solution we want to travel in the direction opposite the *gradient* (*i.e.*, in the direction of the steepest descent). When we get near, we would like to switch over to moving along the curvature (as above), linearizing the solution.

Gradient search: The gradient vector is $\nabla\chi^2 = \sum_{j=1}^{n} \frac{\partial \chi^2}{\partial a_j}\hat{a}_j$, $\hat{a}_j =$ a unit vector in the direction of a_j. Take a step in the direction of steepest descent, *i.e.*, $-\nabla\chi^2$, then re-calculate the gradient (perhaps using the Hessian to do so).

Remember $\beta_k = -\frac{1}{2}\frac{\partial \chi^2}{\partial a_k}$, so that δa_k (the step) = constant $\times \beta_k$. But what is the constant? How do we choose it? This is the clever part of the Levenberg-Marquardt method: β_k has dimension $1/a_k$, so the constant must have dimension a_k^2. So far, only $1/\alpha_{kk}$ has dimension a_k^2. So, choose a step $\delta a_k = \frac{\beta_k}{\lambda \alpha_{kk}}$, or $\beta_k = \lambda \alpha_{kk} \delta a_k$. λ is an adjustable parameter introduced to modulate the α_{kk} scale. The following change, employing an adjustable λ, allows us to vary continuously between a gradient search (steepest descent) and a linearized solution as the minimum is approached:

Instead of $\bar{\beta} = \delta \bar{a}\underline{\alpha}$, choose $\bar{\beta} = \delta \bar{a}\underline{\alpha}'$, where $\alpha'_{jk} = \begin{cases} \alpha_{jk}(1+\lambda), & j = k \\ \alpha_{jk}, & j \neq k \end{cases}$

Large $\lambda \Rightarrow$ steepest descent ($\underline{\alpha}$ is diagonally dominant)
Small $\lambda \Rightarrow$ linearized (Newton's method)

The recipe: Start with $\lambda = 0.001$ (for historical reasons) and starting parameters $\bar{a}$.

1. Compute $\chi^2(\bar{a})$
2. $\delta\bar{a} = \beta(\bar{\alpha}')^{-1}$, compute $\chi^2(a+\delta a)$
3. If $\chi^2(a+\delta a) > \chi^2(a) \Rightarrow \lambda = \lambda \times 10$
   If $\chi^2(a+\delta a) < \chi^2(a) \Rightarrow \lambda = \lambda/10, a = a + \delta a$

Convergence:

1. Preset minimum in χ^2 (sometimes referred to as the "only" way)

2. Relative change in all parameters < preset
3. Maximum iterations

After convergence, set $\lambda = 0$ and calculate $\underline{C} = \underline{\alpha}^{-1}$.

A very nice version of a similar method, with lots of bells and whistles comes from CERN: **elsunc.lc, elsunc.f90** (available on the website).

Caveats: There can be local minima that confuse the solution and broad minima that make convergence slow. There are cases when parameters may be close to degenerate (as in the ozone case mentioned above) where parameters are strongly correlated and where the interplay among parameters slows conversion.

More on retrieval theory

Optimal estimation (and much other retrieval theory, see C. Rodgers references for details) is often derived in terms of:

Weighting functions

$$K_{ij} = \frac{\partial F_i}{\partial a_j}$$

The K_{ij} give a broad idea of information content. They show the part of the atmospheric profile (*e.g.*) that is represented by each measurement. Remember that

$$\chi^2 = \sum_i (y_i - F(x_i))^2,$$

$F(x_i) = \sum_{k=1}^{m} a_k X_k(i)$ (in the linear case). Then, $K_{ij} = X_j(i)$.

Example: SBUV weighting functions

- 10 spectral bands (albedo)
- 1 total ozone construct

Contribution functions

$\underline{D}_y = \frac{\partial \hat{a}}{\partial \bar{y}}$; $\underline{D}_a = \frac{\partial \hat{a}}{\partial \bar{a}^0}$ $\hat{a} \equiv$ final parameters; $\bar{a}^0$: *a priori* parameters

These are sensitivities of the solution vector $\hat{a}$ to the measurements (y) and the *a priori* information $(\bar{a}^0)$. They are normally calculated *after* the solution, to provide a diagnostic.

Averaging kernels

$\underline{A} = \underline{D}_y \underline{K} = \frac{\partial \hat{a}}{dy}\frac{\partial F}{\partial a} = \frac{\partial \hat{a}}{\partial a}$, the way the solution changes, given changes in the atmosphere.

"Each channel contributes in a complicated way to the overall retrieval."

Gives an estimate of the vertical resolution in the case of SBUV retrievals, for example.

Compare with a δ-function or “bump” analysis.

Often (as in the case of several instruments on the NASA EOS satellites) it has become common to use Twomey-Tikhonov/Phillips-Tikhonov *regularization* and do Optimal Estimation-type diagnostics at the end, *i.e.*, $\underline{D}_y$ and $\underline{A}$ at the linearization point.

Why constrain the solution? (*i.e.*, why do regularization?) Measurement noise may easily be amplified in the retrieval process – especially in the inversion of the $\underline{\alpha}$ (curvature) matrix:

$$\bar{a} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\bar{b} \quad \begin{aligned} A_{ij} &= X_j(x_i)/\sigma_i \\ b_i &= y_i/\sigma_i \end{aligned}$$ at the linearization point.

Twomey-Tikhonov regularization

The linear solution from before was $\bar{a} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\bar{b}$.

It can be smoothed by $\bar{a} = (\underline{A}^T\underline{A} + \gamma\underline{H})^{-1}\underline{A}^T\bar{b}$. γ is an adjustable parameter, $\underline{H}$ is a square matrix (*e.g.*, $\underline{H} = \underline{I}$). The purpose of introducing this smoothing contribution is to decrease noise sensitivity. Common choices:

Squared 1st differences, $(f_{n+1} - f_n)^2$, is accomplished by

$$\underline{H} = \begin{bmatrix} 1 & -1 & & & & & & & \\ -1 & 2 & -1 & & & & & & \\ & -1 & 2 & g & & & & & \\ & & g & g & g & & & & \\ & & & g & g & g & & & \\ & & & & g & g & g & & \\ & & & & & g & 2 & -1 & \\ & & & & & & -1 & 2 & -1 \\ & & & & & & & -1 & 1 \end{bmatrix}.$$

This will smooth out differences in $\bar{a}$ (zig-zagging of solution).

Squared 2nd differences, $(\Delta^2 f_n)^2 = (f_{n+2} - 2f_{n+1} + f_n)^2$ is accomplished by

$$\underline{H} = \begin{bmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & g & & & & & \\ & 1 & g & g & g & & & & \\ & & & g & g & g & & & \\ & & & & g & g & g & 1 & \\ & & & & & g & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{bmatrix}.$$

This smoothes "2nd derivatives" in the solution vector.

Outline of several other important methods

First, develop the matrix version of the cost function

$$\chi^2 = \sum_i \left[\frac{y_i - F(x_i, \bar{a})}{\sigma_i} \right]^2 = (y_i - F(x_i, \bar{a}))^T \underline{S}_y^{-1} (y_i - F(x_i, \bar{a})).$$

$\underline{S}_y$ is the *measurement error covariance matrix*, $\underline{S}_y(i,j) = \sigma_i \sigma_j, \sigma_i \sigma_j = 0, i \neq j$ for uncorrelated uncertainties, a common assumption for measurements.

$$\frac{\partial \chi^2}{\partial \bar{a}} = 0 = -2 \left[\frac{\partial F(x_i, \bar{a})}{\partial \bar{a}} \right]^T \underline{S}_y^{-1} (y_i - F(x_i, \bar{a})).$$

Optimal Estimation solution

$$\left[\frac{\partial F(x_i, \bar{a})}{\partial \bar{a}} \right]^T \underline{S}_y^{-1} (y_i - F(x_i, \bar{a})) = 0 \Rightarrow \left[\frac{\partial F(x_i, \bar{a})}{\partial \bar{a}} \right]^T \underline{S}_y^{-1} (y_i - F(x_i, \bar{a})) + \underline{S}_a^{-1} (\bar{a} - \bar{a}^0) = 0.$$

We have added a vector of *a priori* parameters $\bar{a}^0$ and their covariance $\underline{S}_a^{-1}$. Then, if there is no correlation among *a priori* values (index *j*) or measurements (index *i*), the covariance matrices are diagonal and

$\chi^2 = \sum_j \left(\frac{\bar{a}_j - \bar{a}_j^0}{\sigma_j} \right)^2 + \sum_i \left(\frac{y_i - F(x_i, \bar{a})}{\sigma_i} \right)^2$. However, there usually is correlation. The *a priori* values are treated as data. They act to constrain the solution based upon what we know about the problem, *e.g*. ozone values from a climatology. The trick is to estimate $\underline{S}_a$, which is to say, how confident are we about how well we know the *a priori* so that it can be appropriately weighted in the solution. A typical form for $\underline{S}_a$ is

$S_a(i,i) = \sigma_a^2(i); S_a(i,j) = \sigma_a(i)\sigma_a(j)\exp - \left[\frac{z_i - z_j}{h} \right]^2, i \neq j,$ h is the c*orrelation length*.

Then, proceed to develop about a linearization point – if the problem is nonlinear, estimate and then re-linearize as before. Upon taking a step in parameter space, say from $\bar{a}_k$ to $\bar{a}_{k+1}$, χ^2 is re-evaluated as

$\chi^2 = \left\| \underline{\underline{S}}_a^{-1/2} (\bar{a}_{k+1} - \bar{a}^0) \right\|_2^2 + \left\| \underline{\underline{S}}_y^{-1/2} \{ \underline{\underline{K}}_k (\bar{a}_{k+1} - \bar{a}_k) - [\bar{y} - \bar{F}(\bar{x}, \bar{a}_k)] \} \right\|_2^2$, where the notation implies summing the squares of the diagonal elements. Upon convergence, the solution has covariance $\underline{\underline{C}} = (\underline{\underline{S}}_a^{-1} + K^T S_y^{-1} K)^{-1}$, and $\chi^2 = \left\| \underline{\underline{S}}_a^{-1/2} (\bar{a} - \bar{a}^0) \right\|_2^2 + \left\| \underline{\underline{S}}_y^{-1/2} [\bar{y} - \bar{F}(\bar{x}, \bar{a})] \right\|_2^2$.

In more standard notation, for a step in parameter space

$\chi^2 = \left\| \mathbf{S}_{\mathbf{a}}^{-1/2} (\mathbf{a}_{\mathbf{k+1}} - \mathbf{a}_{\mathbf{0}}) \right\|_2^2 + \left\| \mathbf{S}_{\mathbf{y}}^{-1/2} \{ \mathbf{K}_{\mathbf{k}} (\mathbf{a}_{\mathbf{k+1}} - \mathbf{a}_{\mathbf{k}}) - [\mathbf{Y} - \mathbf{F}(\mathbf{a}_{\mathbf{k}})] \right\|_2^2$, and, upon convergence, the solution has covariance $\mathbf{C} = (\mathbf{S}_{\mathbf{a}}^{-1} + \mathbf{K}^{\mathbf{T}} \mathbf{S}_{\mathbf{y}}^{-1} \mathbf{K})^{-1}$, and

$\chi^2 = \left\| \mathbf{S}_{\mathbf{a}}^{-1/2} (\mathbf{a} - \mathbf{a}_{\mathbf{0}}) \right\|_2^2 + \left\| \mathbf{S}_{\mathbf{y}}^{-1/2} [\mathbf{Y} - \mathbf{F}(\mathbf{a})] \right\|_2^2$.

Other methods: Onion-peeling, global fitting – for limb.