CMP 610b Lecture 8
F. Sigworth

Maximum-Likelihood fitting

One of the issues I want to address in this lecture is the fitting of distributions dwell times. We want to find the best curve to draw over a histogram, say a histogram of open times. We know how to do this pretty well, by eye. But how do we do it in an optimum way? Before we tackle this problem, let's review the three best-known distributions of random numbers.

1. The binomial distribution. Suppose you toss a coin that has a probability p of landing heads. (A fair coin would have $p = 1/2$.) The probability of obtaining exactly k heads out of N tries is

$$P(k;p,N) = \binom{N}{k} p^k (1-p)^{N-k}$$

This is the binomial distribution.

2. The Gaussian distribution. Suppose the number of tries N is very large. The expectation value for the number of heads k is $\mu = Np$ and the variance is $\sigma^2 = Np(1-p)$. The binomial distribution can be approximated by the Gaussian probability density function (the probability that

$$P(k;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(k-\mu)^2/2\sigma^2}.$$

3. The Poisson distribution. Suppose that N is very large, but p is very small, such that Np is a fairly small number, which we will call λ. Then a better approximation to the binomial distribution is the Poisson probability,

$$P(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

THE NUMBER OF EVENTS IN A HISTOGRAM BIN

The random number n_i of entries in a bin of a histogram can be thought of arising from the binomial distribution. From last time we said that the expectation value of n_i is

$$\langle n_i \rangle = N\left[F'(t_i) - F'(t_{i+1})\right]$$

where N is the total number of events, and $F'(t_i)$ is the probability that an event exceeds the value t_i corresponding to the lower end of the i^{th} bin. Strictly speaking, the variable n_i follows the binomial distribution; but since there are typically a lot of bins in a histogram, and often not so many events in a single bin, the Poisson distribution is a very good approximation. The Gaussian distribution is not a good approximation, especially in the case where there are only a few entries in a bin. This is especially true for bins that have one or no entries. How do you fit a Gaussian in these cases? Instead, we will consider an entirely different approach to fitting the experimental events, based on computing the likelihood. Maximum-likelihood fitting to histograms is described in the paper by Sigworth and Sine (1989).

THE LIKELIHOOD

When you gather some data (for example, a set of channel open times) you typically want to find the best description of these data in the form of some sort of model, perhaps a kinetic scheme. The best way to approach this problem seems to be to find the "most probable" model, that is the theory that represents the underlying mechanism most likely to give rise to the data you observed. That is, you would like to be able to evaluate the conditional probability

$$P_{\text{theory}'} = \text{Prob}\{\text{ theory } | \text{ data }\} \qquad (1)$$

and find the theory that gives the highest probability. Now this probability is very difficult to evaluate. A starting point is the probability

$$P_{\text{data}'} = \text{Prob}\{\text{ data } | \text{ theory }\}$$

This is not so hard to evaluate: based on your theory, you calculate the probability of actually observing what you observed (we will see how to do this below). By Bayes' theorem we have

$$\text{Prob}\{\text{ theory } | \text{ data }\} = \frac{\text{Prob}\{\text{ theory }\}}{\text{Prob}\{\text{ data }\}}\ \text{Prob}\{\text{ data } | \text{ theory }\}. \quad (2)$$

This means that we can in principle evaluate eqn. (1), if we somehow knew the probability that we got our particular data set (as compared to any other data set) and also how probable our theory is compared to all other theories. No one knows really how to evaluate the first of these, the probability of the data. There might be some *a priori* information, gotten from an independent source, which could be reflected in Prob{theory}. This could consist of rough values for parameters of the theory, in which case Prob{theory} might be a product of broad Gaussian functions centered on these estimates. Or Prob{theory} might reflect a philosophical bias about what kinds

of models are best (e.g. that we prefer simple models to complex ones). Either way, Prob{theory} can be incorporated as a *prior probability* term. If there is no prior probability function, one just guesses that Prob{theory} / Prob{data} is a constant, and defines the *likelihood* of a theory simply as

$$\text{Lik} = k\ \text{Prob}\{\ \text{data}\ |\ \text{theory}\ \} \qquad (3)$$

where k is an undetermined constant.

Computing the likelihood is done as follows. Suppose we had a number of measurements of channel open time which we call t_1, t_2, ... t_n. Suppose we have a theory that these times come from a distribution with probability density

$$f(t_i) = \frac{\text{Prob}\{\ t_i \text{ lies in the interval } (t,\ t+dt)\}}{dt}$$

that, say, is exponential,

$$f(t_i) = \alpha\ e^{-\alpha t}. \qquad (4)$$

Then the probability of observing these data given this theory is something like

$$\text{Prob}\{\ \text{data}\ |\ \text{theory}\ \} = f(t_1)\ dt \times f(t_2)\ dt \times \ldots \times f(t_n)\ dt$$

which is a very infinitesimal number since it contains the infinitesimal factor $(dt)^n$. However, since we have an undetermined constant anyway, we just write

$$\text{Lik} = f(t_1)\ f(t_2)\ f(t_3)\ldots f(t_n) \qquad (5)$$

i.e. the product of the pdf evaluated for each observed dwell time. Even this number can get to be very small, so it is common practice to evaluate the log likelihood,

$$L = \ln(\ \text{Lik}\ ) = \sum_{i=1}^{n} \ln f(t_i) \qquad (6)$$

which is easier to handle because we are forming a sum rather than the product of terms representing each measurement. Our goal is to find the theory that gives the greatest likelihood, so it is just as good to maximize L as to maximize Lik itself.

Returning to our example with the exponential probability density function (4), we can evaluate the log likelihood directly:

$$L = \sum_{i=1}^{n} \ln(\alpha) - \alpha t_i \, .$$

Given that we think that the exponential pdf is the 'right' theory, all we have to do is find the value of α that maximizes *L*. This can be obtained anlytically in this case by finding

$$\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} t_i = 0 \qquad (7)$$

which works out to be α = 1/(average of the t_i).

For more complicated theories the evaluation of *L* is not so easy, and typically has to be done with a computer performing the sum in eqn.(6) by brute force. Once *L* is evaluated for one set of parameters of a theory, an automatic search routine can vary the parameters to find a set that gives a maximum value. What you want is the global maximum value, but what a computer program finds may just as well turn out to be only a local maximum value; so you must be careful. See Colquhoun and Sigworth (1995) for a more complete discussion of maximum-likelihood fitting.

LEAST SQUARES IS A SPECIAL CASE OF MAXIMUM LIKELIHOOD

You are probably more familiar with least-squares fitting of data. For example, standard plotting programs will perform some kinds of least-squares fits for you. Least-squares is sometimes, but definitely not always, equivalent to maximum likelihood. We consider one case where they are equivalent.

Suppose you have some data measurements x_1, x_2, ... x_n which, like samples of current measured in a patch clamp, reflect a true underlying signal but with random noise added to it. Suppose our theory is that the sequence of data measurements reflect a particular sequence s_1, s_2, ... s_n of underlying signal samples. Suppose also that we know that the noise is Gaussian distributed. That means that the probability density function will be something like

$$f(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\left(\frac{(x_i - s_i)^2}{2\sigma^2}\right)$$

and so on for x_2, x_3,... with *k* being a factor that depends on σ but not on *x* or *y*. Now notice what the log likelihood (eqn. 12) becomes:

$$L = -n\ln(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(x_i - s_i)^2}{2\sigma^2} \qquad (8)$$

The log likelihood is just a constant minus something proportional to the sum of the squared deviations of the x_i from the y_i. Maximizing L will be equivalent to minimizing the squared deviations. This is so precisely because we assumed a Gaussian distribution for the noise. If the random variations in the measured values do not follow a Gaussian distribution, then least-squares will give you a result that is not optimum in the sense of being the "most likely" one.

HIDDEN MARKOV MODELS

A particularly ambitious use of the maximum-likelihood technique is the so-called Hidden Markov Model (HMM) analysis of single-channel data. The idea is to compute the probability of the data given the theory, in the case where the data consist of the entire sequence of raw measurements of the membrane current, and the theory is a Markov model (technically, a Hidden Markov model) for the channel gating behavior. A Hidden Markov model is a Markov process (i.e. the switching of a system among discrete states) where the observations do not unambiguously show which state the system is in at a given time. This describes the situation of single-channel recordings very well, where (1) noise sometimes precludes an unambiguous determination of whether the channel is open or closed, and (2) there often are multiple underlying kinetic states corresponding to the "channel closed" or "channel open" conductance levels.

Some very powerful algorithms have been developed for computing and maximizing the likelihood of an HMM. For more than a decade, much work has been done on HMMs for automatic speech recognition; for an introduction, see Rabiner and Juang (1986). Only recently have people noticed that HMM techniques are applicable to ion channel analysis as well (Chung et al. 1990; Chung et al. 1991). Here, we will consider the Forward Algorithm for computing the likelihood.

Suppose we have a set of observations $y_1, y_2, \ldots y_T$ which are measurements of the patch current at evenly spaced times, one observation per sample interval δt. We will compare it with a Markov model having n states. Corresponding to each state there is a channel current μ_i ; if for example state 1 is a closed state, then $\mu_1 = 0$. The background noise is taken to be Gaussian with variance σ^2. To describe the kinetics, we define transition probabilities a_{ij} which is the set of probabilities of going from state s_i to state s_j in one sample interval. (These are approximately related to the rate constants k_{ij} by $a_{ij} \approx k_{ij}\delta t$.)

On the basis of this model, we can define the probability of an observation y_t given that the channel is in state i as a Gaussian function (due to the background noise),

$$b_i(y_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(y_t - \mu_i)^2}{2\sigma^2}\right]. \qquad (9)$$

The probability of a particular *sequence* of observations, given that the underlying Markov process goes through the sequence of states $s_{i_1}, s_{i_2}, \ldots s_{i_T}$ is given by

$$P\{y_1, y_2, \ldots y_T \mid s_{i_1}, s_{i_2}, \ldots s_{i_T}\} = b_{i_1}(y_1) b_{i_2}(y_2) \ldots b_{i_T}(y_T)$$

where p_{i_1} is the probability of starting in state i_1 at the first time point. We are assuming that there is no correlation in the noise so that we can simply form the product of the observations probabilities *b*. Meanwhile, the probability of the state sequence, given the model λ, is given by

$$P\{s_{i_1}, s_{i_2}, \ldots s_{i_T} \mid \lambda\} = p_{i_1} a_{i_1 i_2} a_{i_2 i_3} \ldots a_{i_{T-1} i_T}.$$

The likelihood (i.e. the probability of the observations given λ) can therefore be obtained by a sum over all possible sequences,

$$\begin{aligned} P\{y_1, y_2, \ldots y_T \mid \lambda\} &= \sum_{i_1, i_2, \ldots i_T} P\{y_1, y_2, \ldots y_T \mid s_{i_1}, s_{i_2}, \ldots s_{i_T}\} P\{s_{i_1}, s_{i_2}, \ldots s_{i_T} \mid \lambda\} \\ &= \sum_{i_1, i_2, \ldots i_T} p_{i_1} b_{i_1}(y_1) a_{i_1 i_2} b_{i_2}(y_2) \ldots a_{i_{T-1} i_T} b_{i_T}(y_T) \end{aligned} \qquad (10)$$

Computing this sum is *very* time consuming, because there are n^T possible sequences! If we try to analyze a brief stretch of say 1000 sample points this sum will be impossible to evaluate.

A very elegant approach to evaluating the likelihood is instead to define the quantity α,

$$\alpha_t(i) = P(y_1, y_2 \ldots y_t \text{ and } s_t = i \mid \lambda) \qquad (11)$$

where λ represents the model. It can be seen that

$$\alpha_1(i) = p_i b_i(y_1).$$

Meanwhile, all successive α_i can be evaluated according to

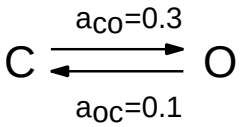
$$\alpha_t(j) = \sum_{i=1}^{n} \alpha_{t-1}(i) a_{ij} b_j(y_t) \qquad (12)$$

The desired likelihood value is obtained finally as

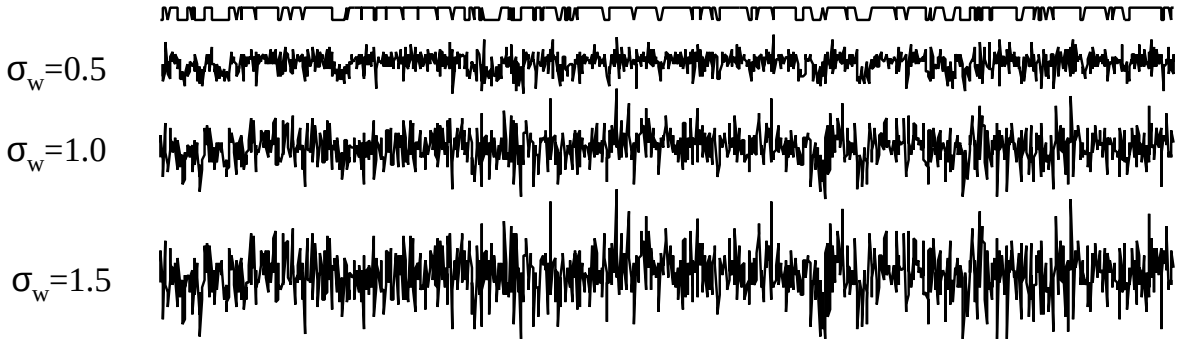
$$L = \sum_{i=1}^{n} \alpha_T(i),$$

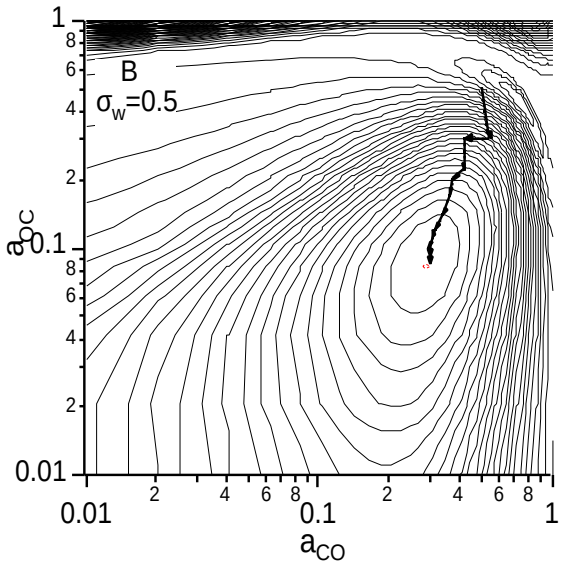
the whole process requiring something on the order of n^2T operations, instead of n^T. This makes evaluation of the likelihood practical.

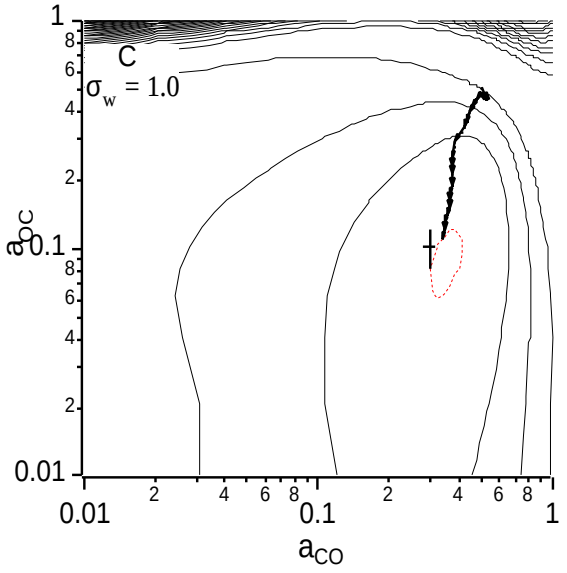
Shown below is a simulation that shows the promise of this sort of approach. A signal from a two-state channel was simulated (a total of 20,000 sample points) with various amounts of white noise added. The likelihood was computed as a function of the two transition probabilities, where the actual two-state scheme was

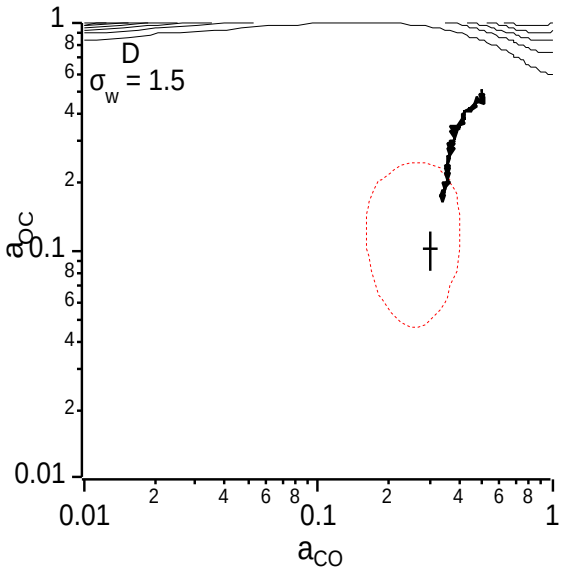


At the "low noise" level, where the noise standard deviation σ_w is half of the single-channel amplitude, the maximum of the likelihood occurs at the correct values. However, even at the extremely high noise level of $\sigma_w = 1.5$ the maximum is near the correct values. The dotted contour gives the 95% confidence interval for the estimates, which is seen to enclose (or nearly enclose) the correct values (indicated by a cross) in each contour plot.




B
σw=0.5
aOC
aCO
1
0.1
0.01
2
4
6
8


C
σw = 1.0
aOC
aCO
1
0.1
0.01
2
4
6
8



From this one can see a general approach to analyzing single-channel data: guess a model (i.e. pick the number of states, the transition probabilities and the currents), and compute the likelihood from the stretch of data. Then vary the parameters of the model until the likelihood is a maximum. This is then, the model that gives the best description of the data. There are two limitations that must be kept in mind for this approach. First, there is a danger that in finding the maximum you have found only a "local maximum". There may be a set of parameters quite different from the ones you settled on that gives a higher likelihood. This is a general problem in all situations where there are many parameters. Second, perhaps you picked the wrong sort of model in the first place. For example, suppose you picked a two-state model to describe data from a three-state channel. You won't know your error unless you happen to try a three-state model; in that case you might see a greatly increased likelihood with the *n*=3 model, but little further increase with *n*=4 or 5 models.

The simple algorithm described here has been extended to deal with various practical issues that are encountered in analyzing actual patch-clamp data, including nonwhite background noise, state-dependent noise, the effect of filtering on the data, and baseline drift. You can read all about it in the thesis by a Yale EE graduate student (Venkataramanan, 1998).

REFERENCES

Chung, S. H., V. Krishnamurthy and J. B. Moore (1991). Adaptive processing techniques based on hidden Markov models for characterizing very small channel currents buried in noise and deterministic interferences. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **334**(1271): 357-84.

Chung, S. H., J. B. Moore, L. G. Xia, L. S. Premkumar and P. W. Gage (1990). Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **329**(1254): 265-85.

Colquhoun, D. and F. J. Sigworth (1995). Fitting and statistical analysis of single-channel records. In: *Single Channel Recording*. 2nd Ed. New York, Plenum.

Rabiner, L. R. and B. H. Juang (1986). An introduction to Hidden Markov Models. *IEEE ASSP Magazine* (January): 4-16.

Sigworth, F.J. and S.M. Sine. Data transformations for improved display and fitting of single-channel dwell time histograms. Biophys. J. 52: 1047-1054 (1987).

Venkataramanan, L. (1998). *Hidden Markov Modelling of Ion Channel Currents*. Ph.D. Thesis, Yale University.
--Part of Lalitha's thesis is contained in the following two articles:

Venkataramanan, L., J. L. Walsh, R. Kuc and F. J. Sigworth. Identification of hidden Markov models for ion channel currents containing colored background noise. IEEE Transactions on Signal Processing 46: 1901-1915, 1998.

Venkataramanan, L., R. Kuc and F. J. Sigworth. Identification of hidden Markov models for ion channel currents containing state-dependent noise. IEEE Transactions on Signal Processing 46:1916-1929, 1998.