Bayesian Model Selection (for FCS)

Jan Krieger

June 10, 2015

Contents

1 Standard FCS data-fitting

1.1 Least-Squares Fit

In standard FCS-Fitting the following least-squares problem is solved:

$$\vec{\beta}^*_{M_g} = \arg\min_{\vec{\beta}} \underbrace{\sum_{i=1}^{n} \left| \frac{g(\hat{\tau}_i;\vec{\beta}) - \hat{g}_i}{\hat{\sigma}_i} \right|^2}_{=:\chi^2(\vec{\beta})}, \tag{1}$$

where $g(\tau;\vec{\beta})$ is the fit model (later also denoted as M_g to enumerate different models) and $\vec{\beta} \in \mathbb{R}^k$ are the k parameters of the model function. The measurement consists of n value-pairs $(\hat{\tau}_i, \hat{g}_i)$ and errors for each datapoint $\hat{\sigma}_i$ that are used to weight the fits.

1.2 Least-Squares Fit Algorithms (Basics)

The least-squares fit problem (1) can be formulated in a more compact form by writing the objective function $\chi^2(\vec{p})$ in the following way:

$$\vec{\beta}^*_{M_g} = \arg\min_{\vec{\beta}} \ \chi^2(\vec{\beta}) = \arg\min_{\vec{\beta}} \ \left\| \vec{F}(\vec{\beta}) \right\|_2^2 = \arg\min_{\vec{\beta}} \ \sum_{i=1}^{n} \left[f_i(\beta_1, \beta_2, ..., \beta_k) \right]^2 \tag{2}$$

Most numerical fit algorithms, that solve this problem, start from an initial guess $\vec{\beta}_0$ and then find the best-fit parameters by proceeding along the steepest descent of χ^2. This is done in an iterative way by solving this linearized problem in each step (Gauss-Newton iteration, probably with additional conditioning, such as in the LM-fit):

$$\vec{\beta}_{i+1} = \vec{\beta}_i + \vec{\delta}_{i+1} = \vec{\beta}_i + \arg\min_{\vec{\delta}} \ \left\| \vec{F}(\vec{\beta}_i) + \mathbf{J}(\vec{\beta}_i)\vec{\delta} \right\|_2^2. \tag{3}$$

Here $\mathbf{J}(\vec{\beta})$ is the jacobi matrix, i.e. the matrix of first derivatives:

$$J_{\nu,\kappa}(\vec{\beta}) = \left. \frac{\partial f_\nu}{\partial p_\kappa} \right|_{\vec{\beta}} = \frac{1}{\hat{\sigma}_\nu} \cdot \left. \frac{\partial g(\hat{\tau}_\nu;\vec{\beta})}{\partial p_\kappa} \right|_{\vec{\beta}} \tag{4}$$

The problem (3) is a linear system of equations, which can be solved as follows by normal equations, which follow from requiring that the gradient of χ^2 equals zero:

$$\begin{aligned} \vec{F}(\vec{\beta}_i) + \mathbf{J}(\vec{\beta}_i)\vec{\delta} \stackrel{!}{=} 0 \qquad &\Leftrightarrow \qquad -\mathbf{J}(\vec{\beta}_i)^{\mathrm{T}} \vec{F}(\vec{\beta}_i) \stackrel{!}{=} \left[\mathbf{J}(\vec{\beta}_i)^{\mathrm{T}} \mathbf{J}(\vec{\beta}_i) \right] \vec{\delta} \\ &\Leftrightarrow \qquad \vec{\delta} = - \left[\mathbf{J}(\vec{\beta}_i)^{\mathrm{T}} \mathbf{J}(\vec{\beta}_i) \right]^{-1} \mathbf{J}(\vec{\beta}_i)^{\mathrm{T}} \vec{F}(\vec{\beta}_i) \end{aligned} \tag{5}$$

1.3 Fit parameter errors/variance-covariance matrix

The next question that arises is, how accurate do we know the parameters in the best-fit parameter vector $\vec{\beta}^*$? To solve this problem, we look again at the least squares problem (1) and write it in terms of data vectors $\hat{\vec{g}} = [\hat{g}_1, \hat{g}_2, ...]^{\mathrm{T}}$ and a vector-valued fit function $\vec{g}(\vec{\beta}) = [g(\hat{\tau}_1; \vec{\beta}), g(\hat{\tau}_2; \vec{\beta}), ...]^{\mathrm{T}}$. If we omit the weights $\hat{\sigma}_i$, we can write for the ideal case of a perfect fit:

$$\hat{\vec{g}} = \vec{g}(\vec{\beta}) \tag{6}$$

Now we have small changes $\vec{\epsilon}$ of the data around the ideal values $\hat{\vec{g}}$. Since these fluctuations are small, it should be possible to account for them by a first-order Taylor approximation of the fit function $\vec{g}(\vec{\beta})$ and therefore a small (linear) variation of the best fit parameters $\vec{\beta}$:

$$\hat{\vec{g}} + \vec{\epsilon} = \vec{g}(\vec{\beta}) + \mathbf{J}\delta\vec{\beta} \tag{7}$$

Using (6) this can be rewritten and solved for $\delta\vec{\beta}$ with the same method as in (5) (only now written in a short form):

$$\vec{\epsilon} = \mathbf{J}\delta\vec{\beta} \qquad \Rightarrow \qquad \delta\vec{\beta} = \left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1}\mathbf{J}^{\mathrm{T}}\vec{\epsilon} \equiv \mathbf{C}\vec{\epsilon} \tag{8}$$

These findings can now be used to calculate an approximation for the variance of the ideal parameters:

$$\begin{aligned}
\mathrm{Var}(\vec{\beta}) = \mathrm{Var}(\delta\vec{\beta}) = \mathrm{Var}\left(\mathbf{C}\vec{\epsilon}\right) = \mathbf{C}^{\mathrm{T}}\,\mathrm{Var}(\vec{\epsilon})\mathbf{C} = \\
= \frac{\chi^2}{n-k} \cdot \left(\left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1}\mathbf{J}^{\mathrm{T}}\right)\left(\mathbf{J}\left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1}\right) = \\
= \frac{\chi^2}{n-k} \cdot \left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1} \underbrace{\mathbf{J}^{\mathrm{T}}\mathbf{J}\left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1}}_{=\mathbb{1}} = \frac{\chi^2}{n-k} \cdot \left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1}
\end{aligned} \tag{9}$$

From the first to the second line we estimate the variance of the data variation as the variance of the residulas $\mathrm{Var}(\vec{\epsilon}) = (\chi^2/(n-k)) \cdot \mathbb{1}$.

So finally we can define the covariance matrix of the non-linear least-squares problem as:

$$\mathbf{\Sigma} = \left[\mathbf{J}^{\mathrm{T}}\mathbf{J}\right]^{-1} \tag{10}$$

and the standard error of a parameter as

$$\mathrm{err}(\beta_i) = \sqrt{\frac{\chi^2}{n-k} \cdot \mathbf{\Sigma}_{ii}}. \tag{11}$$

1.4 Likelihood function

The likelihood function for the problem (1) can be written exactly, if we assume independent errors $\hat{\sigma}_i$ for each measured point $(\hat{\tau}_i, \hat{g}_i)$ on the ACF and a Gaussian error distribution:

$$p(\hat{g}_i | M_g, \vec{\beta}) = \frac{1}{\sqrt{2\pi} \cdot \hat{\sigma}_i} \cdot \exp\left[-\frac{1}{2} \cdot \frac{\left(g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i\right)^2}{\hat{\sigma}_i^2}\right]. \tag{12}$$

This $p(\hat{\tau}_i, \hat{g}_i | M_g, \vec{\beta})$ is the conditional probability to measure the value $\hat{g}_i$ of the ACF at the (given) lag-time $\hat{\tau}_i$, given the specific FCS model function M_g and the parameter vector $\vec{\beta}$. Equation (12) can then be used to derive the likelihood function for this problem, which is basically the net probability to obtain the complete set $i = 1..n$ of measurements, given again the model M_g and the parameter vector $\vec{\beta}$:

$$p(\hat{\vec{g}} | M_g, \vec{\beta}) = \prod_{i=1}^{n} p(\hat{g}_i | M_g, \vec{\beta}) = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\prod_i \hat{\sigma}_i} \cdot \exp\left[-\sum_{i=1}^{n} \frac{1}{2} \cdot \frac{\left(g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i\right)^2}{\hat{\sigma}_i^2}\right]. \tag{13}$$

If the errors are not independent, then they are no longer described by the $\hat{\sigma}_i$, but by an $n \times n$ covariance matrix $\hat{\mathbf{C}}$ that has to be determined from the measurement, or from theoretical considerations. The likelihood is then given by:

$$p(\hat{\vec{g}} | M_g, \vec{\beta}) = \prod_{i=1}^{n} p(\hat{g}_i | M_g, \vec{\beta}) = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sqrt{\det(\hat{\mathbf{C}})}} \cdot \exp\left[-\frac{1}{2} \cdot \left[\hat{\vec{g}} - \vec{g}(\hat{\vec{\tau}}; \vec{\beta})\right]^{\mathrm{T}} \hat{\mathbf{C}}^{-1} \left[\hat{\vec{g}} - \vec{g}(\hat{\vec{\tau}}, \hat{\vec{g}}; \vec{\beta})\right]\right]. \tag{14}$$

Note: Here the single measurements i have been combined into vector, e.g. $\hat{\vec{g}} = [\hat{g}_1, \hat{g}_2, ..., \hat{g}_n]^{\mathrm{T}}$. The model is then also vector-valued with $\vec{g}(\hat{\vec{\tau}}; \vec{\beta}) = \left[g(\hat{\tau}_1; \vec{\beta}), g(\hat{\tau}_2; \vec{\beta}), ..., g(\hat{\tau}_n; \vec{\beta})\right]^{\mathrm{T}}$

Then the least-squares problem (1) can also be written as a maximum-likelihood estimate (MLE):

$$\vec{\beta}^*_{M_g} = \arg\max_{\vec{\beta}} \; p(\hat{\vec{g}}|M_g, \vec{\beta}). \tag{15}$$

Introducing the log-likelihood $\ln\big[p(\hat{\vec{g}}|M_g, \vec{\beta})\big]$, this can be rewritten in terms of sums, not products and allows to remove the exponential functions:

$$\begin{aligned}
\vec{\beta}^*_{M_g} &= \arg\max_{\vec{\beta}} \; p(\hat{\vec{g}}|M_g, \vec{\beta}) = \arg\max_{\vec{\beta}} \; \ln\big[p(\hat{\vec{g}}|M_g, \vec{\beta})\big] = \\
&= \arg\max_{\vec{\beta}} \; \left\{ -\frac{n}{2} \cdot \ln(2\pi) - \sum_i \ln\left[\hat{\sigma}_i\right] - \frac{1}{2} \cdot \sum_{i=1}^{n} \frac{\Big(g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i\Big)^2}{\hat{\sigma}_i^2} \right\} = \\
&= \arg\min_{\vec{\beta}} \; \left\{ \frac{n}{2} \cdot \ln(2\pi) + \sum_i \ln\left[\hat{\sigma}_i\right] + \frac{1}{2} \cdot \sum_{i=1}^{n} \frac{\Big(g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i\Big)^2}{\hat{\sigma}_i^2} \right\} = \\
&= \arg\min_{\vec{\beta}} \; \sum_{i=1}^{n} \frac{\Big(g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i\Big)^2}{\hat{\sigma}_i^2}
\end{aligned}$$

In the last line, constant term and factors (that do not depend on $\vec{\beta}$) were omitted and the problem was reduced again to (1), so the simple least-squares solution equals the maximum likelihood estimator (MLE) of $\vec{\beta}$!

2 Model Selection

So solving the least-squares (or equivalently the MLE) problem gives an estimate of the parameter vector $\vec{\beta}$ for a given model M_g that describes the given measured data best in a least-squares sense (note this is usually not outlier-robust!). This procedure can be repeated for any model M_g, and will result in a best-fit parameter vector $\vec{\beta}^*_{M_g}$ for each of these models, but the question remains: Which model should be taken, especially when the model selection is not obvious from external assumption or knowledge about the sample. In these cases a statistical model selection should be done.

2.1 χ^2 model selection criterion

The simplest model selection is based on the χ^2 criterion, which is simply defined as the sum of squared deviations (sometimes also called residual sum of squares, RSS):

$$\chi^2(\vec{\beta}, M_g) = \mathrm{RSS}(\vec{\beta}, M_g) = \sum_{i=1}^{n} \left| \frac{g(\hat{\tau}_i; \vec{\beta}) - \hat{g}_i}{\hat{\sigma}_i} \right|^2 \tag{16}$$

With this criterion, the decision rule is:

Use the model that has the lowest $\chi^2(\vec{\beta}, M_g)$,
i.e. is closest to the measured data.

Unfortunately this simple model is not ideal, as a more complex model M_g (i.e. with more parameters in $\vec{\beta}$ often also has a lower χ^2-value, since the added model complexity allows to describe more of the noise on the data.

2.2 Other model selection criteria

To overcome these problems, other model selection criteria [2] have been proposed, e.g. the Akaike's information criterion (AIC, [1, 2]):

$$\mathrm{AIC}(\vec{\beta}, M_g) = -2\ln\big[p(\hat{\vec{g}}|M_g, \vec{\beta})\big] + 2k, \tag{17}$$

$$\mathrm{AICc}(\vec{\beta}, M_g) = \mathrm{AIC}(\vec{\beta}, M_g) + \frac{2k \cdot (k+1)}{n-k-1} \quad \text{(corrected for small } n\text{)} \tag{18}$$

or the Bayesian information criterion (BIC, also known as Schwarz's criterion, [5, 2]):

$$\mathrm{BIC}(\vec{\beta}, M_g) = -2\ln\big[p(\hat{\vec{g}}|M_g, \vec{\beta})\big] + k \cdot \ln[n] \tag{19}$$

For both criteria, the model selection rule is:

Use the model that has the smallest (often most negative) value of AIC or BIC.

Since both criteria include the number of parameters k, they also obey (to some extent) the principle of parsimony (or Occam's razor), which states that one should use the model that best describes the data and has the lowest number of parameters/is the simplest.

For any practical purposes, both criteria can be estimated also from the χ^2, but only up to a fixed additive constant, which only depends on the dataset $(\hat{\tau}_i, \hat{g}_i, \hat{\sigma}_i)$ and therefore does not play a role for the selection process. This can be seen, if we write the log-likelihood, assuming Gaussian errors as follows:

$$\ln\left[p(\hat{\vec{g}}|M_g, \vec{\beta})\right] = \underbrace{-\frac{n}{2}\cdot\ln(2\pi) - \sum_i \ln\left[\hat{\sigma}_i\right]}_{=\text{const}} - \frac{1}{2}\cdot\sum_{i=1}^{n}\frac{\left(g(\hat{\tau}_i;\vec{\beta}) - \hat{g}_i\right)^2}{\hat{\sigma}_i^2} = \text{const} - \frac{\chi^2}{2} \tag{20}$$

So we get:

$$\text{AIC}(\vec{\beta}, M_g) = \chi^2 + 2k \qquad \text{and} \qquad \text{BIC}(\vec{\beta}, M_g) = \chi^2 + k\cdot\ln[n] \tag{21}$$

If all $\hat{\sigma}_i = \sigma$ are equal (i.e. a non-weighted fit), then the estimations in (20) change [2]:

$$\ln\left[p(\hat{\vec{g}}|M_g, \vec{\beta})\right] = \frac{n}{2}\cdot\ln(2\pi) - \sum_i \ln\left[\sigma\right] - \frac{1}{2}\cdot\sum_{i=1}^{n}\left(\frac{g(\hat{\tau}_i;\vec{\beta}) - \hat{g}_i}{\sigma}\right)^2 = \tag{22}$$

$$= -\frac{n}{2}\cdot\ln(2\pi) - \frac{n}{2}\cdot\ln\left[\sigma^2\right] - \frac{1}{2\sigma^2}\cdot\underbrace{\sum_{i=1}^{n}\left(g(\hat{\tau}_i;\vec{\beta}) - \hat{g}_i\right)^2}_{=\chi^2} = \tag{23}$$

$$= \text{const} - \frac{n}{2}\cdot\ln\left[\frac{\chi^2}{n}\right] \tag{24}$$

Here, in the last step we used the estimator χ^2/n for the sample variance σ^2. Then the last term is also constant and the AIC and BIC become:

$$\text{AIC}(\vec{\beta}, M_g) = n\cdot\ln\left[\frac{\chi^2}{n}\right] + 2k \qquad \text{and} \qquad \text{BIC}(\vec{\beta}, M_g) = n\cdot\ln\left[\frac{\chi^2}{n}\right] + k\cdot\ln[n] \tag{25}$$

2.3 Bayesian model selection

Another framework for model selection [4, 3] is based in the Bayes theorem, thus the name "Bayesian model selection". The Bayes theorem can be used to calculate the probability $p(M_g|\hat{\vec{g}})$ for a specific model M_g, given a measurement $(\hat{\vec{\tau}}, \hat{\vec{g}}, \hat{\vec{\sigma}})$:

$$p(M_g|\hat{\vec{g}}) = \frac{p(\hat{\vec{g}}|M_g)\cdot p(M_g)}{p(\hat{\vec{g}})}. \tag{26}$$

Here $p(M_g)$ is the prior (probability) of a model M_g, which allows to insert any prior/external information into the selection problem. The probability $p(\hat{\vec{g}})$ is mostly a normalization constant, as no statement about the absolute probability of a given dataset can easily be done. Finally $p(\hat{\vec{g}}|M_g)$ is the conditional probability of obtaining the measurement $\hat{\vec{g}}$, given the specified model M_g. As will be shown later. This can be calculated from the likelihood function (13).

In the special case of model selection, we often do not want to put any prior information into the problem, as we cannot say which model is more likely, a priori. Therefore we will assume a flat prior $p(M_g)$ here. As said above, the probability of the data $p(\hat{\vec{g}})$ is treated as a normalization constant, thus we can simplify (26) to:

$$p(M_g|\hat{\vec{g}}) \propto p(\hat{\vec{g}}|M_g). \tag{27}$$

We will then calculate only the probability $p(\hat{\vec{g}}|M_g)$ for each model M_g. Then the decision rule is simple:

Choose the model M_g with the highest (non-normalized) probability $p(\hat{\vec{g}}|M_g)$.

From the $p(\hat{\vec{g}}|M_g)$, it is also possible to calculate absolute model probabilities, by normalizing the $p(\hat{\vec{g}}|M_g)$ as follows:

$$p_{\text{norm}}(M_g|\hat{\vec{g}}) = \frac{p(\hat{\vec{g}}|M_g)}{\sum_i p(\hat{\vec{g}}|M_i)}, \qquad \text{i.e.} \qquad \sum_g p_{\text{norm}}(M_g|\hat{\vec{g}}) = 1. \tag{28}$$

So finally we are left with the problem of calculating $p(\hat{\vec{g}}|M_g)$. Following [4] this can be done from the likelihood function $p(\hat{\vec{g}}|M_g, \vec{\beta})$ by "integrating out" the parameters $\vec{\beta}$:

$$p(\hat{\vec{g}}|M_g, \vec{\beta}) = \int_{\vec{\beta}} p(\hat{\vec{g}}|M_g, \vec{\beta})\cdot p(\vec{\beta}|M_g)\ \mathrm{d}\vec{\beta} \tag{29}$$

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723, 1974. doi:10.1109/TAC.1974.1100705.

[2] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer, 2 edition, 2002. ISBN 0387953647.

[3] Syuan-Ming Guo, Jun He, Nilah Monnier, Guangyu Sun, Thorsten Wohland, and Mark Bathe. Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: application to simulated and in vitro data. *Analytical Chemistry*, 84(9):3880–3888, 2012. doi:10.1021/ac2034375.

[4] Jun He, Syuan-Ming Guo, and Mark Bathe. Bayesian approach to the analysis of fluorescence correlation spectroscopy data i: Theory. *Analytical Chemistry*, 84(9):3871–3879, 2012. doi:10.1021/ac2034369.

[5] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi:10.1214/aos/1176344136.