Nonparametric density estimation and optimal bandwidth selection for protein unfolding and unbinding data

E. Bura,[1] A. Zhmurov,[2] and V. Barsegov[2,a)]
[1]*Department of Statistics, George Washington University, Washington, D.C. 20052, USA and Biometrics, Vertex Pharmaceuticals, 130 Waverly St., Cambridge, Massachusetts 02139, USA*
[2]*Department of Chemistry, University of Massachusetts, Lowell, Massachusetts 01854, USA*

(Received 13 October 2008; accepted 16 November 2008; published online 7 January 2009)

Dynamic force spectroscopy and steered molecular simulations have become powerful tools for analyzing the mechanical properties of proteins, and the strength of protein-protein complexes and aggregates. Probability density functions of the unfolding forces and unfolding times for proteins, and rupture forces and bond lifetimes for protein-protein complexes allow quantification of the forced unfolding and unbinding transitions, and mapping the biomolecular free energy landscape. The inference of the unknown probability distribution functions from the experimental and simulated forced unfolding and unbinding data, as well as the assessment of analytically tractable models of the protein unfolding and unbinding requires the use of a *bandwidth*. The choice of this quantity is typically subjective as it draws heavily on the investigator's intuition and past experience. We describe several approaches for selecting the "optimal bandwidth" for nonparametric density estimators, such as the traditionally used histogram and the more advanced kernel density estimators. The performance of these methods is tested on unimodal and multimodal skewed, long-tailed distributed data, as typically observed in force spectroscopy experiments and in molecular pulling simulations. The results of these studies can serve as a guideline for selecting the optimal bandwidth to resolve the underlying distributions from the forced unfolding and unbinding data for proteins. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3050095]

I. INTRODUCTION

Mechanical functions of intra- and extracellular proteins play an essential role in diverse biological processes, from ubiquitin-substrate protein degradation[1,2] to cytoskeleton support and cell motility,[3,4] to cell adhesion and formation of extracellular matrix,[5–9] to muscle contraction and relaxation,[10,11] to membrane transport,[12] and to blood clotting.[13–15] Recent advances in dynamic force spectroscopy, which utilize atomic force microscopy[5,16,17] (AFM) and laser and optical tweezer-based force spectroscopy,[18–20] and biomembrane force probes,[21–23] have enabled researchers to study the mechanical properties and unfolding pathways of proteins,[24–27] and the strength of protein-protein complexes and ligand-receptor noncovalent bonds.[5,22,23] These experiments allow one to access the entire distribution of molecular characteristics,[28–30] such as force-induced protein elongation,[17,31,32] unfolding forces and unfolding times[31,33–35] for proteins, and noncovalent bond lifetimes and rupture forces for protein-protein complexes[5,16,18,22] and aggregates.[36–39]

The global unfolding and unbinding transitions in proteins are described by single-step kinetics, $F \rightarrow U$ ($B \rightarrow U$), where F (B) denotes the folded (bound) state and U represents the unfolded (unbound) state, which correspond to *unimodal distributions* of unfolding forces, unfolding times, rupture forces, and bond lifetimes. However, the unfolding and unbinding pathways may also involve formation of intermediate states or dynamic coupling among competing unfolding and unbinding scenarios. For example, recent AFM experiments and computer simulation studies of green fluorescent protein (GFP) revealed the bifurcation in the GFP unfolding pathways.[27] A similar kinetic switch in the unfolding pathways has also been reported for α- and β-tubulin.[40] The transition from "catch" bond to "slip" bond in cell-adhesion complexes between the *P*-, *L*-, and *E*-selectin receptors and their ligands (PSGL-1 and endoglycan), is mediated by the dynamic competition for the forced dissociation from the high affinity bound state (catch bond) and the low affinity bound state (slip bond) of the complex.[5,9,16,41] The interplay between unfolding and unbinding pathways can also be controlled by the amplitude and direction of the applied pulling force. For example, computer simulations of the force-induced dissociation of Aβ peptides from amyloid fibrils, also studied experimentally,[36,37] showed that the dissociation mechanism is highly anisotropic as it depends on whether the pulling force is applied in parallel or perpendicular direction with respect to the Aβ fibril axis.[39] Computer simulations of the forced unfolding of protein tandems show that uncorrelated unfolding transitions of individual protein domains, observed at low forces, may become correlated (dependent) at elevated force levels.[42,43] These examples show that, due to the complexity of the free energy landscape reflected in the multitude of possible unfolding or unbinding transitions, the molecular characteristics of proteins can also be distributed in a *multimodal fashion*.[44]

a)Author to whom correspondence should be addressed. Tel.: 978-934-3661. FAX: 978-934-3013. Electronic mail: valeri_barsegov@uml.edu.

The main goal of statistical analyses of the forced unfolding and unbinding data is to infer the probability density function (pdf) $f(x)$ of a continuous random variable $x=g$ (unfolding forces or rupture forces), or $x=t$ (unfolding times or bond lifetimes). Because, the pdf is, in general, unknown, random samples of data are collected in order to obtain "snapshots" of the population. To describe the pdf $f(x)$, one may have to collect an infinite number of observations. In practice, obtaining large samples is both tedious and costly, sometimes even infeasible. In density estimation, a random sample of moderate size (few hundreds of data points) is drawn from the population and is used to estimate the true population pdf $f(x)$. When a model for the pdf is not available, *nonparametric* density estimation methods are used. Their defining characteristic is that they are fully data driven and that the density estimate at a data point is computed by weighting the points in its neighborhood. The size of the latter is called bandwidth or bin size. Its choice is the central issue in nonparametric density estimation. For example, the histogram, which is the classical and simplest nonparametric density estimator,[45] requires choosing the number of bins or, equivalently, the bin size. Most authors advise that 5–20 bins are sufficient to describe the data;[46] yet, the number of bins chosen may result not only in markedly different but also highly subjective density estimates.

The three widely used "rule-of-thumb" criteria for choosing the bin size for the histogram are due to Sturges, Scott, and Freedman and Diaconis. The Sturges rule[47] uses the Gaussian density as the reference distribution to select the optimal number of bins as $n_{\mathrm{opt}}=1+\log_2 n$, where n is the number of observation, and the optimal bin size

$$h_{\mathrm{opt}}^{\mathrm{St}}=\frac{x_{\max}-x_{\min}}{n_{\mathrm{opt}}}, \tag{1}$$

where $x_{\max}$ ($x_{\min}$) is the largest (smallest) data value. For Scott's rule,[48] which also uses the normal as reference distribution, the optimal bin size is

$$h_{\mathrm{opt}}^{\mathrm{Sc}}=3.5\sigma_x n^{-1/3}, \tag{2}$$

where σ_x is the standard deviation of the data. Both these methods assume the underlying true pdf to be unimodal and symmetric with short tails. Typically, the forced unfolding and unbinding data for proteins are asymmetric and skewed toward longer unfolding or unbinding times and smaller unfolding or rupture forces. Freedman and Diaconis[49] proposed a more robust rule for h_{opt} against outlying observations and lack of symmetry by replacing σ_x in Eq. (2) with the interquartile range (IQR=difference of the 75th and 25th percentiles), so that

$$h_{\mathrm{opt}}^{\mathrm{FD}}=2\mathrm{IQR}n^{-1/3}. \tag{3}$$

To illustrate the effect of the choice of the bin size on the shape of the histogram, we performed Monte Carlo (MC) simulations of the exponential and gamma probability densities, defined by

$$f_E(x)=ke^{-kx} \quad \text{and} \quad f_G(x)=k^{\alpha}x^{\alpha-1}e^{-kx}/\Gamma(\alpha), \tag{4}$$

respectively, where k is the decay rate, α is the shape parameter, and $\Gamma(\alpha)$ is the gamma function. The exponential density, corresponding to single-step unfolding (unbinding) kinetics, $F\rightarrow U$, $(B\rightarrow U)$, is used to describe the forced unfolding times of proteins[25,32,50–52] and bond lifetimes of protein-protein complexes and aggregates.[5,9,16,18,19,41] The gamma density can be used to describe the forced unfolding times for individual protein domains in protein tandems.[42,43] MC simulations are carried out as follows. At each time step Δx, the decay probability $F_E(\Delta x)=1-\exp[-k\Delta x]$ for the exponential density $f_E(x)$ and $F_G(\Delta x)=1-\Gamma(\alpha,k\Delta x)/\Gamma(\alpha)$ for the Gamma density is compared to a uniformly distributed random number F_{ran} from the unit interval $F_{\mathrm{ran}}\in[0,1]$. The decay time $x=s\Delta x$, where s is the number of MC steps, is defined as the instant at which $F_{\mathrm{ran}}<F_{E,G}(\Delta x)$ for the first time. The histograms constructed by using Sturges' rule [Eq. (1)], Scott's rule [Eq. (2)], and Freedman–Diaconis' rule [Eq. (3)] are displayed in Fig. 1. Sturges' rule [Figs. 1(c) and 1(h)] and Scott's rule [Figs. 1(d) and 1(i)] result in smaller numbers of bins (9 and 10) than needed to describe the skewness of the data. Freedman–Diaconis' rule uses 15 and 14 bins and the resulting histograms appear to be closer to the true density [Figs. 1(e) and 1(j)]. We also display histograms with 5 and 20 bins for $f_E(x)$ [Figs. 1(a) and 1(b)], and 5 and 35 bins for $f_G(x)$ [Figs. 1(f) and 1(g)]. The agreement between the underlying pdfs and the histograms varies greatly with the number of bins: using too few bins negatively affects estimation accuracy, and using too many bins results in a noisy histogram that hides the shape of the underlying pdf.

Nonparametric density estimators present considerable improvement over histograms. The most widely used kernel density estimators[53–55] smooth out the contribution of each data point over a neighborhood of that point, and the contribution of a data point to the density estimate at some other point is controlled by a weight function for the distance between the points. As for the histogram, the choice of the bandwidth is most important, especially when assessing a particular model of forced unfolding or unbinding. In this paper, we present a fairly comprehensive study of optimal bandwidth selection methods for constructing nonparametric estimates, such as histograms and kernel density estimates, of the pdfs of unfolding forces and unfolding times for proteins, and bond lifetimes and rupture forces for protein-protein complexes. We employ statistical measures of estimation accuracy, such as the squared error loss function, mean squared error (MSE), and cross validation, to select the optimal bandwidth. These measures are used to assess and compare the performance of histograms and several kernel density estimates at describing the unimodal and multimodal distributions of the unfolding times and unfolding forces, and rupture forces and bond lifetimes. We also introduce and implement *adaptive* bandwidth selection that is shown to be more appropriate for resolving multimodal distributions.

II. STATISTICAL MEASURES OF ESTIMATION ACCURACY

Statistical measures of estimation accuracy assess how well an estimator $\hat{f}(x)$ approximates the true density $f(x)$. For an estimate $\hat{f}(x)$ of the density $f(x)$, the L_2 distance based *squared error loss* function is defined to be $L(f(x),\hat{f}(x))$

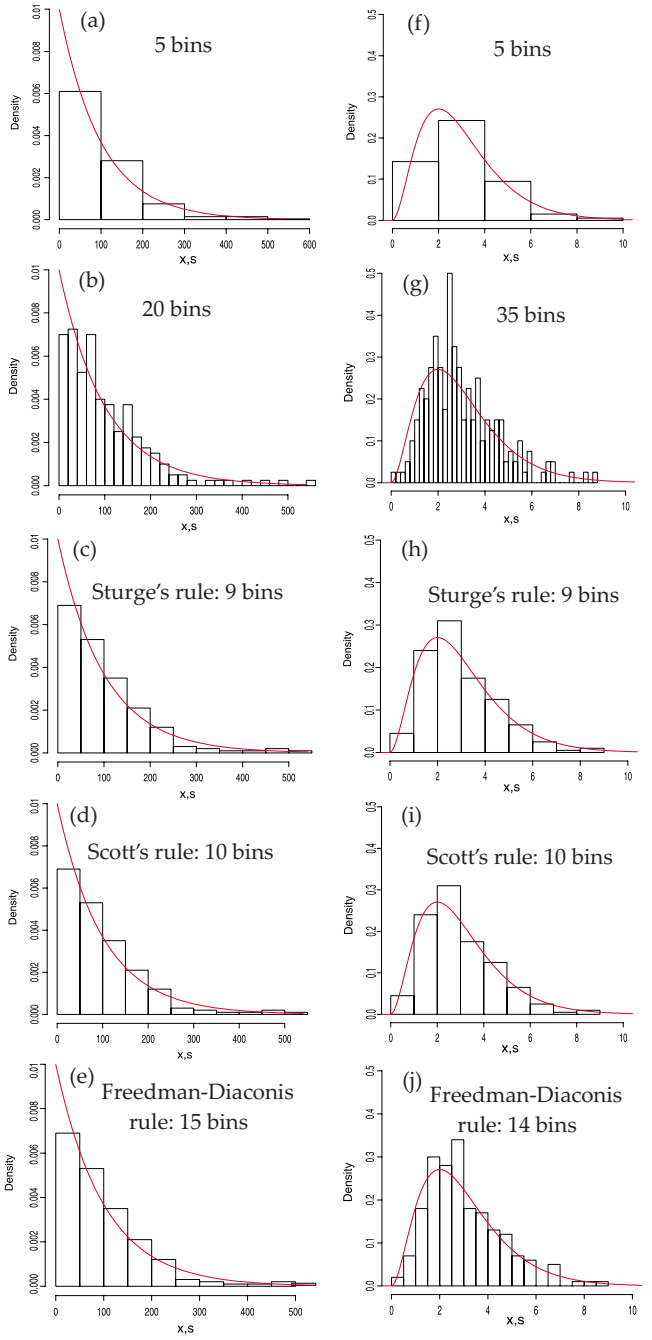

FIG. 1. (Color online) Histograms of the 200 data points (bars) sampled from the exponential density with the decay rate $k=0.01$ s^{-1} [(a)–(e)], and from gamma density [Eq. (4)] with the shape parameter $\alpha=3$ and the decay rate $k=1.0$ s^{-1} [(f)–(j)], generated by carrying out MC simulations (with integration step $\Delta x=10^{-6}$ s). The histograms, overlaid with the exact pdfs (curves), are obtained by using Sturge's rule [Eq. (1), (c) and (h)], Scott's rule [Eq. (2), (d) and (i)], and Freedman–Diaconis rule [Eq. (3), (e) and (j)] for h_{opt}, are compared to the histograms constructed by using too few bins [(a) and (f)] and too many bins [(b) and (g)].

$=(f(x)-\hat{f}(x))^2$. The MSE is defined to be the expected (average) value of the loss function,

$$\text{MSE}[\hat{f}(x)] = E[(\hat{f}(x)-f(x))^2] = \text{var}[\hat{f}(x)] + \text{bias}[\hat{f}(x)]^2. \tag{5}$$

In Eq. (5), $\text{var}[\hat{f}(x)]=E[\hat{f}(x)-E[\hat{f}(x)]]^2$ is the *variance* of $\hat{f}(x)$ around its expected value $E[\hat{f}(x)]$, and $\text{bias}[\hat{f}(x)]=E[\hat{f}(x)]-f(x)$, is the *bias* of $\hat{f}(x)$. The MSE is a pointwise error criterion. The *mean integrated squared error* (MISE),

$$\text{MISE}[\hat{f}(x)] = \int_0^\infty dx \text{MSE}[\hat{f}(x)] = \int_0^\infty dx \text{var}[\hat{f}(x)] + \int_0^\infty dx \text{bias}[\hat{f}(x)]^2, \tag{6}$$

is a global criterion. In Eq. (6), the integrated variance, $\text{IV}=\int_0^\infty dx \text{var}[\hat{f}(x)]$, is a measure of random variation about the mean of the data, and the integrated squared bias, $\text{ISB}=\int_0^\infty dx \text{bias}[\hat{f}(x)]^2$, is an overall measure of the bias of the estimate. An estimator $\hat{f}(x;h)$ is said to be consistent if $\text{MISE}[\hat{f}]\to 0$, as $n\to\infty$ and $h\to 0$.[56] The *asymptotic MISE* (AMISE) is a second order approximation to MISE given by the sum of the first two terms of the Taylor series expansion of MISE around $f(x)$. An estimator $\hat{f}(x;h)$ is consistent in AMISE, if $\text{AMISE}[\hat{f}]\to 0$, as $n\to\infty$ and $h\to 0$. In terms of these measures, the optimal bandwidth (bin size for the histogram) is selected so that it minimizes MISE or AMISE. The main challenge in nonparametric density estimation is the bias-variance trade-off. In both MISE and AMISE, bias and variance depend upon the bandwidth h that controls the overall "smoothness" of the shape of the estimator. When the data are oversmoothed, i.e., $h\to\infty$ [Figs. 1(a) and 1(f)], the bias is large but the variance is small, whereas when the data are undersmoothed, i.e., $h\to 0$ [Figs. 1(b) and 1(g)], the bias is small but the variance is large. Therefore, selecting h_{opt} by minimizing either MISE or AMISE amounts to balancing bias and variance *at the same time*.

We will show in Sec. III that the computation of the optimal MISE and AMISE bandwidth requires knowledge of the unknown true density $f(x)$. Yet, the sole purpose of a density estimator $\hat{f}(x)$ is to use it to infer $f(x)$. This hurdle can be overcome by using a reference distribution for $f(x)$ ("plug-in" method), such as the Gaussian or the exponential. One can also use a fully data based measure of accuracy, such as the cross-validation (CV) estimator, defined by

$$\text{CV}[h] = \int_0^\infty \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i), \tag{7}$$

where $\hat{f}_{-i}$ is the estimator obtained by removing the ith observation. CV estimators are data resampling based estimators of error.[57–59] Equation (7) is motivated by noting that in the integral of the squared error loss function, $\int L(f(x),\hat{f}(x))dx=\int\hat{f}^2(x)dx-2\int\hat{f}(x)f(x)dx+\int f^2(x)dx$, the last term does not depend on h, and so, minimizing the loss is equivalent to minimizing the expected value of $J(h)=\int\hat{f}^2(x)dx-2\int\hat{f}(x)f(x)dx$. That is, $\text{CV}[h]$ in Eq. (7) is the CV estimator of $\text{E}[J(h)]$.[57]

III. OPTIMAL BIN SIZE SELECTION FOR THE HISTOGRAM

A. MSE, MISE, AMISE, and CV of the histogram

First, we consider general rules for computing the optimal bin size for a histogram, which describes the unimodal unfolding and unbinding data. When the forced unfolding and unbinding transitions in proteins follow single-step kinetics, the unfolding and unbinding data are described by skewed unimodal distributions. Suppose $B_k=[x_0+(k-1)h, x_0+kh)$ is the kth bin, where h is the bin size and x_0 is the position of the first bin. Let n_k denote the bin count, i.e., the number of points that fall in B_k. Then, the histogram at $x \in B_k$ is defined as

$$\hat{f}_h(x) = \frac{n_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} 1\{x_i \in B_k\}. \tag{8}$$

Since the bin count n_k is a Binomial random variable, i.e., $n_k \sim B(n,p_k)$, where $p_k=\int_{B_k} f(x)dx$ is the probability that a point falls in B_k, $E[n_k]=np_k$ and $\mathrm{var}[n_k]=np_k(1-p_k)$. The bias of the histogram is $\mathrm{bias}[\hat{f}_h(x)]=E[\hat{f}(x)]-f(x)=E[n_k]/nh-f(x)=p_k/h-f(x)=\int_{B_k} dtf(t)/h-f(x)=\int_{B_k} dt(f(t)-f(x))/h$. Using the first order Taylor expansion of $f(x)$, we obtain

$$E[\hat{f}_h(x)]-f(x) = \frac{1}{h}\int_{B_k} dtf'(x)(t-x)+o(h) \approx f'(x)h+o(h). \tag{9}$$

The variance of the histogram is $\mathrm{var}[\hat{f}_h(x)]=np_k(1-p_k)/n^2h^2=\int_{B_k} dtf(t)(1-\int_{B_k} dtf(t))/nh^2$, and since $\int_{B_k} f(t)dt=\int_{(k-1)h}^{kh} dt(f(x)+f'(x)(t-x)+o(h))=f(x)h+hf'(x)[(k-1/2)h-x]+o(h)$,

$$\mathrm{var}[\hat{f}_h(x)] = \frac{1}{nh}\left(\frac{1}{h}\int_{B_k} dtf(t)\right)[1-O(h)] = \frac{f(x)}{nh}+o\left(\frac{1}{nh}\right). \tag{10}$$

By substituting Eqs. (9) and (10) into Eq. (5), we obtain the MSE for a histogram,

$$\mathrm{MSE}[\hat{f}_h(x)] = (f'(x)h)^2+o(h^2)+\frac{f(x)}{nh}+o\left(\frac{1}{nh}\right). \tag{11}$$

It follows from Eq. (11) that $\mathrm{MSE}[\hat{f}_h(x)]\to 0$ if the sequence $h(n)\to 0$ as $n\to\infty$, in which case $\hat{f}(x)\to f(x)$, and $\hat{f}(x)$ is a consistent estimator of $f(x)$.

Equations (9)–(11) imply that when h is large, the variance and IV are small, whereas when h is small, the bias and ISB are small. Therefore, the optimal bin size, h_{opt}, should balance both components, $\mathrm{bias}[\hat{f}_h(x)]$ and $\mathrm{var}[\hat{f}_h(x)]$, or $\mathrm{ISB}[\hat{f}_h(x)]$ and $\mathrm{IV}[\hat{f}_h(x)]$ at the same time. The fastest rate of convergence of $\hat{f}(x)$ to $f(x)$ can be achieved when bias and variance approach zero at the same speed; otherwise, the slower rate dominates. Therefore, the optimal bin size sequence, h_{opt}, should satisfy the condition, $o(h_{\mathrm{opt}}^2)=o(1/nh_{\mathrm{opt}})$, or equivalently, $h_{\mathrm{opt}}=o(1/n^{1/3})$, which guarantees the fastest rate of convergence of $\mathrm{MSE}[\hat{f}_h(x)]$ to zero and $\hat{f}(x)$ to $f(x)$, i.e., $\mathrm{MSE}[\hat{f}(x);h_{\mathrm{opt}}]=o(1/n^{2/3})$. By using the fastest rate of convergence in Eq. (11), we obtain

$$\lim_{n\to\infty} n^{2/3}\mathrm{MSE}[\hat{f}(x);h_{\mathrm{opt}}] = (f'(x)h_{\mathrm{opt}})^2+\frac{f(x)}{h_{\mathrm{opt}}}, \tag{12}$$

and by taking the derivative of the right hand side of Eq. (12), setting it to zero, and solving for $h_{\mathrm{opt}}(x)$, we find the minimizer of MSE for a histogram

$$h_{\mathrm{opt},H}^{\mathrm{MSE}}(x) = \left(\frac{2f(x)}{nf'(x)^2}\right)^{1/3}. \tag{13}$$

By using Eqs. (6) and (11) we obtain the MISE for a histogram,

$$\mathrm{MISE}[\hat{f}_h(x)] = \int_0^\infty dx\left((f'(x)h)^2+\frac{f(x)}{nh}+o(h^2)+o\left(\frac{1}{nh}\right)\right), \tag{14}$$

so that the minimizer of MISE for a histogram reads[56]

$$h_{\mathrm{opt},H}^{\mathrm{MISE}} = \left(\frac{2}{nR(f')}\right)^{1/3}, \tag{15}$$

where $R(f)=\int_0^\infty dxf^2(x)$ is the roughness function. It follows from Eq. (14) that if $h\to 0$ and $n\to\infty$ $(nh\to 0)$ $\mathrm{MISE}[h]\to 0$. Finally, the AMISE for a histogram is given by[60]

$$\mathrm{AMISE}[\hat{f}_h(x)] = \int_0^\infty dx\left((f'(x)h)^2+\frac{f(x)}{nh}\right). \tag{16}$$

It follows from Eq. (16) that if $h\to 0$ and $nh\to 0$, $\mathrm{MISE}[h]=\mathrm{AMISE}[h]=o(\mathrm{AMISE}[h])$. $\mathrm{AMISE}[\hat{f}_h]$ in Eq. (16) can be expressed as $\mathrm{AMISE}[\hat{f}_h]=1/nh+h^2R(f')/12$,[48,49] where the first term, IV, is of order $O(h^{-1})$, whereas the second term, ISB, is of order $O(h^2)$. By setting the derivative of AMISE with respect to h to zero and solving for $h_{\mathrm{opt}}^{\mathrm{AMISE}}$, we obtain

$$h_{\mathrm{opt},H}^{\mathrm{AMISE}} = 3^{1/3}\left(\frac{2}{nR(f')}\right)^{1/3}. \tag{17}$$

The optimal error of AMISE for $h=h_{\mathrm{opt}}^{\mathrm{AMISE}}$ is $\mathrm{AMISE}[\hat{f}(x);h_{\mathrm{opt}}]=(3^{2/3}4n^{-2/3})[R(f')]^{1/3}$. Thus, the histogram constructed by using $h_{\mathrm{opt}}^{\mathrm{AMISE}}$ converges to the true density at the rate of $O(n^{-2/3})$.

Equations (13), (15), and (17) show that in order to obtain h_{opt} for any of the three criteria, prior knowledge of the unknown density or its derivative is needed. Even when the analytical form of $f(x)$ is known, $f(x)$ and $f'(x)$ typically depend on model parameters, whose values may not be known. It is, thus, important to have a statistical tool for estimating h_{opt} that does not depend on $f(x)$ or $f'(x)$. By using the CV criterion [Eq. (7)], we obtain[59]

$$\mathrm{CV}[h] = \frac{2}{h(n-1)}-\frac{n+1}{h(n-1)}\sum_{k=1}^{m}\frac{n_k^2}{n^2}, \tag{18}$$

where m is number of bins, $h=(x_{\max}-x_{\min})/m$, and n_k is the kth bin count. The optimal bin size can be computed by

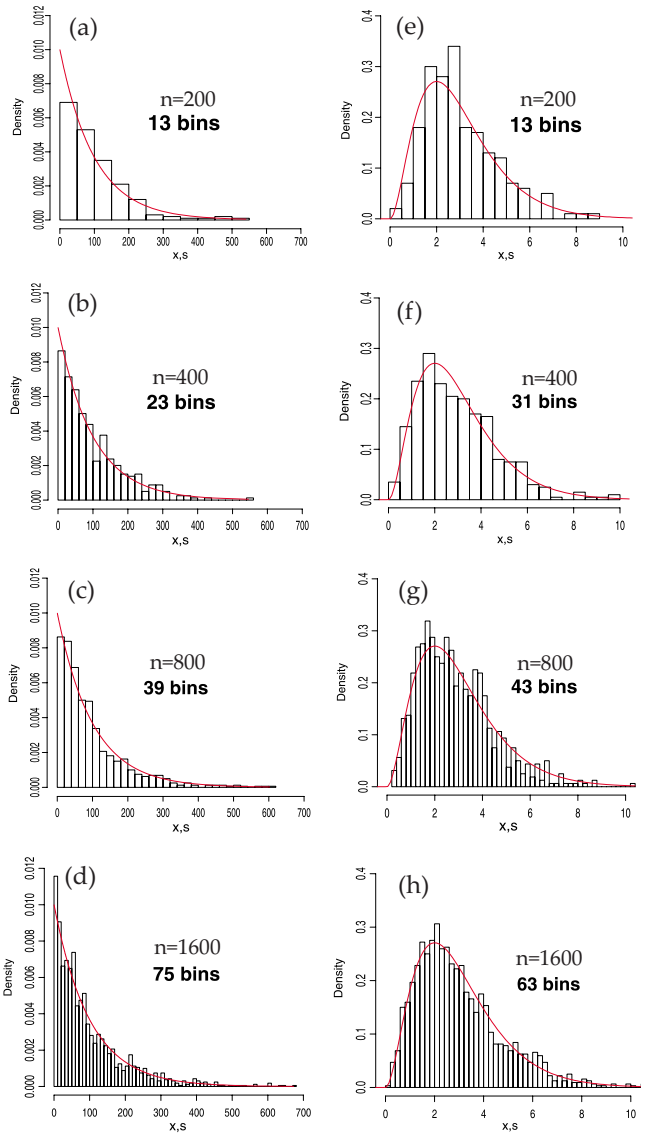

FIG. 2. (Color online) Histograms of the data, obtained by using MC simulations (Sec. I) of the exponential density [(a)–(d)] and gamma density [(e)–(h)] with integration step $\Delta x=10^{-6}$ s. Parameters of the models are given in the caption to Fig. 1. The histograms (bars), constructed by using the CV criterion [Eq. (18)] for $h_{\mathrm{opt}}^{\mathrm{CV}}$, are compared to the exact pdf curves. The number of data points n (sample size) are indicated in the plots.

using a grid search as follows. First, the number of bins is set to $m=m_1=2$ or 3, and the bin size is computed as $h_1=(x_{\max}-x_{\min})/m_1$, which is used to calculate the CV$[h_1]$ value [Eq. (18)]. Next, the number of bins is increased by $k=1$, 2, or 3 (for an exhaustive search, $k=1$), and $m_2=m_1+k$, for which the value of $h_2=(x_{\max}-x_{\min})/m_2$ is used to compute CV$[h_2]$. The process is repeated up to $m_{\max}\leq n$. The value of h_i that corresponds to the minimum of CV$[h]$ defines h_{opt}. For several minima, the largest bin size, $h_{\mathrm{opt}}^{\max}=\max\{h_{\mathrm{opt},1}^{\mathrm{CV}},h_{\mathrm{opt},2}^{\mathrm{CV}},\ldots,h_{\mathrm{opt},s}^{\mathrm{CV}}\}$, is used. We employed this procedure to estimate $h_{\mathrm{opt}}^{\mathrm{CV}}$ in order to construct the histograms for the exponential and Gamma densities [Eqs. (4)], analyzed in Fig. 1. For 200 data points (Fig. 2), 13 bins are used in the histograms of $f_E(x)$ and $f_G(x)$ constructed using the CV-based approach [Figs. 2(a) and 2(e)]. In contrast, the corresponding histograms of $f_E(x)$ and $f_G(x)$ are based on 9 bins [Sturge's rule, see Figs. 1(c) and 1(h)] and 10 bins [Scott's rule, see Figs. 1(d) and 1(i)]. For both densities, the CV-based optimal number of bins is closer to the Freedman–Diaconis' optimal number of bins, i.e., 15 bins for $f_E(x)$ [Fig. 1(e)] and 14 bins for $f_G(x)$ [Fig. 1(j)]. Both $f_E(x)$ and $f_G(x)$ are now resolved better but appear rough in the tails due to data paucity.

B. Adaptive MSE, MISE, AMISE, and CV criteria for multimodal densities

In a large variety of proteins and protein-protein complexes, the forced unfolding and unbinding transitions occur through multiple pathways, which result in *multimodal* distributions of unfolding forces, unfolding times and bond lifetimes, and rupture forces.[5,9,16,41] Resolving the multimodal shape of the density is a challenging task since individual contributions from unfolding or unbinding pathways and nonspecific interaction may partially overlap, resulting in observed broad distributions with long tails. In such cases, the issue of optimal bandwidth selection is crucial, as inaccurate description of the data may result in false conclusions about the underlying mechanism(s) of the protein unfolding or unbinding. To overcome this problem, we propose to use histograms with "adaptive bin size."

Consider a kinetic model that describes the forced unfolding transitions in a protein which occur through two competing unfolding pathways I and II,

$$\text{(I)}\quad U \leftarrow F \rightarrow U \quad \text{(II)}, \tag{19}$$

or a model for the forced rupture of a ligand-receptor complex LR between a receptor R and a ligand L that accounts for two coupled unbinding pathways (I and II),

$$\text{(I)}\ R+L \leftarrow (L\cdot R)_1 \rightleftarrows (L\cdot R)_2 \rightarrow L+R \quad \text{(II)}. \tag{20}$$

The kinetic scheme (19) can be used, e.g., to model the two unfolding pathways for α- and β-tubulins which correspond the unraveling of protein domains initiated in the C-terminal (pathway I) and N-terminal (pathway II),[40] or the bifurcation in the GFP unfolding pathways.[27] The kinetic scheme (20) has been used to describe the forced rupture of cell-adhesion complexes between the P- and L-selectin receptors and the ligand PSGL-1.[5,9,16,41] Corresponding to the kinetic schemes (19) and (20), the distributions of unfolding forces and rupture forces ($x=g$), or unfolding times, and bond lifetimes ($x=t$) are characterized by the bimodal pdf,

$$f(x)=\alpha_1(x)f_1(x)+\alpha_2(x)f_2(x), \tag{21}$$

where the unimodal density $f_1(x)$ ($f_2(x)$) for pathway I (II) is weighted by the function $\alpha_1(x)$ ($\alpha_2(x)$, Appendix A). Using Eq. (11), we obtain the MSE criterion for the bimodal pdf,

$$\mathrm{MSE}[\hat{f}_h(x)]=(\alpha_1(x)f_1'(x)h+\alpha_2(x)f_2'(x)h)^2+o(h^2)+\frac{\alpha_1(x)f_1(x)+\alpha_2(x)f_2(x)}{nh}+o\left(\frac{1}{nh}\right). \tag{22}$$

Suppose the mode of f_1 is to the left of the mode of f_2 in the x-axis, and that f_1 and f_2 weakly overlap in the middle range [Fig. 3(b)]. Since in the left tail of f_1, $f_2\approx 0$, and $f_2'\approx 0$ [region I in Fig. 3(b)], the h_{opt} in region I is given by $h_{1,\mathrm{opt}}^{\mathrm{MSE}}(x)\approx 2^{1/3}n^{-1/3}(f_1(x)/\alpha_1(x)f_1'(x)^2)^{1/3}$ [see Eq. (13)]. In the

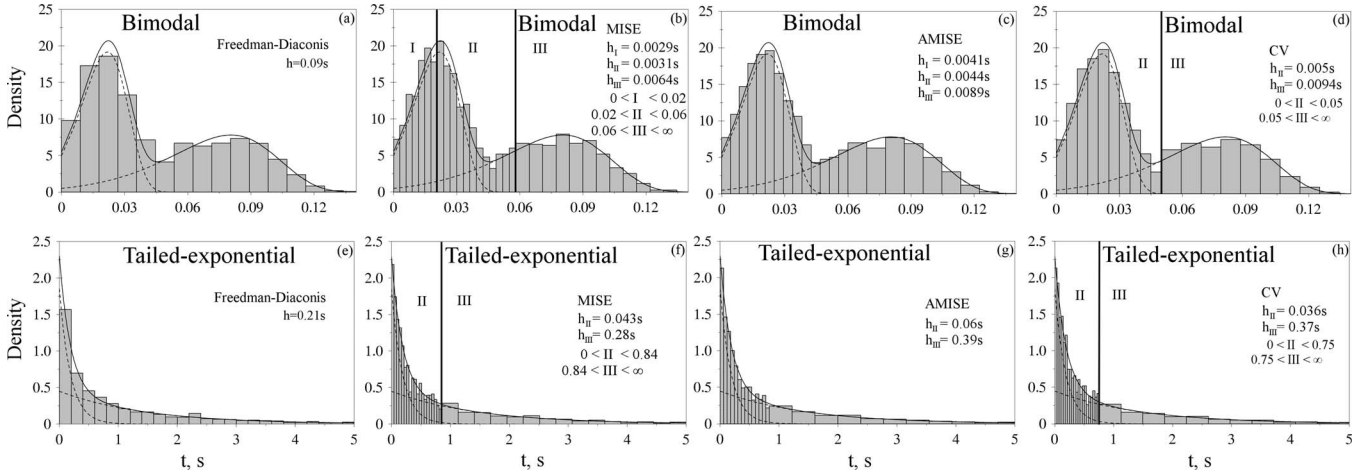

FIG. 3. The distributions of lifetimes of the ligand-receptor complex $L\cdot R$ [scheme (20)] for the force-ramp [(a)–(d)] and force-clamp [(e)–(h)] protocol, sampled from the bimodal density [Eq. (21)]. The histograms of 1600 data points (bars), obtained by using MC simulations (Sec. I) with time step $\Delta t = 10^{-6}$ s, are compared with the exact pdfs (solid curves) and with $\alpha_1 f_1(t)$ and $\alpha_2 f_2(t)$ (dashed curves). The histograms are constructed by using the Freedman–Diaconis rule [Eq. (3), (a) and (e)], adaptive MISE criterion [Eq. (24), (b) and (f)], adaptive AMISE criterion [Eq. (25), (c) and (g)], and CV-based approach [Eq. (18), (d) and (h)]. The regions I, II, and III (II and III) for the bimodal (tailed-exponential) density, used in the MISE and AMISE criteria, are indicated in panel (b) [(f)], and the regions II and III, used in the CV-based approach, are shown in panels (d) and (h). The numerical values for h_{FD}, $h_{\mathrm{opt}}^{\mathrm{MISE}}$, and $h_{\mathrm{opt}}^{\mathrm{AMISE}}$ for regions I–III and the limits for each region are indicated in the graphs.

right tail of f_2, $f_1 \approx 0$ and $f_1' \approx 0$ [region III in Fig. 3(b)], and $h_{\mathrm{III,opt}}^{\mathrm{MSE}}(x) \approx 2^{1/3} n^{-1/3} (f_2(x)/\alpha_2(x) f_2'(x)^2)^{1/3}$. In the middle range [region II in Fig. 3(b)], roughly around the minimum between the two modes where f_1 and f_2 overlap, both densities contribute, and $h_{\mathrm{II,opt}}^{\mathrm{MSE}}(x) = (2/n)^{-1/3} (\alpha_1(x) f_1(x) + \alpha_2(x) f_2(x))^{1/3} (\alpha_1(x) f_1'(x) + \alpha_2(x) f_2'(x))^{-2/3}$.

Consider now the MISE and AMISE criteria. First, by using Eqs. (14) and (21), we obtain the MISE criterion for the bimodal density,

$$\mathrm{MISE}[\hat{f}_h(x)] \approx \int_0^{x_1^{\max}} dx \mathrm{MSE}[\alpha_1 \hat{f}_1(x), h_I] + \int_{x_1^{\max}}^{x_2^{\max}} dx \mathrm{MSE}[\alpha_1 \hat{f}_1 + \alpha_2 \hat{f}_2(x), h_{\mathrm{II}}] + \int_{x_2^{\max}}^{+\infty} dx \mathrm{MSE}[\alpha_2 \hat{f}_2(x), h_{\mathrm{III}}], \tag{23}$$

where $x_1^{\max}$ and $x_2^{\max}$ are the abscissae of the maxima of f_1 and f_2. Equation (23) shows that h_{opt} for the histogram must be estimated separately for each density region I–III, so that

$$h_{\mathrm{opt,I}}^{\mathrm{MISE}} = \left(\frac{2}{n}\right)^{1/3} \left(\frac{\int_0^{x_1^{\max}} dx \alpha_1(x) f_1(x)}{\int_0^{x_1^{\max}} dx (h a_1(x) f_1'(x))^2} \right)^{1/3},$$

$$h_{\mathrm{opt,II}}^{\mathrm{MISE}} = \left(\frac{2}{n}\right)^{1/3} \left(\frac{\int_{x_1^{\max}}^{x_2^{\max}} dx (\alpha_1(x) f_1(x) + \alpha_2(x) f_2(x))}{\int_{x_1^{\max}}^{x_2^{\max}} dx (\alpha_1(x) f_1'(x) + \alpha_2(x) f_2'(x))^2} \right)^{1/3},$$

$$h_{\mathrm{opt,III}}^{\mathrm{MISE}} = \left(\frac{2}{n}\right)^{1/3} \left(\frac{\int_{x_2^{\max}}^{\infty} dx \alpha_2(x) f_2(x)}{\int_{x_2^{\max}}^{\infty} dx (\alpha_2(x) f_2'(x))^2} \right)^{1/3}. \tag{24}$$

Finally, by using Eq. (24) and Eqs. (15) and (17), we obtain the corresponding expressions for $h_{\mathrm{opt}}^{\mathrm{AMISE}}$ for the histogram:

$$h_{\mathrm{opt},r}^{\mathrm{AMISE}} = 3^{1/3} h_{\mathrm{opt},r}^{\mathrm{MISE}}, \quad r = \mathrm{I}, \mathrm{II}, \quad \text{and} \quad \mathrm{III}. \tag{25}$$

Equations (24) and (25) show that due to the bimodal nature of the density, both $h_{\mathrm{opt}}^{\mathrm{MISE}}(x)$ and $h_{\mathrm{opt}}^{\mathrm{AMISE}}(x)$ vary along the data range $x \in [0;\infty)$. Indeed, for shorter $0 < x < x_1^{\max}$ (region I) and longer $x_1^{\max} < x < \infty$ (region III) the kinetics of forced unfolding [scheme (19)] and forced unbinding [scheme (20)] is controlled by the pathways I and II, respectively, and h_{opt} is determined by the statistical properties of $f_1(x)$ and $f_2(x)$ alone. In the middle range $x_1^{\max} < x < x_2^{\max}$ (region II), pathways I and II compete, and hence, h_{opt} is determined by $f_1(x)$ and $f_2(x)$.

Let us consider the lifetimes for a ligand-receptor complex $L\cdot R$, which undergoes conformational transitions between the two bound states, $(L\cdot R)_1 \rightleftharpoons (L\cdot R)_2$, with rates r_{12} and r_{21}. The complex is subjected to pulling force, which results in its dissociation from state $(L\cdot R)_1$ or state $(L\cdot R)_2$ with the off-rate k_1 and k_2, respectively [scheme (20)]. We employ the constant force protocol, $g = g_0$ (force clamp), and the time-dependent force protocol, $g(t) = r_g t$ (force ramp), where r_g is the loading rate. We describe the force dependence of the off-rates k_1, k_2, and transition rates r_{12}, and r_{21} by using the Bell model, i.e., $k_{1,2} = k_{1,2}^0 e^{y_{1,2} g/k_B T}$ and $r_{ij} = r_{ij}^0 e^{x_{1,2} g/k_B T}$ $(i,j = 1,2)$, where $k_{1,2}^0$ and r_{ij}^0 are the attempt frequencies, and $x_{1,2}$ and $y_{1,2}$ are the distances from the minima of the states $(L\cdot R)_1$ and $(L\cdot R)_2$ to the transition states for conformational fluctuations and unbinding, respectively.[9,41] For the kinetic model (20), the distribution

of bond lifetimes $f(t)$ for both force protocols are described by the bimodal density in Eq. (21) and Eqs. (A1) and (A2) (Appendix A), which we refer to as "tailed-exponential" density (force clamp) and "bimodal" density (force ramp). To generate $n=1600$ bond lifetimes for each force regime, we carried out MC simulations of the bimodal density by using the following model parameters: $k_1=5.0\ \mathrm{s}^{-1}$, $k_2=0.5\ \mathrm{s}^{-1}$, and $r_{12}=r_{12}=0.3\ \mathrm{s}^{-1}$ (force clamp), and $k_1^0=1.0\ \mathrm{s}^{-1}$, $k_2^0=10.0\ \mathrm{s}^{-1}$, $r_{12}^0=r_{21}^0=0.01\ \mathrm{s}^{-1}$, $y_1=0.5$ nm, $y_2=1.0$ nm, and $x_1=x_2=0.02$ nm (force ramp). The loading rate was set to $r_g=250$ pN/s ($k_BT=4.14$ pN nm).

We used the adaptive MISE [Eq. (24)] and AMISE criteria [Eq. (25)] to calculate h_{opt}. Because the use of Freedman–Diaconis rule for h_{opt} resulted in better approximations to the true pdf curves for $f_E(x)$ and $f_G(x)$ (Fig. 1), we compared the performance of adaptive MISE and AMISE criteria with the histograms constructed based on the Freedman–Diaconis rule [Eq. (3)]. The histograms are presented in Fig. 3. We partitioned the range of integration in Eq. (24) into regions II and III (I, II, and III) for the tailed-exponential (bimodal) density, shown in Fig. 3(f) [Fig. 3(b)]. The histograms constructed by using the adaptive MISE and AMISE criteria resolve better the pdf curves compared to the histogram constructed by using the Freedman–Diaconis rule [Figs. 3(a) and 3(e)], especially in the range of shorter lifetimes where the tailed-exponential density shows a sharp decrease, and in the crossover region. Deviations from the pdf curves are smaller for the histograms with adaptive $h_{\mathrm{opt}}^{\mathrm{MISE}}$ and $h_{\mathrm{opt}}^{\mathrm{AMISE}}$, which permits to resolve more clearly the modes of the bimodal density and abscissae for both $f_1(t)$ and $f_2(t)$, as well as the position of the minimum [Fig. 3(c)].

These results demonstrate that for more complex probability densities, such as bimodal densities, and, in general, long-tailed densities and multimodal densities with multiple peaks and valleys, the histograms constructed by using the adaptive MISE and AMISE criteria with varying bin size perform better at describing the true pdf curves compared to the histograms constructed by using standard rules for bin size selection. The optimal bin size varies from one density region to another as it adjusts to local changes in probability mass. Adaptive criteria use more (fewer) bins for describing the pdf curves where the probability density changes faster (slower), thus offering more flexibility at resolving the density in the regions of maxima and minima (Fig. 3). For example, for $n=1600$ data points, the Freedman–Diaconis rule prescribes to use 15 (41) bins for the bimodal (tailed-exponential) density, whereas the adaptive MISE criterion partitions the data into 31 (49) bins and the AMISE criterion uses 23 (35) bins. Of these, 19 bins (MISE) and 14 bins (AMISE) are used to describe the narrower $f_1(t)$ portion of the bimodal pdf, and 20 bins (MISE) and 14 bins (AMISE) are used to resolve the sharply decaying part of $f_1(t)$ for the tailed-exponential pdf.

The use of adaptive MISE and AMISE criteria requires prior knowledge about the roughness function, $R(f')$ [Eqs. (15), (17), (24), and (25)], that quantifies the amplitude of variation of the true pdf $f(t)$. To overcome this limitation, the data-driven adaptive CV criterion can be employed. We used the adaptive CV criterion to compute the varying optimal bin size [Eq. (18)] for tailed-exponential and bimodal densities. Because the application of MISE and AMISE criteria showed that $h_{\mathrm{opt,I}}\approx h_{\mathrm{opt,II}}$, we partitioned the lifetime data into two groups [regions II and III in Figs. 3(d) and 3(h)]. The grid search was used to find the minima of functions $\mathrm{CV}[h^{\mathrm{II}}]$ and $\mathrm{CV}[h^{\mathrm{III}}]$, where $h^{\mathrm{II,III}}=(x_{\max}^{\mathrm{II,III}}-x_{\min}^{\mathrm{II,III}})/m$ and $m=m_1=2$, and an exhaustive search ($k=1$) was employed to compute $h_{\mathrm{opt},i}^{\mathrm{II,III}}$ and $h_{\mathrm{opt}}^{\max}$. We see that the adaptive CV-based histograms are closer to the true pdf curves compared to the histograms, obtained by using the Freedman–Diaconis rule for h_{opt}. As in the case of the MISE and AMISE, the adaptive CV criterion for h_{opt} requires more bins, i.e., 10 (21) bins, for describing the narrower (rapidly decaying) $f_1(t)$-part, and fewer bins, namely, 9 (12) bins, for resolving the wider (slowly decaying) $f_2(t)$-portion of the bimodal (tailed-exponential) pdf [Figs. 3(d) and 3(h)]. Hence, the use of adaptive CV criterion allows to resolve better the pdfs with long tails compared to the Freedman–Diaconis approach.

IV. KERNEL DENSITY ESTIMATION

A. General methodology

Although the histograms constructed based on adaptive MISE, AMISE, and CV optimal bin size perform better at describing the true pdf curves, compared to the histograms with fixed bin size, they remain discrete approximations to the continuous pdfs. In addition, the CV-based histogram construction involves computationally intensive algorithms.

One of the most widely used techniques of nonparametric density estimation is kernel density estimation.[55] The idea behind this method is that instead of assigning equal weight of $1/n$ to every point, as in the construction of a histogram density estimator, this weight is smoothly redistributed in the vicinity of each point. The kernel density estimate is defined by

$$\hat{f}_K(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right), \tag{26}$$

where the kernel $K(x)$ (weight function) is a normalized ($\int dxK(x)=1$) and symmetric ($\int dxK(x)=0$) function with finite second moment $\sigma_K^2=\int dxx^2K(x)$. A widely used kernel function is the Gaussian kernel $K(x)=\exp(-x^2/2)/\sqrt{2\pi}$. As in the case of histograms, the bandwidth h is the most important characteristic of a kernel density estimate. The MSE of a kernel density estimate $\hat{f}_K(x)$ is given by [Eq. (5)],[57]

$$\begin{aligned}\mathrm{MSE}[\hat{f}_K(x)]&=\mathrm{var}[\hat{f}_K(x)]+\mathrm{bias}[\hat{f}_K(x)]^2\\&=\frac{f(x)}{nh}R(K(x))+\frac{1}{4}h^4\sigma_K^4(f''(x))^2+O\left(\frac{1}{n}\right)\\&\quad+O(h^6),\end{aligned} \tag{27}$$

and the MISE of $\hat{f}_K(x)$ is given by [Eq. (6)]

$$\begin{aligned}\mathrm{MISE}[\hat{f}_K(x)]&=\frac{1}{nh}R(K(x))+\frac{1}{4}h^4\sigma_K^4R(f''(x))+O\left(\frac{1}{n}\right)\\&\quad+O(h^6).\end{aligned} \tag{28}$$

The $\mathrm{AMISE}[\hat{f}_K]$ is simply the sum of the first two terms in

Eq. (28), i.e., $\mathrm{AMISE}[\hat{f}_K(x)]=R(K(x))/nh+h^4\sigma_K^4R(f''(x))/4$. In both MISE and AMISE, the squared bias, given by the h^4-term in Eqs. (27) and (28), is proportional to the bandwidth and large bandwidth values may result in oversmoothed pdf estimates that obscure the fine structure of the data. The variance, given by the first term in Eqs. (27) and (28), is inversely proportional to the bandwidth and hence, small bandwidth values may result in undersmoothed wiggly curves with artificial modes.

By following the same line of argument used in deriving the AMISE based optimal bin size for the histogram [Eq. (17)], one can show that the bandwidth that minimizes the AMISE for a kernel density estimate is given by

$$h_{\mathrm{opt},K}^{\mathrm{AMISE}}=\left(\frac{R(K(x))}{\sigma_K^4R(f''(x))}\right)^{1/5}n^{-1/5}. \tag{29}$$

A comparison of Eq. (17) with Eq. (29) shows that $h_{\mathrm{opt},K}^{\mathrm{AMISE}}$ decreases at the faster rate of $n^{-1/5}$ as opposed to the $n^{-1/3}$-rate for $h_{\mathrm{opt},H}^{\mathrm{AMISE}}$ for the histogram. Since the rate of convergence of the MISE and AMISE based kernel density estimates to the true pdf are asymptotically equivalent in the large n limit, we can substitute the expression for $h_{\mathrm{opt},K}^{\mathrm{AMISE}}$ [Eq. (29)] into Eq. (28) in order to estimate the optimal rate of convergence of the MISE for a kernel density estimate. We find that the convergence rate scales as $O(n^{-4/5})$ compared to the $n^{-2/3}$ scaling law for the MISE for a histogram. Hence, kernel density estimates are not only smoother but they also converge to the true pdf at a faster rate compared to histograms.

Equation (29) also indicates that $h_{\mathrm{opt},K}^{\mathrm{AMISE}}$ depends on the underlying pdf through its second derivative, $f''(x)$, and hence, the use of $h_{\mathrm{opt},K}^{\mathrm{AMISE}}$ is limited. As in the case of histograms, the optimal bandwidth can be chosen either by using the CV approach, or by using the so-called plug-in or reference distribution methods. The original plug-in approach is based on the assumption that there is some knowledge about the underlying pdf. If the true density is symmetric, then the normal reference rule sets the Gaussian with standard deviation σ as the target density, and the optimal bandwidth is given by

$$h_{\mathrm{opt},K,G}^{\text{plug-in}}=\left(\frac{8\sqrt{\pi}}{3}\right)^{1/5}\left(\frac{R(K(x))}{\sigma_K^4}\right)^{1/5}\sigma n^{-1/5}. \tag{30}$$

When $K(x)$ is the standard normal kernel, $h_{\mathrm{opt},N,K}^{\text{plug-in}}=1.06n^{-1/5}\hat{\sigma}$, where $\hat{\sigma}=\min\{\sigma_x,\mathrm{IQR}/1.34\}$ and σ_x is the sample standard deviation. We will refer to this approach as the plug-in method.

Among the second generation plug-in methods, the most widely used is the Sheather–Jones (SJ) "solve-the-equation plug-in approach."[61] The main idea of this approach is to plug-in an estimate for $R(f''(x))$ into Eq. (29) for $h_{\mathrm{opt},K}$ and then solve the equation

$$h_{\mathrm{opt},K}^{\mathrm{SJ}}=\left(\frac{R(K(x))}{n\sigma_K^4R(\hat{f}''_{\hat{\beta}(h)}(x))}\right)^{1/5} \tag{31}$$

for $h_{\mathrm{opt},K}^{\mathrm{SJ}}$. This involves the calculation of the bandwidth $\hat{\beta}(h)$ for the estimation of the roughness function $R(f''(x))$, which is done by finding an analog of h_{opt} in Eq. (29) for estimating $R(f'')$.[61] The algorithm for obtaining the SJ based optimal bandwidth $h_{\mathrm{opt},K}^{\mathrm{SJ}}$ is presented in Appendix B.

The most studied among automatic databased bandwidth selectors is the least-squares CV.[59] The CV function for a kernel density estimate is given by (Appendix B)

$$\mathrm{CV}[h]=\frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}K\star K\left(\frac{x_j-x_i}{h}\right)-\frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(\frac{x_j-x_i}{h}\right), \tag{32}$$

where $K\star L(x)=\int duK(u)L(x-u)$, and the optimal CV bandwidth, $h_{\mathrm{opt},K}^{\mathrm{CV}}$, is the minimizer of Eq. (32). Because $\mathrm{CV}[h]$ often has several minima,[62] a grid search (Sec. III) is preferred over standard optimization techniques such as the Newton–Raphson method.

B. Application of kernel density estimation methods

1. Tailed-exponential and bimodal data

We assessed the performance of the plug-in, CV, and SJ kernel density estimates at approximating the tailed-exponential and bimodal pdf of the bond lifetimes [Eq. (21)], generated by using MC simulations (Sec. I) of these densities with the same parameter values as in the case of the adaptive MISE, AMISE, and CV criteria (Sec. III B, Fig. 3). The obtained kernel density curves are compared in Fig. 4 for varying sample size n with the true tailed-exponential and bimodal pdfs, and with histogram estimates based on the Freedman–Diaconis rule for h_{opt}. The numerical values of h_{opt} for all three kernel density estimates are summarized in Table I. We see that all three density estimates deviate somewhat from the true tailed-exponential pdf at shorter lifetimes [Figs. 4(a)–4(d)], especially for $n=400$ [Fig. 4(a)]. Specifically, the plug-in, CV, and SJ based curves underestimate the true pdf curves in the range of the faster decaying $f_1(t)$-part of the tailed-exponential density. However, the improvement grows with n, and the slower $f_2(t)$-portion of the density is resolved fairly well. The CV and SJ based estimates show closer agreement with true tailed-exponential and bimodal pdfs [Figs. 4(e)–4(h)] compared to the plug-in estimate, and for $n=3200$ the agreement is perfect [Figs. 4(d) and 4(h)]. In the case of the CV and SJ based estimates of the bimodal density, the location of modes, the minimum separating $f_1(t)$ from $f_2(t)$, and the widths of $f_1(t)$ and $f_2(t)$ are well resolved; yet the heights of the modes are lower compared to those for the true bimodal pdf [Fig. 4(g)].

To provide a numerical assessment of the relative performance of kernel density estimates and the histogram based estimate, we computed the MSE

$$\mathrm{MSE}=\sum_{i=1}^{n}(\hat{f}_{K,h}(x_i)-f(x_i))^2/n, \tag{33}$$

which measures the L_2-distance between the kernel (histogram) estimate $\hat{f}_K(x)$ $(\hat{f}_h(x))$ and the true density $f(x)$. The obtained MSE values are given in Table II. For the tailed-exponential pdf, the CV and SJ based kernel density esti-

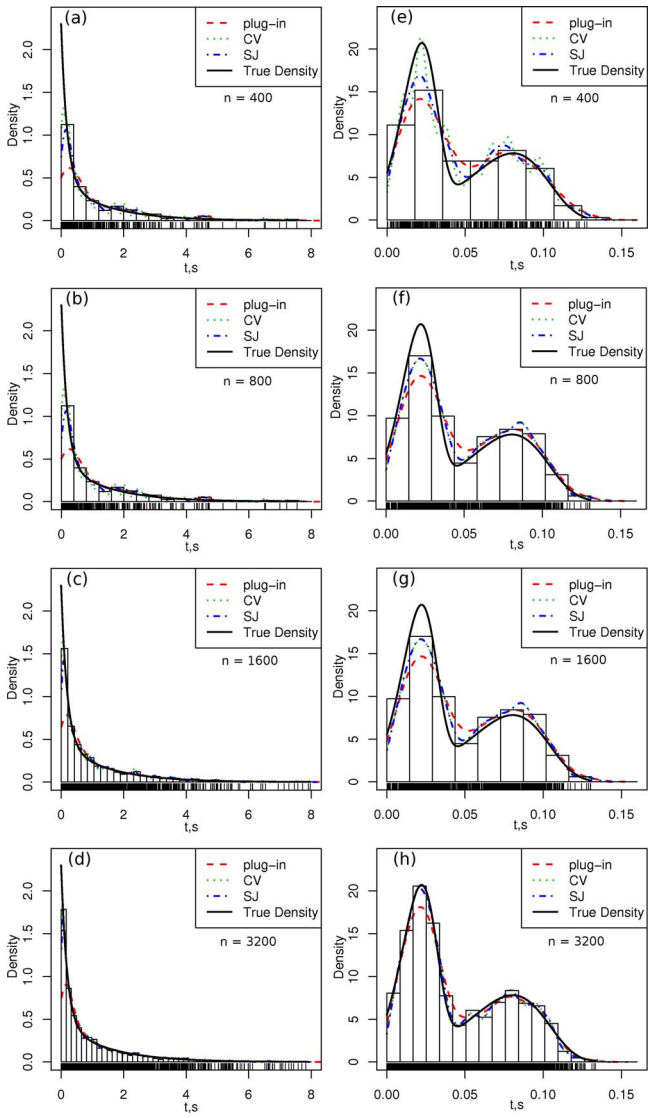

FIG. 4. (Color online) The true tailed-exponential pdf of bond lifetimes $f(t)$ [panels (a)–(d)] and the bimodal pdf of rupture forces $f(g)$ [panels (e)–(h)] for the ligand-receptor complex [scheme (20), solid curves], computed by using Eq. (21) (Appendix A), are compared for varying sample size n with the kernel density estimates of these densities (dashed curves), obtained by using the plug-in, CV, and SJ methods for the optimal bandwidth selection, and with the histograms (bars), constructed by using the Freedman–Diaconis rule for h_{opt} [Eq. (3)].

mates show smaller estimation errors for all values of n, compared to the plug-in based estimate. The CV-based estimate shows similar agreement with the true pdf as the histogram constructed by using the Freedman–Diaconis rule for h_{opt} (Table II). This implies that the CV method is more useful for describing the crossover region, from the faster decay at shorter lifetimes to the slower decay at longer lifetimes, which characterizes the tailed-exponential density [Figs. 4(a)–4(d)]. For the bimodal density, the CV and SJ based kernel density estimates show better agreement with the true pdf compared to the histogram based estimate, implying that the SJ based bandwidth is preferred for resolving multiple peaks and valleys of the multimodal densities.

2. Forced unraveling of the Rouse chain

We performed Langevin simulations of the forced unraveling of the Rouse chain[63] by using the force-clamp and force-ramp protocols (Appendix C). Very close approximations to the true pdfs of unfolding times, $f(t)$, and unfolding forces, $f(g)$, were obtained by collecting large data set ($n=5000$) of unfolding measurements for each force protocol. The data were used to construct the cumulative distribution functions (cdfs), $F(t)$ and $F(g)$, which were then used to estimate numerically the pdfs, $f(x)=dF(x)/dx$ ($x=t$ or g). We set the number of monomers in the chain $N=100$, the covalent bond distance $a=0.4$ nm, which sets the total length of the chain $L=Na=40$ nm, and the diffusion constant $D=250$ nm$^2/\mu$s. We assumed that the end-to-end distance B at which the chain is in the unfolded state is $B=0.9$ L, and used constant force $g_0=30$ pN and time-dependent force $g(t)=r_g t$ ramped up with the loading rate $r_g=3$ pN/μs. A comparison of the plug-in, CV, and SJ kernel density estimates with the histograms constructed by using the Freedman–Diaconis rule for h_{opt}, and with the numerically determined pdfs $f(t)$ and $f(g)$ is presented in Fig. 5. The h_{opt} values for the plug-in, CV, and SJ kernel density estimates are given in Table I. The MSE values for the kernel density estimates and for the histograms (Table III), show that all three kernel density estimates perform better at describing the skewed distributions of unfolding times [Figs. 5(a)–5(d)] and unfolding forces [Figs. 5(e)–5(h)] compared to histograms for all values of n. Indeed, here histograms are largely inaccurate rough estimates of $f(t)$ and $f(g)$. Even a visual inspection of the graphs shows that the agreement between kernel density estimates and the pdf curves is very good for $n=800$ [Figs. 5(c) and 5(g)], and that the agreement is almost quantitative for $n=1200$ data sample [Figs. 5(d) and 5(h)]. All kernel density estimates capture the locations of maxima, describe well the width of the densities, and resolve the long right (left) tail of $f(t)$ ($f(g)$). The density tails are due to outlying observations that can be analyzed by using statistics of extremes.[64]

3. Forced dissociation of the protein-protein complex

We also performed Langevin simulations of the Brownian particle in the harmonic potential subjected to pulling force. This model was used in Ref. 52 to describe the forced dissociation of the protein-protein complex $P_1 \cdot P_2$ formed by proteins P_1 and P_2, $P_1 \cdot P_2 \rightarrow P_1 + P_2$, in which the particle position describes the extension of the noncovalent bond distance. The bond lifetime and the rupture force data samples of varying size n were generated by employing force-clamp and force-ramp protocols, respectively (Appendix C). Very close approximations to the pdfs of bond lifetimes $f(t)$ and unbinding forces $f(g)$ were obtained by collecting large data set ($n=10\,000$) for each force protocol and constructing the cdfs $F(t)$ and $F(g)$, which were used to estimate the pdfs. We set the molecular spring constant, which quantifies the curvature of the equilibrium free energy landscape for unbinding, $\kappa_m=10$ pN/nm, the cantilever spring constant $\kappa=1$ pN/nm, and the diffusion constant $D=1.0$ nm$^2/\mu$s. We assumed that the critical extension of the noncovalent bond, y^*, at which the bond ruptures is $y^*=1.0$ nm. We applied constant force $g_0=3$ pN, and time-dependent force $g(t)$ increasing linearly with the pulling speed $\nu_0=1.0$ nm/μs.

TABLE I. The optimal bandwidth values for the plug-in, CV, and SJ kernel density estimates of the tailed-exponential pdf of bond lifetimes, the bimodal pdf of rupture forces for the ligand-receptor complex (Appendix A), the pdf of unfolding times and unfolding forces for the Rouse chain (Appendix C), and the pdf of bond lifetimes and rupture forces for the protein-protein complex (Appendix C) as a function of sample size n.

Density	Tailed-exponential (s)				Bimodal (s)			
n	400	800	1600	3200	400	800	1600	3200
Plug-in	0.345	0.257	0.227	0.188	0.010	0.009	0.008	0.007
CV	0.048	0.036	0.034	0.029	0.003	0.005	0.004	0.003
SJ	0.121	0.082	0.066	0.048	0.006	0.005	0.004	0.003
Density	Unfolding times (μs)				Unfolding forces (pN)			
n	200	400	800	1200	200	400	800	1200
Plug-in	0.128	0.115	0.096	0.082	0.522	0.480	0.424	0.383
CV	0.032	0.092	0.089	0.073	0.537	0.515	0.455	0.379
SJ	0.122	0.096	0.081	0.72	0.535	0.478	0.416	0.377
Density	Bond lifetimes (μs)				Rupture forces (pN)			
n	200	400	800	1600	200	400	800	1600
Plug-in	0.463	0.421	0.337	0.310	0.520	0.469	0.378	0.338
CV	0.194	0.081	0.083	0.093	0.559	0.504	0.275	0.352
SJ	0.278	0.227	0.154	0.130	0.576	0.513	0.354	0.348

The plug-in, CV, and SJ based kernel density estimates are compared to the histogram based estimate, constructed by using the Freedman–Diaconis rule for h_{opt}, and with the pdf curves of bond lifetimes $f(t)$ and rupture forces $f(g)$ in Fig. 6. The optimal bandwidth values are summarized in Table I. The pdf curves of bond lifetimes and rupture forces for the protein-protein complex appear more noisy compared to the pdf curves of unfolding times and unfolding forces for the Rouse chain (Fig. 5). This is due to a weaker cantilever spring constant ($\kappa=1$ pN/nm) compared to a stiff molecular spring constant ($\kappa_m=10$ pN/nm), and is due to cusplike potential for unbinding used in pulling simulations (Appendix C). The $\kappa=1$ pN/nm value was chosen to assess the performance of kernel density estimators at describing noisy data. Though the curves for all three kernel density estimates disagree with the pdf $f(t)$ at shorter lifetimes [Figs. 6(a)–6(d)], due to poor sampling of the probability density (few data points), they describe well the decaying portion and the width of $f(t)$, which carries information about the unbinding rate. All three density kernel estimators resolve better the pdf of rupture forces $f(g)$ [Figs. 6(e)–6(h)] compared to the lifetime pdf, $f(t)$. This is because in the force-ramp protocol the applied pulling force increases gradually resulting in slower variation of $f(g)$. The MSE values (Table IV) indicate that the CV and SJ based estimates resolve better $f(t)$ and $f(g)$ compared to the histogram, and capture the location and height of the peak of $f(t)$ and $f(g)$.

V. DISCUSSION

We presented a comprehensive analysis of optimal bandwidth selection for nonparametric estimates of the pdfs that describe the forced unfolding data for proteins and the forced unbinding data for protein-protein complexes. The histogram is the classical nonparametric density estimator dating back to the mortality studies of Graunt in 1662.[45] By construction, the histogram is very sensitive to the choice of the bin size, and also depends on the "histogram origin," i.e., location of

TABLE II. The MSE values [Eq. (33)] for the histogram, constructed using the Freedman–Diaconis rule for h_{opt} [Eq. (3)], and for the plug-in, CV, and SJ kernel density estimates (Sec. IV) of the tailed-exponential pdf of bond lifetimes and the bimodal pdf of rupture forces for the ligand-receptor complex (Appendix A) as a function of sample size n.

Density	Tailed-exponential (s^2)				Bimodal (s^2)			
n	400	800	1600	3200	400	800	1600	3200
Plug-in	0.380	0.307	0.324	0.284	10.694	8.799	3.877	1.784
CV	0.108	0.083	0.049	0.037	4.290	4.278	1.448	0.307
SJ	0.202	0.128	0.101	0.063	4.468	4.209	1.428	0.295
FD histogram	0.135	0.082	0.053	0.035	9.752	6.750	3.941	2.336

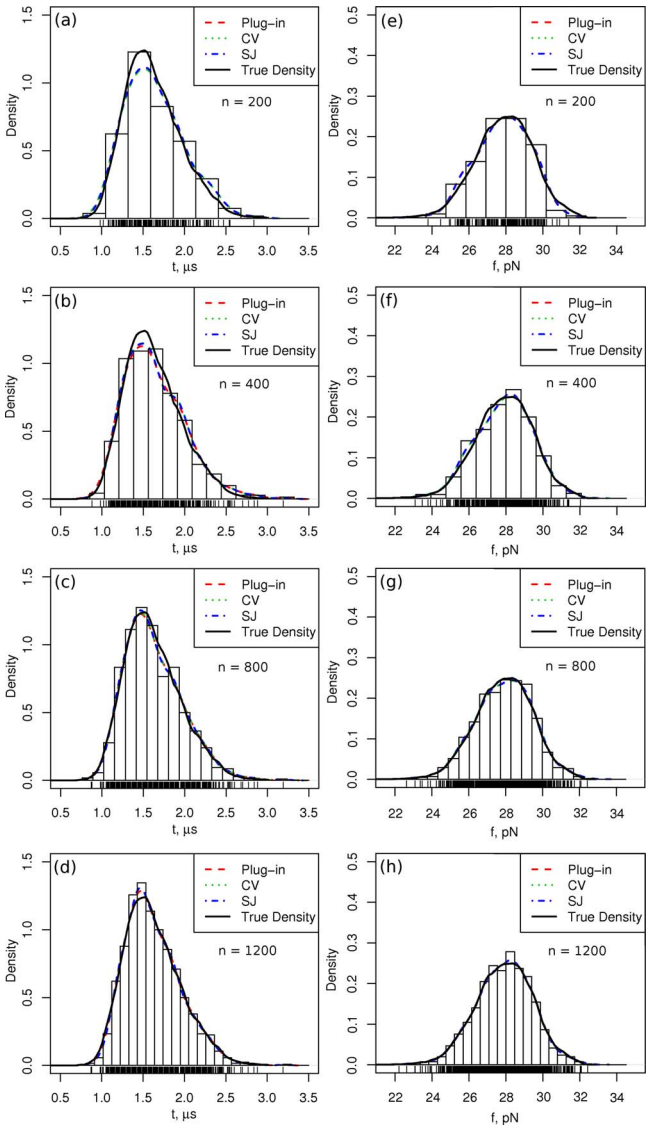

FIG. 5. (Color online) The pdf curves of the unfolding times $f(t)$ [(a)–(d)] and the unfolding forces $f(g)$ [(e)–(h)] for the Rouse chain (Appendix C) are compared for varying sample size n with the kernel density estimates of these densities (dashed curves), obtained by using the plug-in, CV, and SJ methods, and with the histograms (bars), constructed by using the Freedman–Diaconis rule for h_{opt} [Eq. (3)].

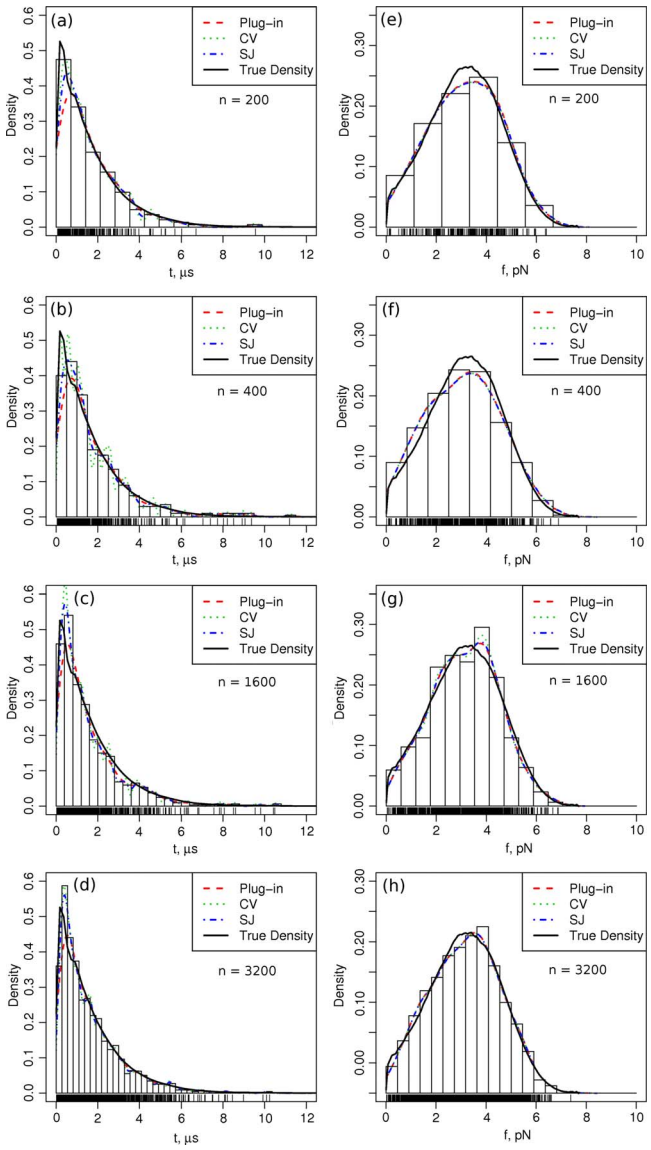

FIG. 6. (Color online) The pdf curves of the bond lifetimes $f(t)$ [(a)–(d)] and the rupture forces $f(g)$ [(e)–(h)] for the protein-protein complex (Appendix C), are compared for varying sample size n with the kernel density estimates of these densities (dashed curves), obtained by using the plug-in, CV, and SJ methods, and with the histograms (bars), constructed by using the Freedman–Diaconis rule for h_{opt} [Eq. (3)].

the first bin. Most authors of Statistics textbooks advise that 5–20 bins are usually adequate for describing real data sets, and that the origin of the first bin should be chosen so that the data do not fall on the bin boundaries.[46] The question of the bin origin is of lower order effect on the histogram compared to the bin size. One may simply select several bin origins, and, by using the same bin size, average the resulting histograms, which results in the so-called averaged shifted histogram.[56,65]

TABLE III. The MSE values [Eq. (33)] for the histogram, constructed using the Freedman–Diaconis rule for h_{opt} [Eq. (3)], and for the plug-in, CV, and SJ kernel density estimates (Sec. IV) of the pdf of unfolding times and unfolding forces for the Rouse chain (Appendix C) as a function of sample size n.

Density	Unfolding times (μs^2)				Unfolding forces (pN^2)			
n	200	400	800	1200	200	400	800	1200
Plug-in	0.002	0.002	0.001	0.0002	2.53×10^{-5}	2.13×10^{-5}	5.37×10^{-6}	2.54×10^{-6}
CV	0.003	0.002	0.001	0.0004	2.45×10^{-5}	2.07×10^{-5}	6.10×10^{-6}	2.63×10^{-6}
SJ	0.002	0.002	0.001	0.0004	2.46×10^{-5}	2.14×10^{-5}	5.27×10^{-6}	2.66×10^{-6}
FD histogram	0.011	0.006	0.006	0.003	0.0002	0.0001	5.99×10^{-5}	5.12×10^{-5}

TABLE IV. The MSE values [Eq. (33)] for the histogram, constructed using the Freedman–Diaconis rule for h_{opt} [Eq. (3)], and for the plug-in, CV, and SJ kernel density estimates (Sec. IV) of the pdf of bond lifetimes and rupture forces for the protein-protein complex (Appendix C) as a function of sample size n.

Density	Bond lifetimes (μs^2)				Rupture forces (pN^2)			
n	200	400	800	1600	200	400	800	1600
Plug-in	0.0009	0.0009	0.0005	0.0005	0.0001	0.0003	8.22×10^{-5}	6.85×10^{-5}
CV	0.0003	0.0007	0.0006	0.0005	0.0001	0.0003	0.0002	6.69×10^{-5}
SJ	0.0005	0.0006	0.0005	0.0004	0.0001	0.0003	9.58×10^{-5}	6.73×10^{-5}
FD histogram	0.0008	0.0008	0.001	0.0007	0.0007	0.0006	0.0006	0.0002

The first attempt to derive a formal rule for choosing the bin size for the histogram is due to Sturges[47] who proposed a simple rule for classifying a series of n observations. He observed that normally distributed random variables can be appropriately divided so that the bin counts comprise a binomial series for all n which are even powers of 2. As the number of bins increases, an ideal histogram tends to a normal density with mean $(r-1)/2$ and variance $(r-1)/4$, for which the kth bin count B_k equals the binomial coefficient $\binom{B_k=r-1}{k}$, for $k=0,\ldots,r-1$ and a total of r bins of width 1. The sum of bin counts then equals the total sample size, $n=2^{r-1}$. Consequently, the optimal number of bins is $n_{opt}=1+\log_2 n$, and the optimal bin size is given by Eq. (1).[47] Sturges' rule has become a guideline for researchers, and is the standard, default value for bin size used in statistical software packages. Scott[48] proposed an alternative choice of the bin size that minimizes the MSE over the entire data range. Because Scott's rule also uses the Gaussian distribution as reference, the bin size is proportional to the sample standard deviation σ_x [Eq. (2)]. However, the forced unfolding data for proteins and unbinding data for protein-protein complexes are of "lifetime type," i.e., skewed, in which case Sturge's and Scott's rules-of-thumb are inappropriate as they use too few bins to resolve long tails of such distributions. Freedman and Diaconis'[49] rule takes into account the asymmetry of the data and sets the bin size to be proportional to the IQR [Eq. (3)]. The application of Sturges' rule, Scott's rule, and Freedman–Diaconis' rule to describing exponential and gamma densities [Eq. (4)] showed that the use of the Freedman–Diaconis rule results in better histogram based approximations to skewed, lifetimelike distributions (Fig. 1).

By employing several model-driven statistical measures of estimation accuracy, such as the global MISE and AMISE measures, we derived analytical expressions for the optimal bin size for the histograms. We also presented a numerical algorithm for the estimation of the optimal bin size using a data-driven CV estimator of accuracy. The advantage of the CV-based approach, which belongs to the class of automatic bandwidth selection methods,[66] over the MISE and AMISE based methods is that its implementation *does not require any knowledge about the underlying pdf*. Their construction, although, employs computationally intensive algorithms that are more difficult to implement compared to the MISE and AMISE based approaches. We also developed adaptive MISE, AMISE, and CV-based approaches for the *varying optimal bin size* selection for histograms. "Adaptive histograms" perform better at describing the true pdf curves of the tailed-exponential and bimodal density compared to the histograms constructed by using standard MISE, AMISE, and CV-based rules. Our results show that, compared to histograms with fixed bin size, the histograms with adaptive (varying) bin size appear wider in the tails of the tailed-exponential and bimodal density and narrower in the regions where the density is changing rapidly (Fig. 3).

To this day, the histogram remains a simple yet important statistical tool for displaying and summarizing experimental and simulated data. However, the histogram is a "discrete" approximation to a continuous pdf, sensitive to the choice of the bin size, in which bin counts are constant over the intervals of bin length. As a result, the histogram requires large data samples to capture the main features of the underlying pdf. Our results indicate that 400–800 data points are needed in order to obtain a good histogram based approximation to the unimodal distributions (Figs. 1 and 2), and that 1000–1200 data points are needed to resolve more complex distributions such as the tailed-exponential and the bimodal density (Fig. 3). Furthermore, in order to assess the validity of a model for protein unfolding or unbinding, a goodness-of-fit test is needed. The existing chi-square goodness-of-fit test is, practically, the only available such tool.[67] However, for a relatively small data set of, say, a few hundreds of data points or less, sampled from a skewed distribution, the test oftentimes fails because of empty bins (Figs. 1 and 2). As an alternative, we propose to use nonparametric kernel density estimators.[53–55] These estimators are characterized by a faster rate of convergence, and present considerable improvement over histograms. In effect, the rate of convergence of kernel density estimates to the true pdf is the fastest across all nonparametric estimators.[68] Hence, smaller data sets are needed in order to obtain accurate estimates of the underlying pdf.

In this paper, we focused on the model-driven plug-in, as well as data-driven CV and SJ methods for optimal bandwidth selection for kernel density estimates. We employed these methods to compute kernel density estimates for the tailed-exponential and bimodal data (Fig. 4), for the unfolding times and unfolding forces for a Rouse chain (Fig. 5), and for the bond lifetime and rupture force data for the protein-protein complex (Fig. 6). The results of MSE-based analysis indicate that the model-free, data-driven CV and SJ methods perform better at approximating tailed-exponential and bimodal distributions (Table II), and the forced unfolding data (Table III) and unbinding data (Table IV) compared to histograms, with the SJ based density estimates showing the best performance, which agrees with previous findings.[69]

The SJ and CV methods do not require any knowledge about the unknown pdf; however, the bandwidth, selected by the CV method, converges slowly to the optimal bandwidth and has large variance compared with the bandwidth, computed by using the SJ approach. These findings agree with results of recent studies.[70,71]

VI. CONCLUSION: HISTOGRAMS VERSUS KERNEL DENSITY ESTIMATORS

The obtained results and algorithms for optimal bandwidth selection for nonparametric density estimates, such as histograms and the plug-in, CV, and SJ method based kernel density estimates, can be used by experimentalists and theoreticians to model the pdfs that underlie the forced unfolding data for proteins (unfolding times and unfolding forces), and forced dissociation data for protein-protein complexes and aggregates (bond lifetimes and rupture forces). The MISE, AMISE, and CV-based rules of the optimal bin size selection and the closed form expressions for h_{opt} (Sec. III A) can be used to construct histograms of the unfolding and unbinding data sampled from unimodal distributions. The adaptive MISE, AMISE, and CV-based rules and the formulas for h_{opt} (Sec. III B) can be employed to construct histograms of the data described by bimodal distributions, and can be generalized to accommodate multimodal distributions as well. The plug-in, CV, and SJ method (Sec. IV A) can be used to construct kernel density estimates of the pdfs for the protein unfolding and unbinding data. The functions for computing optimal bandwidths for all three methods are built-in in the R software package.[72]

For both histograms and kernel density estimates, the choice of a particular optimal bin size or bandwidth method depends on three factors: prior knowledge about the underlying distribution, the overall complexity of the distribution, and size of the data sample. When information about the underlying pdf is available, the MISE and AMISE based expressions for the optimal bandwidth can be used to estimate h_{opt} in order to describe the data sampled from the unimodal distributions, and the adaptive MISE and AMISE based expressions for h_{opt} can be utilized to resolve the multimodal densities with long tails and multiple peaks. When the underlying density is unknown, CV approaches can be used instead. Although histograms are simpler to construct compared to kernel density estimates, they require large data samples in order to provide good estimates of unimodal and multimodal densities. On the other hand, both for unimodal and multimodal densities, kernel density estimates can be employed to analyze smaller samples of just a few hundreds of data points. Combining several density estimation methods may be needed for accurate description of the unfolding or unbinding data for proteins, and for the assessment of analytically tractable models of unfolding or unbinding.

ACKNOWLEDGMENTS

The last two authors were supported by a start-up fund from the University of Massachusetts at Lowell.

APPENDIX A: DERIVATION OF THE BIMODAL PDF FOR KINETIC MODEL (20)

1. The lifetime pdf for constant force protocol

In the constant force regime (force clamp), the kinetic rates k_1, k_2, r_{12}, and r_{21} do not vary in time. The pdf of lifetimes for the ligand-receptor complex $L\cdot R$ was obtained in Refs. 9 and 41. The lifetime pdf, $f(t)$, is given by the tailed-exponential density [Eq. (21)] with $f_1(t)$ and $f_2(t)$ given by

$$f_1(t)=e^{-K_1 t} \quad \text{and} \quad f_2(t)=e^{-K_2 t}, \tag{A1}$$

where $K_{1,2}=(k_1+k_2+r_{12}+r_{21}\pm\sqrt{D})/2$, and $\alpha_1=K_1(P_{10}(K_1-k_2)+P_{20}(K_1-k_1)-r_{12}-r_{21})/\sqrt{D}$ and $\alpha_2=K_2(P_{10}(k_2-K_2)+P_{20}(k_1-K_2)+r_{12}+r_{21})/\sqrt{D}$. The initial populations of the bound states $(L\cdot R)_1$ and $(L\cdot R)_2$, P_{10} and P_{20}, are given by $P_{10}=r_{21}/(r_{12}+r_{21})$ and $P_{20}=r_{12}/(r_{12}+r_{21})$ and $D=(k_1+k_2+r_{12}+r_{21})^2-4(k_1k_2+k_1r_{21}+k_2r_{12})$.

2. The lifetime pdf for time-dependent force protocol

Under the simplifying assumption that $r_{12},r_{21}\ll k_1,k_2$, the lifetime density is given by the bimodal pdf [Eq. (21)] with $f_1(t)$ and $f_2(t)$ given by

$$f_1(t)=q_1(t)e^{-k_1 t} \quad \text{and} \quad f_2(t)=q_2(t)e^{-k_2 t}, \tag{A2}$$

where $q_{1,2}(t)=r_f(r_{12}x_1+r_{21}x_2)/kT(r_{12}+r_{21})+k_{1,2}+(r_f/kT)\times(k_{1,2}y_{1,2}t-x_{2,1})$, and $\alpha_1(t)=r_{21}(t)/(r_{12}(t)+r_{21}(t))$ and $\alpha_2(t)=r_{12}(t)/(r_{12}(t)+r_{21}(t))$.

APPENDIX B: COMPUTATION OF h_{opt} FOR SHEATHER–JONES METHOD AND THE DERIVATION OF Eq. (32)

1. Computation of $h_{\mathrm{opt}}^{\mathrm{SJ}}$

The algorithm for the computation of $h_{\mathrm{opt}}^{\mathrm{SJ}}$ is adapted from Ref. 61, and is built-in in the R software package:[72]

Step 1. Suppose that the set of unfolding or unbinding data comprise n observations, $x_1,\ldots,x_n$. Set parameters $a=0.920\times\mathrm{IQR}n^{-1/7}$ and $b=0.912\times\mathrm{IQR}n^{-1/9}$.

Step 2. Construct functions $\hat{T}(b)=-(1/n(n-1)b^{1/7})\Sigma_{i,j=1}^{n}\phi((x_i-x_j)/b)$ and $\hat{R}(a)=(1/n(n-1)a^5)\Sigma_{i,j=1}^{n}\phi((x_i-x_j)/a)$, where $\phi(t)=\exp(t^2/2)/\sqrt{2\pi}$.

Step 3. Set $\hat{\beta}(h)=1.357(\hat{R}(a)/\hat{T}(b))^{1/7}h^{5/7}$, and use the Newton–Raphson optimization algorithm to solve the equation $(R(K(x))/\sigma_K^4\hat{R}(\hat{\beta}(h)))^{1/5}n^{-1/5}-h=0$ for h_{opt}.

2. Derivation of Eq. (32)

We integrate the squared error loss function (Sec. II), for the kernel density estimate, $\hat{f}_K(x)$, $L(f(x),\hat{f}_K(x))$. The first term in the integral of $L(f(x),\hat{f}_K(x))$ is given by

$$\int dx(\hat{f}_K(x))^2 = \int dx\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right)^2$$
$$= \frac{1}{n^2h^2}\sum_i \sum_j \int dx K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)$$
$$= \frac{1}{n^2h}\sum_i \sum_j K(s)K\left(\frac{x_j-x_i}{h}-s\right),$$

whereas the second term is given by

$$\int dx\hat{f}_K(x)f(x) = E(\hat{f}_K(x)) \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{K,-i}(x_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(n-1)}\sum_{i\neq j}K\left(\frac{x_i-x_j}{h}\right.$$
$$= \frac{1}{hn(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}K\left(\frac{x_i-x_j}{h}\right) \quad .$$

By using these expressions we obtain Eq. (32) of the main text.

APPENDIX C: FORCED UNRAVELING OF THE ROUSE CHAIN, AND FORCED DISSOCIATION OF THE PROTEIN-PROTEIN COMPLEX

1. Rouse chain

Consider a Rouse chain of N connected beads with the bond distance a.[63] The first monomer is fixed, and a pulling force $\mathbf{g}=g\mathbf{y}$ is applied to the Nth monomer in the direction parallel to the end-to-end vector $\mathbf{y}$. The Hamiltonian is given by $H=3k_BT/2a^2\Sigma_{n=2}^{N}(\mathbf{R}_n-\mathbf{R}_{n-1})^2+\mathbf{g}(\mathbf{R}_N-\mathbf{R}_1)$, where $\mathbf{R}_j=\{R_j^x,R_j^y,R_j^z\}$ is the jth monomer position. To obtain the kinetics of forced unraveling of the chain, we integrate numerically Langevin equations for each monomer position in the overdamped limit, $\xi(d/dt)\mathbf{R}_j=-\nabla_{\mathbf{R}_j}H(\{\mathbf{R}_j\})+\mathbf{G}_j$, where $\nabla_{\mathbf{R}_j}\equiv\partial/\partial\mathbf{R}_j$, $\xi=D/k_BT$ is the friction coefficient (D is the diffusion constant), and $\mathbf{G}_j$ is Gaussian random force: $\langle G_j^\alpha(t)\rangle=0$ and $\langle G_j^\alpha(t)G_{j'}^\beta(0)\rangle=2\xi k_BT\delta_{jj'}\delta_{\alpha\beta}\delta(t)$ $(\alpha,\beta=x,y,z)$.

2. Protein-protein complex

We model the forced rupture of the protein-protein complex $P_1\cdot P_2$ by the escape of the particle, along the pulling direction y, evolving on the potential of mean force $U(y,t)=U_0(y)+U_g(y)$, where $U_0(y)$ is the binding potential and $U_g(y)$ is the potential due to applied force. We take $U_0(y)$ to be a harmonic function of y with a cusplike barrier, i.e., $U_0(y)=\frac{1}{2}\kappa_m y^2$ for $y<y^*$ and $U_0(y)=-\infty$ for $y\geq y^*$, where κ_m is the spring constant and y^* is the critical bond extension at which the complex dissociates. For the force-ramp protocol, we use $U_g(y)=1/2\kappa(y-\nu_0t)^2$, where κ is the cantilever spring constant and ν_0 is the pulling speed ($r_g=\kappa\nu_0$), and for the force clamp we set $\partial U(g)/\partial y=g_0$. We describe the bond rupture kinetics by the diffusive motion of the bond extension y on the potential $U(y,t)$ by following Langevin equation in the overdamped regime, $\xi(dy/dt)=-(dU/dy)+G$, where G is Gaussian random force.

[1] A. Matouschek, Curr. Opin. Struct. Biol. **13**, 98 (2003).
[2] C. M. Pickart, Annu. Rev. Biochem. **70**, 503 (2001).
[3] T. P. Stossel, J. Condeelis, L. Cooley, J. H. Hartwig, A. Noegel, M. Schleicher, and S. S. Shapiro, Nat. Rev. Mol. Cell Biol. **2**, 138 (2001).
[4] D. Discher and P. Carl, Cell. Mol. Biol. Lett. **6**, 593 (2001).
[5] B. T. Marshall, M. Long, J. W. Piper, T. Yago, R. P. McEver, and C. Zhu, Nature (London) **423**, 190 (2003).
[6] B. Geiger, A. Bershadsky, R. Pankov, and K. M. Yamada, Nat. Rev. Mol. Cell Biol. **2**, 793 (2001).
[7] D. Leckband, Curr. Opin. Struct. Biol. **14**, 524 (2004).
[8] R. P. McEver, Curr. Opin. Cell Biol. **14**, 581 (2002).
[9] V. Barsegov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **102**, 1835 (2005).
[10] S. Labeit and B. Kolmerer, Science **270**, 293 (1995).
[11] W. Linke, M. Ivemeyer, N. Olivieri, B. Kolmerer, J. Ruegg, and S. Labeit, J. Mol. Biol. **261**, 62 (1996).
[12] M. E. Chicurel, C. S. Chen, and D. E. Ingber, Curr. Opin. Cell Biol. **10**, 232 (1998).
[13] J. W. Weisel, Biophys. Chem. **112**, 267 (2004).
[14] J. W. Weisel, Science **320**, 456 (2008).
[15] S. T. Lord, Curr. Opin. Hematol. **14**, 236 (2007).
[16] E. Evans, A. Leung, V. Heinrich, and C. Zhu, Proc. Natl. Acad. Sci. U.S.A. **101**, 11281 (2004).
[17] A. E. X. Brown, R. I. Litvinov, D. E. Discher, and J. W. Weisel, Biophys. J. **92**, L39 (2007).
[18] R. I. Litvinov, O. V. Gorkun, S. F. Owen, H. Shuman, and J. W. Weisel, Blood **106**, 2944 (2005).
[19] R. I. Litvinov, O. V. Gorkun, D. K. Galanakis, S. Yakovlev, L. Medved, H. Shuman, and J. W. Weisel, Blood **109**, 130 (2007).
[20] A. E.-M. Clemen, M. Vilfan, J. Jaud, J. Zhang, M. Baermann, and M. Rief, Biophys. J. **88**, 4402 (2005).
[21] D. A. Simson, F. Ziemann, M. Strigl, and R. Merkel, Biophys. J. **74**, 2080 (1998).
[22] E. Evans, Annu. Rev. Biophys. Biomol. Struct. **30**, 105 (2001).
[23] R. Merkel, P. Nassoy, A. Leung, K. Ritchie, and E. Evans, Nature (London) **397**, 50 (1999).
[24] R. Perez-Jimenez, S. Garcia-Manyes, S. R. K. Ainavarapu, and J. M. Fernandez, J. Biol. Chem. **281**, 40010 (2006).
[25] J. Brujić, R. I. Z. Hermans, K. A. Walther, and J. M. Fernandez, Nat. Phys. **2**, 282 (2006).
[26] C. P. Johnson, H.-Y. Tang, C. Carag, D. W. Speicher, and D. E. Discher, Science **317**, 663 (2007).
[27] M. Mickler, R. Dima, H. Dietz, C. Hyeon, D. Thirumalai, and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **104**, 20268 (2007).
[28] V. Barsegov, V. Chernyak, and S. Mukamel, J. Chem. Phys. **116**, 4240 (2002).
[29] V. Barsegov and S. Mukamel, J. Chem. Phys. **116**, 9802 (2002).
[30] V. Barsegov and S. Mukamel, J. Chem. Phys. **117**, 9465 (2002).
[31] J. M. Fernandez and H. Li, Science **303**, 1674 (2004).
[32] J. Brujić, R. I. Z. Hermans, S. Garcia-Manyes, K. A. Walther, and J. M. Fernandez, Biophys. J. **92**, 2896 (2007).
[33] H. Dietz and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **103**, 1244 (2006).
[34] F. Oesterhelt, D. Oesterhelt, M. Pfeiffer, A. Engel, H. E. Gaub, and D. J. Mueller, Science **288**, 143 (2000).
[35] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Science **276**, 1109 (1997).
[36] M. S. Z. Kellermayer, L. Grama, A. Karsai, A. Nagy, A. Kahn, Z. L. Datki, and B. Penke, J. Biol. Chem. **280**, 8464 (2005).
[37] C. McAllister, M. A. Karymov, Y. Kawano, A. Y. Lushnikov, A. Mikheikin, V. N. Uversky, and Y. L. Lyubchenko, J. Mol. Biol. **354**, 1028 (2005).
[38] S. Guo and B. B. Akhremitchev, Biomacromolecules **7**, 1630 (2006).
[39] E. P. Raman, T. Takeda, V. Barsegov, and D. K. Klimov, J. Mol. Biol. **373**, 785 (2007).
[40] R. Dima and H. Joshi, Proc. Natl. Acad. Sci. U.S.A. **105**, 15743 (2008).
[41] V. Barsegov and D. Thirumalai, J. Phys. Chem. B **110**, 26403 (2006).
[42] E. Bura, D. K. Klimov, and V. Barsegov, Biophys. J. **93**, 1100 (2007).
[43] E. Bura, D. K. Klimov, and V. Barsegov, Biophys. J. **94**, 2516 (2008).
[44] C. Bustamante, Q. Rev. Biophys. **38**, 291 (2005).
[45] H. Westergaard, *Contributions to the History of Statistics* (Agathon, New York, 1968).
[46] A. Haber and R. P. Runyon, *General Statistics* (Addison-Wesley, Reading, Massachusetts, 1969).
[47] A. Sturges, J. Am. Stat. Assoc. **21**, 65 (1926).

[48] D. W. Scott, Biometrika **66**, 605 (1979).
[49] D. Freedman and P. Diaconis, Z. Wahrscheinlichkeitstheor. Verwandte Geb. **57**, 453 (1981).
[50] H. B. Li, A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **97**, 6527 (2000).
[51] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Proedel, J. Clarke, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **96**, 3694 (1999).
[52] G. Hummer and A. Szabo, Biophys. J. **85**, 5 (2003).
[53] M. Rosenblatt, Ann. Math. Stat. **27**, 832 (1956).
[54] E. Parzen, Ann. Math. Stat. **33**, 1065 (1962).
[55] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
[56] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (Wiley, New York, 1992).
[57] L. Wasserman, *All of Nonparametric Statistics* (Springer, New York, 2006).
[58] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans* (SIAM, Philadelphia, 1982).
[59] M. Rudemo, Scand. J. Stat. **9**, 65 (1982).
[60] J. S. Marron and M. P. Wand, Ann. Stat. **20**, 712 (1992).
[61] S. J. Sheather and M. C. Jones, J. R. Stat. Soc. Ser. B (Methodol.) **53**, 683 (1991).
[62] P. Hall and J. S. Marron, J. R. Stat. Soc. Ser. B (Methodol.) **53**, 245 (1991).
[63] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, New York, 1994).
[64] B. H. Lavenda, *Thermodynamics of Extremes* (Albion, Chichester, 1995).
[65] G. R. Terrell and D. W. Scott, Ann. Stat. **20**, 1236 (1992).
[66] B. A. Turlach, Discussion Paper No. 9317, Institut de Statistique, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, 1993.
[67] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference* (Dekker, New York, 2003).
[68] A. W. Van Der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, 1998).
[69] M. C. Jones, J. S. Marron, and S. J. Sheather, Comput. Stat. **11**, 337 (1996).
[70] P. Hall, Ann. Stat. **11**, 1156 (1983).
[71] C. J. Stone, Ann. Stat. **12**, 1285 (1984).
[72] R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2007.