χ^2 Minimization with Correlated Errors and Principle Components Analysis

F. Masci, June 2005

Adapted from the appendix of a draft of a paper by F. Masci et al. 2005:
http://web.ipac.caltech.edu/staff/fmasci/home/paper8R.pdf

Here we present the method used to estimate the parameters and uncertainties of our *linear* model: (α, β) where $y = \alpha + \beta x$ (see eqs. [38] and [39]). As discussed in § 5.1, a determination of these parameters by directly minimizing the χ^2 function in equation (40) is not possible due to the *quasi-singular* nature of the covariance matrix $\mathbf{C}$. The mathematics of linear fits with correlated data points is not new. In fact, Fisher et al. (1994) and Bernstein (1994) have discussed this in the context of fitting models to the two-point correlation function with "bootstrap" derived covariance matrices which in general could be singular, or close to it. We expand on the methods presented therein below.

First, we recast equation (40) in matrix form:

$$\chi^2(\alpha, \beta) = (\tilde{y} - \tilde{y}_m)^T \, \mathbf{C}^{-1} \, (\tilde{y} - \tilde{y}_m) \,, \tag{B1}$$

where $\tilde{y}$ is a column vector of the data $y_i = y_1 \ldots y_N$, and $\tilde{y}_m$ the corresponding vector of "expected" model values $y_{m_i} = \alpha + \beta x_i$. The first step of *principal component analysis* (PCA; e.g., Kendall 1980) is to find a set of linear combinations of the measured values y_i which are linearly independent. In other words, the goal is to find a new "basis" (coordinate system) in which correlations are non-existent. This amounts to finding a matrix $\mathbf{M}$ that diagonalizes the covariance matrix $\mathbf{C}$:

$$\mathbf{M}^T \, \mathbf{C} \, \mathbf{M} \, = \, \mathbf{D}, \tag{B2}$$

where $\mathbf{D}$ is diagonal. In particular, the symmetry of $\mathbf{C}$ ($\equiv \mathbf{C}^T$) guarantees the existence of a diagonalizing matrix $\mathbf{M}$ whose columns form an *orthogonal* set of N eigenvectors for $\mathbf{C}$ with corresponding eigenvalues along the diagonal of $\mathbf{D}$. Once $\mathbf{C}$ is diagonolized, the power of

PCA is in the second step where we select only those eigenvectors (basis components) which are the most stable, or have relatively large eigenvalues to ensure stability in χ^2. This will become more apparent below.

The specific procedure is as follows. If there are N elements in $\tilde{y}$, we form the $N \times M$ matrix $\mathbf{M}$ from the M stable eigenvectors of $\mathbf{C}$. We then define a new dataset (of M "observables") and corresponding expected (model) values in this new basis,

$$\tilde{z} = \mathbf{M}^T \tilde{y}; \tag{B3}$$

and

$$\tilde{z}_m = \mathbf{M}^T \tilde{y}_m, \tag{B4}$$

respectively, where the new components are guaranteed to be linearly independent with a covariance matrix which is diagonal (i.e., with vanishing covariances) given by $\mathbf{D}$ (eq. [B2]). We can now define a new "simplified" χ^2 function,

$$\chi^2(\alpha, \beta) = (\tilde{z} - \tilde{z}_m)^T \mathbf{D}^{-1} (\tilde{z} - \tilde{z}_m). \tag{B5}$$

As discussed above, stability in χ^2 is ensured by selecting only those eigenvectors of $\mathbf{C}$ with the largest eigenvalues. The eigenvalues (diagonal elements of $\mathbf{D}$) are also actually the variances for the new dataset $\tilde{z}$, $\sigma_i^2 \equiv D_{ii}$. In other words, $\mathbf{D}^{-1}$ in equation (B5) is none other than

$$\begin{pmatrix} \frac{1}{\sigma^2(z_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma^2(z_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^2(z_M)} \end{pmatrix}, \tag{B6}$$

and on expanding equation (B5), the χ^2 reduces to the simple definition:

$$\chi^2(\alpha, \beta) = \sum_{i=1}^{M} \frac{[z_i - z_{m_i}(\alpha, \beta)]^2}{\sigma^2(z_i)}. \tag{B7}$$

By selecting the largest eigenvalues (variances), we therefore avoid erroneously inflating χ^2 and making it unstable against changes in α and β.

Since the model for $y_k = \log \bar{w}(\theta_k)$ is linear (eq. [38]), the minimization of χ^2 is analytic. From equations (B3) and (B4), components of the new data vectors $\tilde{z}$ and $\tilde{z}_m$ can be expanded as:

$$z_i = \sum_{k=1}^{N} M_{ki} y_k, \tag{B8}$$

and

$$\begin{aligned} z_{m_i} &= \sum_{k=1}^{N} M_{ki} y_{m_k} \\ &= \alpha \sum_{k=1}^{N} M_{ki} + \beta \sum_{k=1}^{N} M_{ki} x_k \\ &\equiv \alpha u_i + \beta v_i, \end{aligned} \tag{B9}$$

respectively, where the M_{ki} represent the matrix elements of the diagonalizing matrix $\mathbf{M}$ and $i = 1 \ldots M$. We use u_i, v_i as shorthand notation to represent the coefficient sums of α and β. Applying the minimization conditions to equation (B7): $\partial\chi^2/\partial\alpha = 0$; $\partial\chi^2/\partial\beta = 0$, and solving the simultaneous system for α and β, we have

$$\begin{aligned} \alpha &= \frac{S_1 S_5 - S_3 S_4}{S_2 S_5 - S_3^2}; \\ \beta &= \frac{S_2 S_4 - S_1 S_3}{S_2 S_5 - S_3^2}; \\ \sigma_\alpha^2 &= \frac{S_5}{S_2 S_5 - S_3^2}; \\ \sigma_\beta^2 &= \frac{S_2}{S_2 S_5 - S_3^2}; \\ \mathrm{cov}(\alpha, \beta) &= \frac{-S_3}{S_2 S_5 - S_3^2}, \end{aligned} \tag{B10}$$

where the variances and covariance follow from the inverse of the coefficient matrix representing the simultaneous system with determinant $S_2 S_5 - S_3^2$. The S_n are defined by

$$\begin{aligned} S_1 &= \sum_{i=1}^{M} D_{ii}^{-1} u_i z_i \\ S_2 &= \sum_{i=1}^{M} D_{ii}^{-1} u_i^2 \\ S_3 &= \sum_{i=1}^{M} D_{ii}^{-1} u_i v_i \\ S_4 &= \sum_{i=1}^{M} D_{ii}^{-1} v_i z_i \\ S_5 &= \sum_{i=1}^{M} D_{ii}^{-1} v_i^2, \end{aligned} \tag{B11}$$

where u_i, v_i, and z_i were defined in equations (B8) and (B9) and as discussed above, $D_{ii} \equiv \sigma^2(z_i)$ (the diagonal elements of $\mathbf{D}$).

Note that the above χ^2 minimization procedure (eq. [B7]) does not depend in any way on the errors in the y_i (or residuals $z_i - z_{m_i}$) being Gaussian. It is only important if one wants to treat the χ^2 function as having a *true* χ^2-distribution (with two degrees of freedom) with a likelihood $L = \exp(-\chi^2/2)$ for the purpose of setting confidence limits on model parameters, or, to assign a probability for the goodness of fit. Nonetheless, the central limit theorem ensures that the underlying distribution of errors in the original $\bar{w}_2(\Theta_i)$ will be approximately Gaussian. Assuming that the $\sigma[\bar{w}_2(\Theta_i)]$ are Gaussian, what about the errors in $y_i = \log \bar{w}_2(\Theta_i)$, or, linear combinations thereof (eq. [B8]) used to define the χ^2 in equation (B7])?

We find that the normalized PDF for the logarithm of a Gaussian random variable $\bar{w}_{2_i} = \bar{w}_2(\Theta_i)$, can be written:

$$\begin{aligned} P(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2(\bar{w}_{2_i})}} \exp\left[y_i - \frac{1}{\sigma^2(\bar{w}_{2_i})}(e^{y_i} - \langle\bar{w}_{2_i}\rangle)^2\right], \\ &= \frac{1}{\sqrt{2\pi\sigma^2(\bar{w}_{2_i})}} \bar{w}_{2_i} \exp\left[-\frac{1}{\sigma^2(\bar{w}_{2_i})}(\bar{w}_{2_i} - \langle\bar{w}_{2_i}\rangle)^2\right], \\ &\approx \langle\bar{w}_{2_i}\rangle\, G(\bar{w}_{2_i}) \;\text{ for }\; \frac{\sigma(\bar{w}_{2_i})}{\bar{w}_{2_i}} \ll 1, \end{aligned} \tag{B22}$$

where the $\langle\bar{w}_{2_i}\rangle$, is a boot-strapped ensemble average of some $\bar{w}_{2_i}$ with variance $\sigma^2(\bar{w}_{2_i})$. The last expression follows from the substitution $y_i = \log \bar{w}_{2_i}$ where $G(\bar{w}_{2_i})$ is the Gaussian form, and the approximation in the last step is valid when the $\bar{w}_{2_i}$ are close to their ensemble averages (i.e., the variance is small). At all angular separations, we typically have $\sigma(\bar{w}_{2_i})/\bar{w}_{2_i} \lesssim 0.05$, so that this approximation is justified to reasonable accuracy. In fact, we find that the approximation $P(y_i) \approx \langle\bar{w}_{2_i}\rangle\, G(\bar{w}_{2_i})$ differs from the true expression for $P(y_i)$ (first line in eq. [B22]) by at most 10% in the tails, i.e., at $\gtrsim 3\sigma(\bar{w}_{2_i})$. Since any linear combination of a Gaussian random variable is also Gaussian (e.g., the z_i in eq. [B8]), we can then treat the simple χ^2 function in equation (B7) as a *true* χ^2-distribution for the purpose of computing goodness-of-fit probabilities and assigning confidence limits.

The quantity in equation (B7) is then distributed like χ^2 with $\nu = M - 2$ degrees of freedom about it's minimum value. The absolute goodness of fit is given by the probability $Q(\chi^2_{min}|\nu)$ that some measured realization of the data, z_i, will yield a χ^2 which exceeds the observed value χ^2_{min} by chance (i.e., expected on the basis of random Gaussian fluctuations alone). This probability is given by the incomplete gamma function (e.g., Press et al. 1999, p. 216):

$$Q(\chi^2_{min}|\nu) = \frac{1}{\Gamma(\nu/2)} \int_{\chi^2_{min}/2}^{\infty} \mathrm{e}^{-t} t^{\nu/2-1} \mathrm{d}t \tag{B23}$$

For the full sample all subsamples, we have the range $\nu = 7 - 10$ (or $M = 9 - 12$ principle components; see above) with $\chi^2_{min} \simeq 5.8 - 12.6$ respectively. This range corresponds to $Q \simeq 0.56 - 0.24$, indicating that our power-law models for $\bar{w}_2(\theta)$ are an adequate representation of the data. Also, the values of χ^2_{min} are in good agreement with those expected from the χ^2 statistic, i.e., $\langle\chi^2_{min}\rangle \approx \nu$, and within standard deviation $\sqrt{2\nu}$.