

Chapter 5

Fitting in the Age of Single-Molecule Experiments: A Guide to Maximum-Likelihood Estimation and Its Advantages



Behrouz Eslami-Mosallam, Iason Katechis and Martin Depken

5.1 Introduction

Single-molecule (SM) experiments allow us to peer deep into the molecular dynamics that drive biology at the microscopic scale [4, 14]. Though observing the dynamics of a single molecule is an amazing feat in and of itself, the information gleaned is limited by the small number of observables that can be simultaneously tracked, and the resolution at which this can be done. Faced with such limitations, mechanistic modeling and parameter estimation are often used to extract as much quantitative information as possible.

Using SM fluorescence or Förster resonance energy transfer (FRET) [15], it is possible to generate time distributions for reactions, such as the unbinding-time distributions of ligands unbinding from a single receptor (Fig. 5.1). Such distributions are particularly useful when the pathway includes multiple steps, as they can be quite complex and information rich. Faced with systems exhibiting several characteristic times, least-squares (LS) fitting is often brought to bear on the problem. Though popular and often useful, there are situations in which standard LS approaches fail, and unfortunately often do so in quite non-obvious ways. To help the reader understand and avoid such pitfalls, we here explore some of these situations through the lens of ML estimation, an alternative approach that has become very popular in the physical sciences [1–3, 7, 8, 13, 16, 17, 24–26, 28].

As it is straight forward, adaptable, and well suited to SM experiments, we here provide a self-contained introduction to ML estimation. We heuristically show that ML estimation should generally outperform LS fitting and explicitly show this to be the case in relevant SM FRET examples. We close with a discussion of how to use bootstrapping to estimate the standard deviation of fit parameters. The presentation

B. Eslami-Mosallam · I. Katechis · M. Depken (✉)

Department of BioNanoScience, Kavli Institute of NanoScience, Delft University of Technology, 2629 HZ Delft, The Netherlands

e-mail: s.m.depken@tudelft.nl

© Springer Science+Business Media, LLC, part of Springer Nature 2019

C. Joo and D. Rueda (eds.), *Biophysics of RNA-Protein Interactions*,

Biological and Medical Physics, Biomedical Engineering,

https://doi.org/10.1007/978-1-4939-9726-8_5

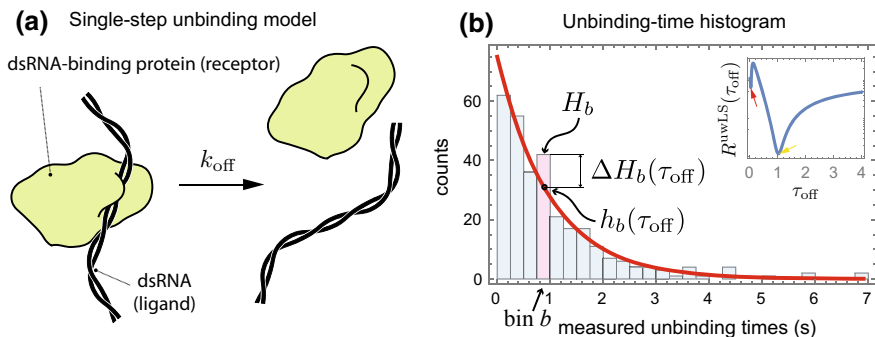


Fig. 5.1 **a** A single-step ligand-receptor unbinding model. A dsRNA-binding protein releases dsRNA at a characteristic rate k_{off} . For this model, we expect an exponential distribution of unbinding time, with the average unbinding time $\hat{\tau}_{\text{off}} = 1/k_{\text{off}}$. **b** A histogram (bars) formed from 300 unbinding times picked from an exponential distribution with the true average unbinding time $\hat{\tau}_{\text{off}} = 1$ s. The predicted bin counts for a model with average unbinding time $\tau_{\text{off}} = 1$ s are shown as a red curve, and the notation used in Eq. (5.1) is indicated for bin b (pink bar). In the inset, we show the unweighted LS residue $R^{\text{uWLS}}(\tau_{\text{off}})$ (in log-scale) as a function of the model parameter τ_{off} . The function displays a global minimum close to the true average unbinding time (yellow arrow), as well as a local minimum for short times (red arrow). Beware that local minima can sometimes trap numerical minimization algorithms, leading them to erroneously report the local minimum as the sought after global minimum

is intended for SM experimenters who find fitting data indispensable to their work, but might find the advantages/limitations/rationale of various approaches hard to ascertain.

5.2 Prerequisites

In an effort to be self-contained, we start by discussing LS fitting, as well as error estimation and some crucial concepts in probability theory. These sections can be skipped by the initiated reader.

5.2.1 LS Fitting and the Distance Between Model and Data

LS fitting comes in several flavors, depending on how statistical fluctuations in bin counts are accounted for. The fitting is generally performed by collecting the available data into bins $b = 1, 2, \dots, B$, and finding the model parameter values that minimize the total square deviation between actual bin counts (H_b) and model predictions for bin counts (h_b) (Fig. 5.1b), normalized with the true standard deviation of the bin count (σ_b). We will refer to this approach as true LS (tLS) fitting. For unbinding

times in the simple RNA-protein example of Fig. 5.1a, tLS fitting consists of finding the model parameter τ_{off} (the average unbinding time of the model) that minimizes the total residue

$$R^{\text{tLS}}(\tau_{\text{off}}) = \sum_{b=1}^B \left(\frac{\Delta H_b(\tau_{\text{off}})}{\sigma_b} \right)^2, \quad \Delta H_b(\tau_{\text{off}}) = H_b - h_b(\tau_{\text{off}}). \quad (5.1)$$

Minimizing the total residue R^{tLS} makes intuitive sense, as it penalizes parameter values that give large deviations between predictions and measurements, in a manner scaled by the size of statistical fluctuations in each bin. A perfect estimate in a bin ($H_b = h_b(\tau_{\text{off}})$) results in zero residue, while any positive (weighted) residue gives a measure of the “statistical distance” between model and data in that bin. By summing the residues in Eq. (5.1), we get a measure of the total distance between model and data; tLS fitting aims to minimize this distance.

Unfortunately, we do not often have access to the true standard deviation of counts in each bin, and various approximations to Eq. (5.1) must be deployed. For ease of presentation, we will here focus on two cases: In the first case, we assume that count fluctuations are almost constant over all bins, and we use unweighted LS (uwLS) residues by taking σ_b to be constant¹ (e.g., see inset in Fig. 5.1b); in the second case, we assume a fixed total number (N) of independent measurements, such that the count fluctuations in each bin are binomially distributed, with $\sigma_b = \sqrt{\langle H_b \rangle (1 - \langle H_b \rangle / N)} \approx \sqrt{\langle H_b \rangle}$. Here, the angle brackets represent the statistical average over a large number of experiments, and we have in the last step assumed bins to be small enough that no bin on average contains a large fraction of the total number of observations (i.e., $\langle H_b \rangle \ll N$ for all bins). With no better estimate at hand, the statistical average of bin counts is often approximated with the observed bin count by setting $\sigma_b \approx \sqrt{H_b}$ in Eq. (5.1). We will refer to this approach as weighted LS (wLS).

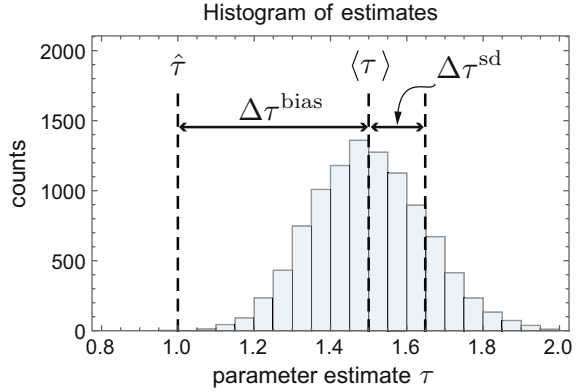
Both wLS and uwLS fitting can be problematic. Using uwLS, we assume fluctuations in bin counts to be uniform over bins. As we shall see, this is often a poor approximation for systems with multiple characteristic timescales. Using wLS, we instead use individual bin counts to estimate the standard deviation of counts in that bin. As individual bin counts can be small, relative fluctuations can be large, resulting in large approximation errors when using $\sigma_b \approx \sqrt{H_b}$ in Eq. (5.1).

5.2.2 Error Estimation, Variation, and Systematic Bias

For any estimation method applied to an experiment with a finite set of measurements, the estimated parameter value (τ) will deviate from the true value ($\hat{\tau}$). To compare

¹Note that we do not need to know the actual constant value of σ_b , as it will not affect the position of the minimum of R^{uwLS} .

Fig. 5.2 A histogram of estimates for a hypothetical process with the true parameter value $\hat{\tau} = 1$. The systematic bias $\Delta\tau^{\text{bias}}$ and the typical size of fluctuations $\Delta\tau^{\text{sd}}$ around the average estimate $\langle\tau\rangle$ are indicated



two methods, we need to understand the distribution of parameter estimates that each approach would yield were it to be repeated many times. Over a large number of experiments, the typical error can be measured by the mean square error, $\text{MSE} = \langle(\tau - \hat{\tau})^2\rangle$. To understand the nature of estimation errors, consider the bias $\Delta\tau^{\text{bias}} = \langle\tau\rangle - \hat{\tau}$, capturing how the average estimate deviates from the true parameter value, as well as the standard deviation $\Delta\tau^{\text{sd}} = \sqrt{\langle(\tau - \langle\tau\rangle)^2\rangle}$, capturing the typical spread of estimates around their average (Fig. 5.2). Conveniently, the bias and standard deviation add in quadrature to form the MSE [9]

$$\text{MSE} = (\Delta\tau^{\text{bias}})^2 + (\Delta\tau^{\text{sd}})^2.$$

The smaller the MSE the better, and we should seek to minimize both the bias and standard deviation as far as possible. A large bias can be introduced by the estimation method itself, while a large standard deviation typically results from a lack of data and/or accuracy of the measurements.

5.2.3 Bayes' Equation and Observation Frequencies

To explain the rationale behind ML estimation [9], we first introduce Bayes' equation by way of Venn diagrams and the frequentist interpretation of probability. According to this interpretation, probabilities can be seen as the asymptotic frequency of outcomes, recorded over a large number of repetitions [12]. For concreteness, imagine a steady rainfall with water drops hitting the yellow (event A)- and blue (event B)-striped shapes shown in Fig. 5.3. Further imagine keeping track of the number of raindrops that falls on the section with just yellow stripes (N_A), just blue stripes (N_B), both yellow and blue stripes ($N_{A\&B}$), or anywhere ($N_{\text{tot}} = N_A + N_B$). Among these various counts, the relationship

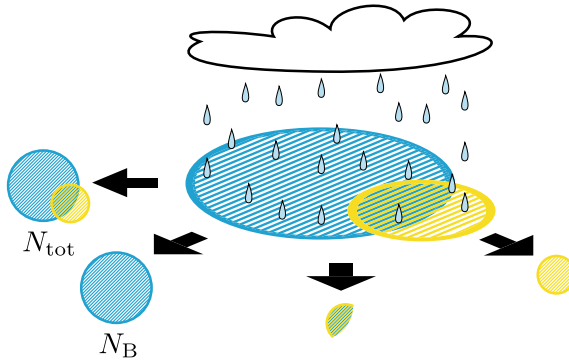


Fig. 5.3 Imagine exposing the blue and yellow shapes to rain, while keeping track of the number of raindrops that hit each differently striped area. If the rainfall is steady, we can use the frequentist interpretation of probability to relate the different fractions of raindrops landing on the various areas to probabilities. The trivial Eq. (5.2) then becomes Bayes' equation as expressed in Eq. (5.3)

$$\frac{N_{A,B}}{N_{\text{tot}}} = \frac{N_{A\&B}}{N_B} \frac{N_B}{N_{\text{tot}}} = \frac{N_{A\&B}}{N_A} \frac{N_A}{N_{\text{tot}}} \quad (5.2)$$

holds trivially true, as can be seen by canceling the first denominator with the second numerator after each equal sign. If we collect enough raindrops, the fraction of raindrops that has so far landed on a particular section will approach the probability that also the next raindrop will land in that same section. Taking the frequentist approach, we can translate Eq. (5.2) into Bayes' equation for probabilities

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A). \quad (5.3)$$

In the above, $P(A, B) = N_{A\&B}/N_{\text{tot}}$ is the joint probability that both A and B occur, $P(A) = N_A/N_{\text{tot}}$ is the probability that A occurs irrespective of whether B occurs or not, $P(A|B) = N_{A\&B}/N_B$ gives the conditional probability that A occurs, given that B occurs, and so on swapping A and B .

5.2.4 Continuous Outcomes and Probability Densities

We are ultimately interested in measurements that produce real numbers (such as unbinding times), while Bayes' equation (Eq. 5.3) is valid for probabilities of discrete events. For outcomes that can take a continuous value, the relevant concept is that of the probability density function (PDF). For two concurring continuous outcomes, denote recording the respective values in interval I_a and I_b , centered around a and b , as event A and B . For very short interval lengths ΔI_a and ΔI_b , the probability to end up in the interval (denoted by upper case P) is simply the relevant PDF (denoted by a lower case p) multiplied with the relevant interval length(s)

$$P(A, B) = p(a, b)\Delta I_a\Delta I_b,$$

$$P(A) = p(a)\Delta I_a, P(A|B) = p(a|b)\Delta I_a, P(B) = p(b)\Delta I_b, P(B|A) = p(b|a)\Delta I_b.$$

The above relations can be plugged into Bayes' equation for probabilities of discrete events (Eq. 5.3), giving the sought after Bayes' equation for PDFs of continuous outcomes

$$p(a, b) = p(a|b)p(b) = p(b|a)p(a). \quad (5.4)$$

With the prerequisites covered, we are now ready to address the rationale behind ML estimation and assess how it compares and relates to LS fitting.

5.3 Maximum Likelihood

To keep the discussion in general, consider an experiment where we collect N -independent measurements $\{t\}_N = \{t_1, t_2, \dots, t_N\}$ and modeled it as process with M parameters $\{\tau\}_M = \{\tau_1, \tau_2, \dots, \tau_M\}$. Based on our one experiment, we would like to determine the parameter values that gave rise to the data. As the stochasticity of the data makes it impossible to precisely determine the parameters exactly, our best bet would by definition be to find the most probable set of parameter values, *given* the data. In the language of conditional PDFs, this corresponds to finding the model parameters which maximize the PDF $p(\{\tau\}_M|\{t\}_N)$ of having a model with parameters $\{\tau\}_M$ given the measured data $\{t\}_N$ (for a lighthearted and instructive discussion of the meaning of the probability of a model, see [18]). Unfortunately, we do not have direct access to this conditional PDF. Still, we can make considerable progress by using Bayes' equation and introducing a few additional assumptions.

5.3.1 The Most Likely Model

Through Bayes' equation for PDFs (Eq. 5.4), we can relate the unknown PDF of interest to PDFs about which we do have some knowledge, or regarding which we can at least make some reasonable assumptions. Letting $a = \{t\}_N$ and $b = \{\tau\}_M$ in Eq. (5.4), we have²

$$p(\{\tau\}_M|\{t\}_N) = \frac{p(\{\tau\}_M)}{p(\{t\}_N)} p(\{t\}_N|\{\tau\}_M).$$

²A more intuitive way of writing this might be in the form $p(\text{model}|\text{data}) = \frac{p(\text{model})}{p(\text{data})} p(\text{data}|\text{model})$.

With the aim to maximize the left-hand side of the above expression with respect to the model parameters, we note that the denominator on the right-hand side does not depend on the model parameters and therefore will not influence which parameter value maximizes the left-hand side; we promptly ignore the denominator. The numerator can be interpreted as encoding what we knew of the correct parameter values before our experiments. If we assume little or no prior knowledge, it makes sense to also assume this prior PDF to be roughly uniform and thus largely independent of the model parameters³; we promptly ignore also the numerator. The last term on the right-hand side of the equation describes the PDF of a particular set of measurements, given the model parameters. This conditional PDF *can* be calculated if we have a model of the system!

Through the above argument, we conclude that by maximizing the *likelihood function* $p(\{t\}_N|\{\tau\}_M)$, we can find an estimate for the model parameter values that best describe the data. Equivalently, we could choose to minimize the *log-likelihood function*⁴ $L^{\text{ML}}(\{\tau\}_M) = -\ln p(\{t\}_N|\{\tau\}_M)$, which has a global minimum for the same parameter values as the likelihood function has a global maximum. As we assume *independent* measurements, the PDF of the whole experimental outcome $\{t\}_N$ can simply be written as the product of the PDFs for each measurement. The log-likelihood function then has the convenient property that it turns into a sum over measurements,

$$L^{\text{ML}}(\{\tau\}_M) = -\ln\left(\prod_{n=1}^N p(t_n|\{\tau\}_M)\right) = -\sum_{n=1}^N \ln p(t_n|\{\tau\}_M). \quad (5.5)$$

Finding the parameter values that globally minimize Eq. (5.5) constitutes ML parameter estimation, and we now apply it to a few simple but illustrative examples to familiarize the reader with the approach.

5.3.2 ML Estimation for an Exponential Process

To demonstrate ML estimation in practice, we return to ligand–receptor unbinding. For simple unbinding kinetics, the unbinding times are exponentially distributed with the PDF $p(t|\tau_{\text{off}}) = e^{-t/\tau_{\text{off}}}/\tau_{\text{off}}$. Inserting this PDF into Eq. (5.5), we see that the log-likelihood function is given by

³There are subtleties here relating to variable changes [18], but these lie outside our present scope.

⁴It should be noted that as the logarithm takes a unit-less argument, while the PDF has units (inverse time in case of the unbinding experiments). Strictly, we therefore need to multiply the PDF with some constant that renders the argument of the logarithm unit less in the definition of $L^{\text{ML}}(\{\tau\}_M)$. As the value of this constant does not affect the position of the minimum, we drop it for notational convenience.

$$L^{\text{ML}}(\tau_{\text{off}}) = N \left(\ln \tau_{\text{off}} + \frac{\bar{t}}{\tau_{\text{off}}} \right), \quad \bar{t} = \frac{1}{N} \sum_{n=1}^N t_n.$$

The ML estimate ($\tau_{\text{off}}^{\text{ML}}$) is now arrived at by minimizing $L^{\text{ML}}(\tau_{\text{off}})$ with respect to τ_{off} . In this simple example, we can find the ML estimate analytically by using the zero-derivative test for finding an optimum,

$$0 = \frac{\partial L^{\text{ML}}}{\partial \tau_{\text{off}}}(\tau_{\text{off}}^{\text{ML}}) = N \left(\frac{1}{\tau_{\text{off}}^{\text{ML}}} - \frac{\bar{t}}{(\tau_{\text{off}}^{\text{ML}})^2} \right) \Rightarrow \tau_{\text{off}}^{\text{ML}} = \bar{t}. \quad (5.6)$$

Consequently, ML estimation confirms the well-known result that the characteristic time of an exponential process can be estimated by the average event time observed in the data; or simply, the off-rate estimate is $k_{\text{off}}^{\text{ML}} = 1/\bar{t}$. Note that we did not need to perform any binning to extract this estimate, which constitutes a clear advantage over standard LS fitting methods.

5.3.3 ML Estimation for an Exponential Process with a Time Cutoff

The simplest additional characteristic time to consider is possibly that introduced by photobleaching in FRET experiments. With photobleaching, the experimental signal in our unbinding example can, in addition to unbinding, also be lost due to the stochastic degradation of fluorophores over time. We can account for photobleaching by interpreting the estimated characteristic rate ($1/\tau_{\text{off}}^{\text{ML}}$) of the PDF (which is still exponential), not purely as the unbinding rate, but as the sum of the unbinding and bleaching rate. As the bleaching rate can usually be independently measured, we can often readily estimate the unbinding rate by subtracting the bleaching rate from the estimated total rate.

Next, consider having a hard cutoff time T_{cut} limiting the duration of each measurement. Slightly more complex than photobleaching, this scenario will serve to demonstrate that the ML approach often allows us to utilize extra information in a rational manner. Though we cannot know the precise duration for any binding event lasting longer than T_{cut} , there is information in the number of unbinding events that exceeded it. We start by noting that the simple ML recipe used in Eq. (5.6) does not work, as losing long unbinding times will clearly lead us to underestimate the characteristic unbinding time. Instead, we would like to keep the information regarding the number of measurements that exceeded the finite measurement time window. Combining the probability densities of the measured unbinding times ($\{t\}_{N_{\text{rec}}}$) with the probabilities of the missed times ($\{t'\}_{N_{\text{cut}}}$), the relevant likelihood function is

$$\underbrace{\prod_{n=1}^{N_{\text{rec}}} p(t_n | \tau_{\text{off}})}_{\text{PDF of the } N_{\text{rec}} \text{ recorded events}} \underbrace{\prod_{n'=1}^{N_{\text{cut}}} P(t_{n'} > T_{\text{cut}} | \tau_{\text{off}})}_{\text{probability of the } N_{\text{cut}} \text{ missed events}} = \prod_{n=1}^{N_{\text{rec}}} p(t_n | \tau_{\text{off}}) \left(\int_{T_{\text{cut}}}^{\infty} dt' p(t' | \tau_{\text{off}}) \right)^{N_{\text{cut}}}.$$

The corresponding log-likelihood function will now become a sum over both probability densities (for the N_{rec} recorded times) and probabilities (for the N_{cut} missed times)

$$\begin{aligned} L^{\text{ML}}(\tau_{\text{off}}) &= - \sum_{n=1}^{N_{\text{rec}}} \ln p(t_n | \tau_{\text{off}}) - N_{\text{cut}} \ln \int_{T_{\text{cut}}}^{\infty} dt p(t | \tau_{\text{off}}) \\ &= N_{\text{rec}} \left(\ln \tau_{\text{off}} + \frac{\bar{t}}{\tau_{\text{off}}} \right) + \frac{N_{\text{cut}} T_{\text{cut}}}{\tau_{\text{off}}}. \end{aligned} \quad (5.7)$$

The ML estimate can once again be found analytically through the zero-derivative condition, yielding the simple formula

$$0 = \frac{\partial L^{\text{ML}}}{\partial \tau_{\text{off}}}(\tau_{\text{off}}^{\text{ML}}) \Rightarrow \tau_{\text{off}}^{\text{ML}} = \bar{t} \left(1 + \frac{T_{\text{cut}} N_{\text{cut}}}{\bar{t} N_{\text{rec}}} \right) \quad (5.8)$$

to correct for the cutoff-induced bias. Note that the correction only becomes significant when the lower bound of the total duration of cut events ($T_{\text{cut}} N_{\text{cut}}$) becomes comparable to the total time of recorded events ($\bar{t} N_{\text{rec}}$).

5.3.4 ML Estimation for a Double-Exponential Process

The unbinding process itself might have several characteristic times. We next consider the case where the model yields a double-exponential PDF of unbinding times and where the maximal measurement duration is large enough to be ignored. For the unbinding problem discussed above, such PDFs could originate in two interconvertible binding modes: a loose binding mode where the ligand first binds, and eventually unbinds from, and a tight binding mode from which the ligand cannot unbind directly (see Fig. 5.4a). Alternatively, it could result from two protein populations with different unbinding rates. The PDF for either system can be written as (Fig. 5.4b)

$$p(t | \tau_1, \tau_2, P_1) = \frac{P_1}{\tau_1} e^{-t/\tau_1} + \frac{1 - P_1}{\tau_2} e^{-t/\tau_2} \quad (5.9)$$

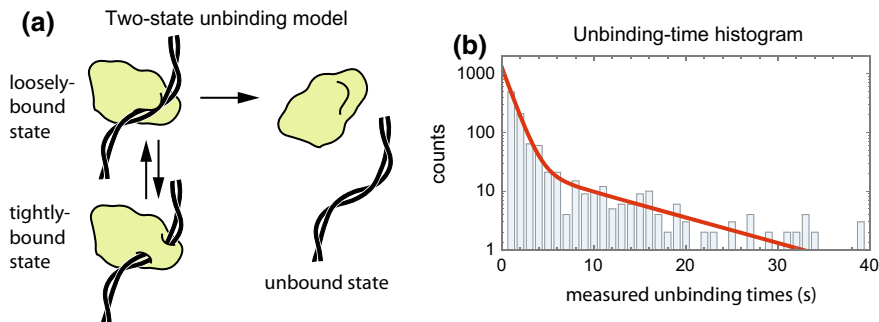


Fig. 5.4 **a** A dsRNA-binding protein exhibiting two bound states, resulting in a double-exponential PDF for the unbinding time. **b** Histogram (bars, in log-scale) formed by picking 1000 unbinding times from a double-exponential distribution with a PDF characterized by $\hat{\tau}_1 = 1$ s, $\hat{\tau}_2 = 5$ s, and $\hat{P}_1 = \hat{P}_2 = 0.5$. The predicted bin counts for a model with $\tau_1 = 1$ s, $\tau_2 = 5$ s, and $P_1 = P_2 = 0.5$ are shown as a red curve

where the characteristic times τ_1 and τ_2 , as well as the population fraction P_1 associated with τ_1 , can be directly related to the microscopic rates of the relevant system. Attempting to use the PDF of Eq. (5.9) to calculate the log-likelihood function according to Eq. (5.5), it quickly becomes clear that we can no longer find a simple analytic solution to the minimization problem. This is quite generally the case, and one has to perform the minimization numerically, as we will do when comparing LS and ML approaches on simulated data below.

5.3.5 Coarse-Grained Likelihood

Though ML estimation has the clear advantage of requiring no binning of the data, for large data sets, it often becomes computationally demanding to numerically minimize a log-likelihood function with as many terms as there are measurements (see sum in Eq. 5.5). The computational efficiency can be drastically increased by considering the likelihood over bins, which should be a reasonable approximation as long as we choose the bin size small enough for there to be little change in the PDF over each bin. The probability P_b of a particular measurement ending up in bin b can then be related to the model PDF and used to calculate the predicted bin count h_b as

$$h_b(\{\tau\}_M) = N P_b(\{\tau\}_M), \quad P_b(\{\tau\}_M) = \int_{t_b - \Delta t_b/2}^{t_b + \Delta t_b/2} dt p(t|\{\tau\}_M) \approx \Delta t_b p(t_b|\{\tau\}_M), \quad (5.10)$$

where the integral runs over the whole width Δt_b of bin b centered around t_b .

Splitting the sum over measurements in the definition of the log-likelihood function (Eq. 5.5) into a sum over bins and a sum over measurements in each bin, it can

be approximated by the *coarse-grained* (cg) log-likelihood function

$$L^{\text{ML}}(\{\tau\}_M) = - \sum_{b=1}^B \sum_{\substack{t_n \text{ in} \\ \text{bin } b}} \ln p(t_n | \{\tau\}_M) \approx - \sum_{b=1}^B H_b \ln h_b(\{\tau\}_M) = L^{\text{cgML}}(\{\tau\}_M). \quad (5.11)$$

Here, the last equality is a definition, and we have dropped constant terms and factors not affecting the minimizing parameter values. Note that the results of using cgML estimation can always be made arbitrarily close to the original ML estimate by choosing the bin widths small enough.

5.3.6 The Connection Between LS and ML

We will now show that ML estimation can be seen as another approximation of tLS and, importantly, one that is generally expected to do better than both uwLS and wLS. The connection between LS and ML estimation has been studied for the case of independent and Gaussian-distributed data with equal variance [10], but in an effort to understand the differences in estimates more generally, we here employ a heuristic approach with wide applicability.

For any data set $\{t\}_N$ and model with parameter set $\{\tau\}_M$, we seek to compare tLS fitting to ML estimation. As the tLS scheme is based on binned data sets, we opt to compare it to equally binned cgML. The zero-derivative condition for finding the tLS parameter estimates $\{\tau^{\text{tLS}}\}_M$ from Eq. (5.1) is

$$0 = \frac{\partial R^{\text{tLS}}}{\partial \tau_m}(\{\tau^{\text{tLS}}\}_M) \approx \sum_b \frac{\Delta H_b(\{\tau^{\text{tLS}}\}_M)}{\langle H_b \rangle} \frac{\partial \Delta H_b(\{\tau^{\text{tLS}}\}_M)}{\partial \tau_m}, \quad m = 1, \dots, M. \quad (5.12)$$

Similarly, differentiating Eq. (5.11), and using the normalization of probabilities $\left(\sum_b h_b = N\right)$, the condition for finding the cgML estimate $\{\tau^{\text{cgML}}\}_M$ can be written as

$$\begin{aligned} 0 &= - \frac{\partial L^{\text{cgML}}}{\partial \tau_m}(\{\tau^{\text{cgML}}\}_M) \\ &= \sum_b \frac{\Delta H_b(\{\tau^{\text{cgML}}\}_M)}{h_b(\{\tau^{\text{cgML}}\}_M)} \frac{\partial \Delta H_b(\{\tau^{\text{cgML}}\}_M)}{\partial \tau_m}, \quad m = 1, \dots, M. \end{aligned} \quad (5.13)$$

Interestingly, though the functions that are minimized during tLS (Eq. 5.1) and cgML (Eq. 5.11) estimation are quite different, their minima are located in close proximity. From above, it is clear that the cgML minimization condition (Eq. 5.13) can be seen as an approximation to the tLS minimization condition (Eq. 5.12) with $\langle H_b \rangle \approx h_b(\{\tau^{\text{cgML}}\}_M)$.

The cgML approximation ($\langle H_b \rangle \approx h_b(\{\tau^{\text{cgML}}\}_M)$) should be compared to the wLS approximation ($\langle H_b \rangle \approx H_b$). The wLS approximation includes only the data of each bin when estimating the variance in each bin. The cgML approximation takes into account the data in all bins, since $\{\tau^{\text{cgML}}\}_M$ is estimated from the whole data set by definition. As increasing the number of measurements generally reduces both the variance and systematic bias of estimates, we typically expect the cgML approach to outperform the wLS approach. It should be noted that the ML approach is *not* equivalent to setting $\sigma_b \approx \sqrt{h_b(\{\tau^{\text{cgML}}\}_M)}$ already in Eq. (5.1), as we would then need to know the optimal parameters before we have minimized the residue to find them. ML estimation elegantly bypasses this problem by enforcing the same approximation, not on the function to be minimized but directly on the condition defining the minimum (Eq. 5.13).

Having argued that we should generally expect (cg)ML to outperform wLS, we explicitly compare their performance, together with that of uwLS, on the examples used above.

5.4 Comparing LS and ML Through Simulations

Having established that uwLS, wLS, and cgML can all be seen as tLS approximations of various severity, we here numerically explore the consequences of these approximations. By generating data with a known distribution, we can quantify the success of the different approaches at estimating known parameter values. We do not discuss the numerical minimization schemes we use when analytics fail, further than stating that it is implemented in Mathematica™, using a simulated-annealing algorithm [22] to minimize the risk of finding a local rather than global minimum (see inset in Fig. 5.1b, e.g., of a local (red arrow) and global (yellow arrow) minimum). There are many powerful software packages available with the required numerical optimization capabilities.

Without a sharp cutoff time for the measurements, we always expect many long-time bins to be empty in the tail end of the PDF. A zero count in any bin is catastrophic for wLS, as it gives a zero estimate for the standard deviation and so introduces infinite terms in Eq. (5.1). In an attempt to circumvent such issues, various re-binning procedures or reassignments of weights can be performed. Though such approaches avoid infinite terms in Eq. (5.1), they do change the details of the estimation method depending on the observed data, and so risk introducing a strong bias. For simplicity, we will here only consider the interval between the highest and lowest measured data

points generated, and for wLS we choose the minimum constant bin size that leaves *no* empty bins in the intervening interval.

5.4.1 Method Comparison for an Exponential Process

Though trivial, we start with the simple exponential process. Using Eq. (5.10), we can calculate the predicted bin counts $h_b(\tau_{\text{off}})$ from the PDF. It should be noted that we could in principle estimate both N and τ_{off} by optimizing with respect to both in any LS or ML approach. Though this is often done, it is not advisable as it will increase the MSE compared to if we heed the fact that N is *known* and precisely dictates the translation from probability to histogram counts in Eq. (5.10).

In Fig. 5.5, we show the results of using uwLS, wLS, and ML estimation on 10,000 exponentially distributed data sets of 100 measurements each ($\hat{\tau}_{\text{off}} = 1$ s). Even after eliminating the zero bins for wLS (see above), the wLS estimate remains biased due to the unavoidable presence of the low-count bins [11, 19–21, 23, 27]. This bias has been shown to be inversely proportional to the average occupancy of the bins [11]. The fact that uwLS estimation introduces a much smaller—if not vanishing—bias compared to wLS estimation might seem strange, given that the latter estimates the standard deviations in bins based on the data, while the former ignores the data and assumes them all equal. The explanation can likely be found in that though the weighted approach clearly employs better approximation for bins

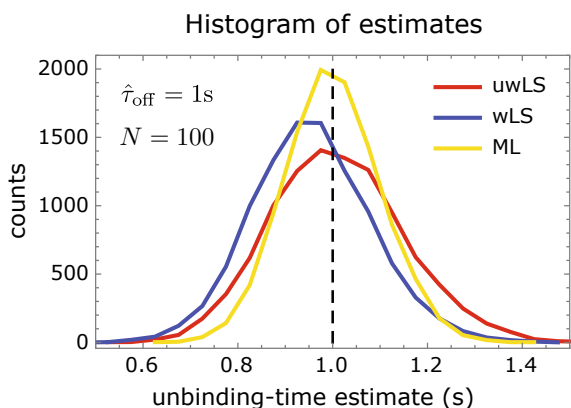


Fig. 5.5 Distribution of unbinding-time estimates from 10,000 exponentially distributed data sets containing 100 samples each. There is a clear bias for wLS estimation (-0.04 s), while little bias is apparent for uwLS (0.009 s) and ML (0.0002 s) estimation. The standard deviation of ML estimation (0.10 s) is less than for wLS estimation (0.13 s), which in turn is less than for uwLS estimation (0.14 s). Notwithstanding the larger absolute bias, the $\sqrt{\text{MSE}}$ for wLS estimation (0.13 s) outperforms that for uwLS estimation (0.14 s), while ML estimation outperforms both other methods (0.10 s)

with many counts, the relative errors in low-count bins can be very large, outstripping the error made when assuming the variance of counts to be equal in all bins. Among the three approaches, ML is clearly preferable as both bias and standard deviation are the smallest.

5.4.2 *Method Comparison for an Exponential Process with a Cutoff*

Next, we consider a measurement that is limited by a maximum measurement time T_{cut} . If this cutoff time is largely compared to the average unbinding time, we effectively have no cutoff, which we covered in the previous section. If we instead have a cutoff time that is comparable to the average unbinding time, there is information in the number of unbinding events that exceeded the maximal duration of the measurements. With a measurement cutoff time, the unbinding times are still exponentially distributed, but the number of experimental observations $N = N_{\text{rec}} + N_{\text{cut}}$ has to be split into the N_{rec} events where the time was recorded, and the N_{cut} events for which we know only that they lasted longer than T_{cut} . For both wLS and uwLS, we explicitly fit only the N_{rec} measurements falling within the observation window, while for ML estimation, we include also the information regarding the cut events, according to Eq. (5.8).

Though we lose data, introducing a short-time cutoff has the benefit of removing bins that are likely to have zero counts, and thus, we decrease the need to re-bin data for wLS estimation. For small data sets (Fig. 5.6a, b), the counts in each bin will still have large (relative) fluctuations, and it is not surprising that we see a substantial error in wLS estimation. This error decreases as the cutoff is lowered and progressively fewer low-count bins are included (c.f. Fig. 5.6a with b), even though a higher fraction of measurements falls outside the observation window. For the cutoff time close to the characteristic unbinding time, uwLS and ML estimation are comparable, as the variance in bin counts is roughly constant among bins below the cutoff time. This shows a scenario where uwLS outperforms wLS, though ML estimation consistently remains the better alternative.

As we increase the size of the data sets by a factor 100 (Fig. 5.6c, d), we expect the relative fluctuations around the predicted bin counts to decrease, bringing wLS estimation closer to ML estimation. This effect can be seen clearly seen in Fig. 5.6c, d. It is interesting to note that for these large data sets, the extra information regarding the cut measurements included in the ML estimation had little effect on the fit, as all fits roughly coincide in Fig. 5.6c, d.

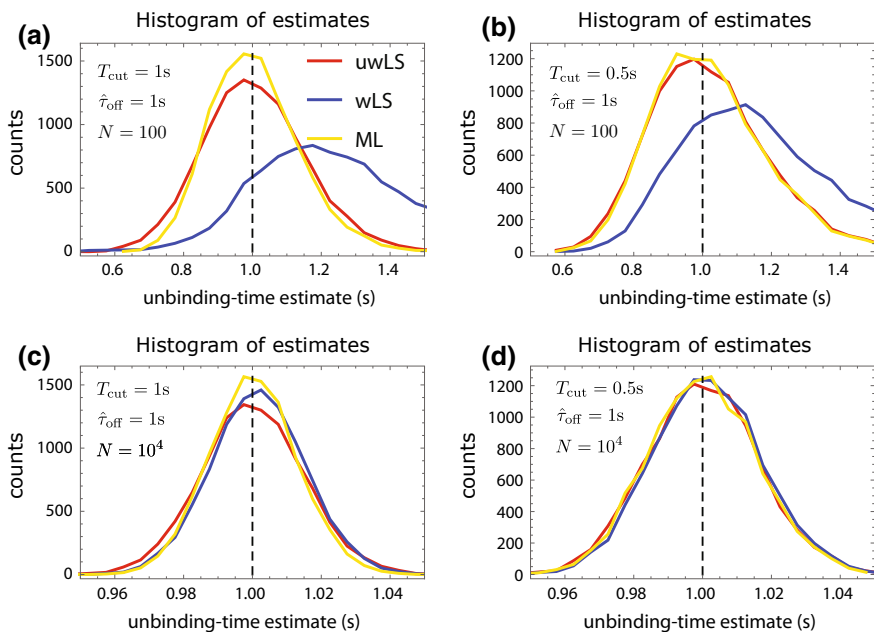


Fig. 5.6 Parameter estimation for 10,000 exponentially distributed data sets with a cutoff. **a** For sets with 100 measurements, and a $T_{\text{cut}} = 1$ s, we see a clear bias in wLS estimation, while uwLS estimation has a somewhat larger standard deviation than ML estimation. **b** For a lower $T_{\text{cut}} = 0.5$ s, the bias for wLS estimation decreases slightly, while uwLS approaches ML estimation. **c** Increasing the size of the data sets to 10,000 measurements and considering a moderate cutoff time, the difference between wLS estimation and ML estimation diminishes and both methods marginally outperform uwLS estimation. **d** For large data sets and a low cutoff time, all methods converge

5.4.3 Method Comparison for a Double-Exponential Process

For data distributed according to the double-exponential PDF of Eq. (5.9), we need to fit out two characteristic times ($\hat{\tau}_1$ and $\hat{\tau}_2$), together with the fraction of events belonging to each ($\hat{P}_1, \hat{P}_2 = 1 - \hat{P}_1$). In Fig. 5.7, we show the results of 10,000 fits to data sets of size 10,000, for a process with moderately separated characteristic times ($\hat{\tau}_1 = 1$ s, $\hat{\tau}_2 = 3$ s) and for three different population fractions ($\hat{P}_1 = 0.1$ Fig. 5.7a–c, $\hat{P}_1 = 0.5$ Fig. 5.7d–e, $\hat{P}_1 = 0.9$ Fig. 5.7g–h). In each case, we report the $\sqrt{\text{MSE}}/s$ within parenthesis in the legend.

The error in the short-timescale estimate (τ_1) is dominated by the variance around the average for all methods, and all methods perform better the larger the fraction of events corresponding to the shorter timescale are (Fig. 5.7a, d, and g). The error in the long-timescale estimates (τ_2) is also dominated by the variances, which is particularly large in uwLS estimation (Fig. 5.7b, e, and h). This can likely be traced back to the fact that the constant variance assumption of uwLS suppresses the relative influence of long timescales, introducing a relatively low penalty for variation here.

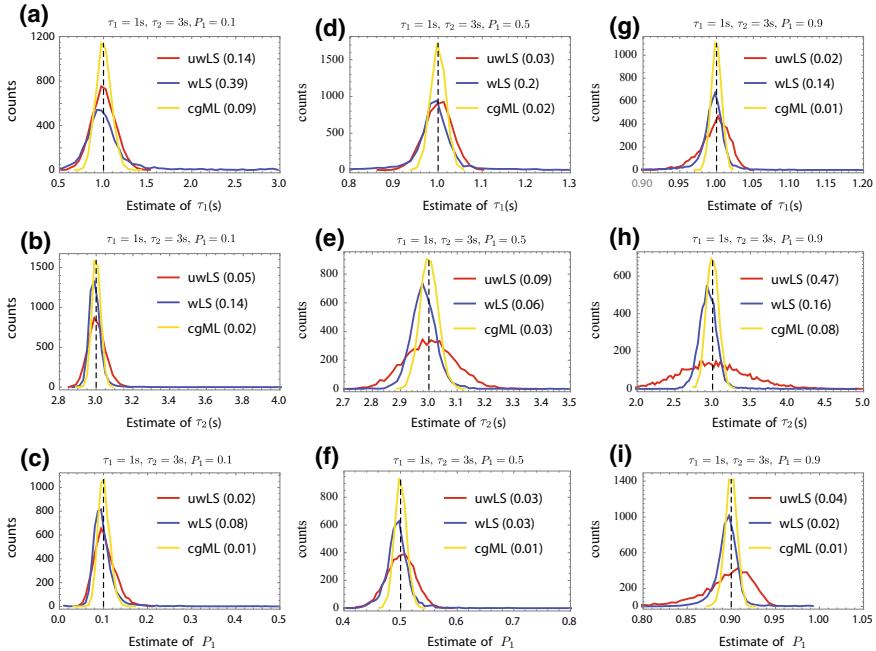


Fig. 5.7 Parameter estimation over 10,000 double-exponentially distributed data sets of size 10,000. Each column corresponds to parameter estimate distributions for a particular value of P_1 , and each row corresponds to a particular model parameter. **a–c** $P_1 = 0.1$. **d, e** $P_1 = 0.5$. **f, h** $P_1 = 0.9$. In each case, we report the $\sqrt{\text{MSE}}$ s within parenthesis in the legend. In all considered situations, ML estimation is clearly the preferable choice as it has the lowest $\sqrt{\text{MSE}}$

The error in the estimation of the fraction of measurements belonging to the short timescale (P_1) is also dominated by the variance, and uwLS is particularly effected due to the poor accounting for the change in variance going from short to long timescales (Fig. 5.7c, f, and i). For all parameter values considered, cgML estimation again clearly outperforms the other methods as was expected from our theoretical developments.

5.5 Fitting Experimental Data

In the previous section, we have examined the performance of LS and ML estimation on well-specified data sets without experimental noise. Though a proper treatment of experimental noise is outside our present scope, it is still interesting to apply the three fitting methods on experimental data to see to what extent they agree. Considering experimental data will also give us the opportunity to comment on how to estimate the variance of parameter estimates through bootstrapping.

5.5.1 All Fits Different, but All Naively Plausible

Continuing with our RNA–protein unbinding example, we now analyze SM total internal reflection microscopy (TIRFM) data. The experiments measure the unbinding time of double-stranded (ds) RNA from viral RNA-binding proteins involved in protecting the viral genome from the hosts’ RNA interference-based defenses [6]. The viral suppressors of RNA interference (VSR) proteins are immobilized on a glass surface, and the binding/unbinding of fluorescently tagged dsRNAs to the immobilized VSRS is followed (for more information on the biological aspects and the interpretation of the data, see [6]).

The unbinding-time data of 50 nucleotide dsRNA-binding VSR is fitted with uwLS, wLS, and cgML methods in Fig. 5.8a–c. In this particular system, and presumably due to the existence of weak and very strong binding modes, it is common to have a population of VSRS that unbind quickly, as well as a population that remain bound for the duration of the measurement. In the latter case, the apparent unbinding time will report on the photobleaching time of the fluorophores, as discussed previously. In such situations, the appropriate PDF is double exponential (Eq. 5.9), and the information regarding the number of molecules still bound and fluorescing at the end of the experiment (N_{cut}) can be incorporated in the ML estimation along

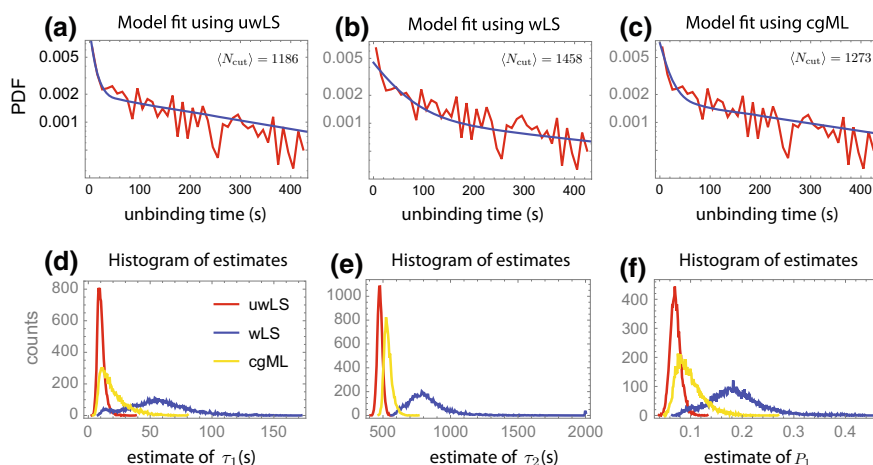


Fig. 5.8 **a** The measured distribution of unbinding times (red) together with the uwLS fit (blue). **b** The measured distribution of unbinding times (red) together with the wLS fit (blue). **c** The measured distribution of unbinding times (red) together with the cgML fit (blue). In **a–c**, the average number of measurements predicted to fall outside the observation window for the optimal fit is given as an inset. This should be compared to $N_{\text{cut}} = 1298$ in the fitted data set. **d** Histogram of estimates for the short timescale generated over 10,000 bootstrapped data sets. **e** Histogram of estimates for the long timescale generated over 10,000 bootstrapped data sets. **f** Histogram of estimates of the fraction of unbinding times originating in the short timescale, generated from 10,000 bootstrapped data sets. The parameter distributions vary significantly between data sets, even though all fits look plausible in **a–c**

the lines of Eq. (5.7)

$$L^{\text{ML}}(\tau_1, \tau_2, P_1) = - \sum_{n=1}^{N_{\text{rec}}} \ln \left(\frac{P_1}{\tau_1} e^{-t_n/\tau_1} + \frac{1-P_1}{\tau_2} e^{-t_n/\tau_2} \right) - N_{\text{cut}} \ln \left(P_1 e^{-T_{\text{cut}}/\tau_1} + (1-P_1) e^{-T_{\text{cut}}/\tau_2} \right). \quad (5.14)$$

The information regarding N_{cut} is ignored in standard uwLS and wLS approaches, where the data is binned and fitted based on Eq. (5.9) only within the window capturing the N_{rec} unbinding times.

As can be seen in Fig. 5.8a–c, the three methods considered give very different results, all naively appearing to describe the data well. Lacking an objective way to evaluate the goodness of fit across scenarios, we can only point to the fact that our general developments and our numerical investigation suggest that the ML approach gives the best estimate of the fit parameters.

The insets in Fig. 5.8a–c report the average number of measurements that the best fit predicts should fall outside the measurement window. This average should be compared to the $N_{\text{cut}} = 1298$ measurements that actually fell outside the observation window. From this, it is clear that the extra information included in the ML estimation regarding the cut data does increase its predictive capabilities in this case, which was not visibly the case for the fits in Fig. 5.6c, d.

5.5.2 *Bootstrapping: Doing the Best We Can with Limited Resources*

To determine the standard deviation of our parameter estimates, we would ideally like to establish their distribution by repeating the same experiment many times—much like we did in our earlier numerical comparison between estimation methods. A common practice is to report the standard deviation of fit estimates over a triplicate of identical experiments. However, not having a statistically significant sample can result in significant errors in estimating the standard deviation. Unfortunately, repeating the same experiment a sufficient number of times is often too time-consuming and costly, and we have to rely on other means.

If we could perform repeat experiments, we would in effect draw new unbinding times from the *true* PDF describing the unbinding kinetics. Instead of repeating the experiments by drawing from the true PDF, we here repeatedly draw from our best estimate of the true PDF: the original data set. This approach is called bootstrapping the data [5]. To generate each “new experiment,” we randomly draw N unbinding times from our original data set (also of size N), *allowing for repeated draws* of the same data instance (this is known as random sampling with replacement). We then fit our bootstrapped data set in the same manner as we fit our original data sets. By repeating this process many times, we build up the desired distributions of fit

parameters. In Fig. 5.8d–f, the distributions of the double-exponential fit parameters are plotted, using uwLS, wLS, and cgML methods over 10,000 bootstrapped data sets.

Contrary to the situation with our simulated data sets, we here do not know the true values of the model parameters and so cannot establish the bias nor the MSE and thus lack an objective metric by which to compare the different approaches. In light of this, it is important to stress that the fact that the standard deviation is consistently smallest for uwLS is not a good argument for why this approach should be preferable. Given the disparate results of the various methods—even though all fits naively look good (Fig. 5.8a–c)—it is clear that at least two of the three methods can go astray in very non-obvious ways, and that caution is warranted. Our heuristic arguments and simulations suggest ML estimation to be generally preferable.

5.6 Conclusion

We have provided an introduction to ML estimation as a powerful alternative to conventional LS fitting methods. Focusing on exponential distributions as examples, we showed how the ML method provides a general way to estimate the model parameters from stochastic data, in principle without the need for binning. We also showed that uwLS, wLS, and ML can all be thought of as approximations to tLS, utilizing various estimates for the a priori unknown standard deviation of bin counts. The main upshots of both our heuristic argument and numerical investigation are:

1. wLS becomes unreliable as soon as there are bins with low counts, as should always be expected in the tail end of distributions without a severe experimental cutoff time.
2. uwLS often outperforms wLS for processes with a single characteristic time, but for processes with multiple characteristic times, it becomes unreliable as it fails to appropriately weigh the contribution of data on different timescales.
3. (cg)ML consistently outperforms both wLS and uwLS by estimating bin-count variations from the whole data set, rather than ignoring them (uwLS) or estimating them on a bin-to-bin basis (wLS).

The two first points significantly limit the global applicability of both uwLS and wLS methods. The maximum-likelihood method is generally applicable though, needs no binning—but if binned, is not sensitive to empty bins—and outperforms both uwLS and wLS in all examples discussed. Although we focused on exponentially distributed data, our conclusions are general and should apply irrespective of the particular distribution describing the data. These advantages, together with the adaptability of the approach, have convinced the authors that ML estimation is the preferable choice for dealing with SM data; we hope our presentation has gone some way toward convincing the reader of the same.

Acknowledgements We thank Tao Ju (Thijs) Cui, Misha Klein, and Olivera Rakic for careful reading of the manuscript and thoughtful feedback. B. Eslami-Mosallam acknowledges financial support through the research program *Crowd management: the physics of genome processing in complex environments*, which is financed by the Netherlands Organisation for Scientific Research. I. Katechis acknowledges financial support from the Netherlands Organisation for Scientific Research, as part of the Frontiers in Nanoscience program.

References

1. Aartsen, M. G., Abraham, K., Ackermann, M., Adams, J., Aguilar, J. A., Ahlers, M., et al. (2015). A combined maximum-likelihood analysis of the high-energy astrophysical neutrino flux measured with icecube. *Astrophysical Journal*, *809*(1), 1–15.
2. Avdis, E., & Wachter, J. A. (2017). Maximum likelihood estimation of the equity premium. *Journal of Financial Economics*, *125*(3), 589–609.
3. Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *5*(2), 179–190.
4. Dulin, D., Berghuis, B. A., Depken, M., & Dekker, N. H. (2015). Untangling reaction pathways through modern approaches to high-throughput single-molecule force-spectroscopy experiments. *Current Opinion in Structural Biology*.
5. Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. Chapman & Hall.
6. Fareh, M., van Lopik, J., Katechis, I., Bronkhorst, A. W., Haagsma, A. C., van Rij, R. P., & Joo, C. (2018). Viral suppressors of RNAi employ a rapid screening mode to discriminate viral RNA from cellular small RNA. *Nucleic Acids Research* (March), 1–11.
7. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376.
8. Forney, G. D. (1972). Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. *IEEE Transactions on Information Theory*, *18*(3), 363–378.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. The mathematical intelligencer* (2nd ed.). New York: Springer.
10. Hauschild, T., & Jentschel, M. (2001). Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments. *Nuclear Instruments and Methods in Physics Research A*, *457*(1–2), 384–401.
11. Humphrey, P. J., Liu, W., & Buote, D. A. (2009). χ^2 and Poissonian data: BIASES even in the high-count regime and how to avoid them. *The Astrophysical Journal*, *693*(1), 822–829.
12. Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
13. Johansen, S., & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration—With applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, *52*(2), 169–210.
14. Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., & Ha, T. (2008). Advances in single-molecule fluorescence methods for molecular biology. *Annual Review of Biochemistry*, *77*, 51–76.
15. Joo, C., & Ha, T. (2012). Single-molecule FRET with total internal reflection microscopy. *Cold Spring Harbor Protocols*, *7*(12), 1223–1237.
16. Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, *9*(2), 171–185.
17. Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica. Section D, Biological Crystallography*, *53*(3), 240–255.

18. Nelson, P. (2015). *Physical models of living systems*. New York: W. H. Freeman.
19. Nishimura, G., & Tamura, M. (2005). Artefacts in the analysis of temporal response functions measured by photon counting. *Physics in Medicine & Biology*, *50*(6), 1327–1342.
20. Nørrelykke, S. F., & Flyvbjerg, H. (2010). Power spectrum analysis with least-squares fitting: Amplitude bias and its elimination, with application to optical tweezers and atomic force microscope cantilevers. *Review of Scientific Instruments*, *81*(7).
21. Nousek, J. A., & Shue, D. R. (1989). Chi-squared and C statistic minimization for low count per bin data. *Astrophysical Journal*, *342*, 1207–1211.
22. Press, W., Teukolsky, S., Vetterling, W., Flannery, B., Ziegler, E., Press, W., et al. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). Cambridge: Cambridge University Press.
23. Santra, K., Zhan, J., Song, X., Smith, E. A., Vaswani, N., & Petrich, J. W. (2016). What is the best method to fit time-resolved data? A comparison of the residual minimization and the maximum likelihood techniques as applied to experimental time-correlated, single-photon counting data. *Journal of Physical Chemistry B*, *120*(9), 2484–2490.
24. Scholten, T. L., & Blume-Kohout, R. (2018). Behavior of the maximum likelihood in quantum state tomography. *New Journal of Physics*, *20*, 023050.
25. Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690.
26. Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, *44*(W1), W232–W235.
27. Turton, D. A., Reid, G. D., & Beddard, G. S. (2003). Accurate analysis of fluorescence decays from single molecules in photon counting experiments. *Analytical Chemistry*, *75*(16), 4182–4187.
28. Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, *18*(5), 691–699.