Supplement: An information-based approach to Change-Point Analysis with applications to biophysics and cell biology.

Paul A. Wiggins
*Departments of Physics, Bioengineering and Microbiology, University of Washington, Box 351560. 3910 15th Avenue Northeast, Seattle, WA 98195, USA**

Contents

I. MODEL DEFINITIONS

A. State model probability distribution

The noise model in a particular state is (i) Markovian, (ii) gaussian and (iii) Stationary. Markovian implies that the memory of the noise is only a single time step. Stationary implies that the parameters describing the noise are constant in any given state, although these parameters can clearly change due to state transitions. Gaussian refers to the distribution of the noise around the mean value. Together these conditions imply that the model for the probability distribution (q) for observation $\vec{x}_i$, given $\vec{x}_{i-1}$ takes the following form:

$$q(x_i|x_{i-1};\boldsymbol{\theta}) = \left(\frac{k}{2\pi}\right)^{D/2} e^{-\frac{1}{2}\xi_i^2}, \tag{1}$$

$$\xi_i \equiv k^{1/2}\left(x_i - \overline{x}_i\right), \tag{2}$$

$$\overline{x}_i \equiv \varepsilon(x_{i-1}+\alpha) + (1-\varepsilon)(\mu + \alpha\Delta t_i^*),, \tag{3}$$

$$\Delta t_i^* \equiv t_i - t^*, \tag{4}$$

*Electronic address: pwiggins@uw.edu; URL: http://mtshasta.phys.washington.edu/

where D is the dimension of the observation $\vec{x}$, $\boldsymbol{\theta}$ is the vector of model parameters and we have made the observation vector sign implicit for clarity. Clearly these equations can be recast as a discrete stochastic process:

$$x_i = \overline{x}_i + k^{-1/2}\xi_i, \tag{5}$$
$$\xi_i \sim \mathcal{N}_D(0, \mathbb{1}_D), \tag{6}$$

where ξ_i are i.i.d. normally distributed random variables with variance one per dimension D. The connection between these discrete-time parameters and the underlying physical parameters used to describe the continuous-time process are discussed here [1].

II. ANALYTIC RESULTS FOR MLE PARAMETERS

A. Preliminaries

In this section we will write the algebraic expressions for the Maximum Likelihood Estimator (MLE) model parameters $\hat{\boldsymbol{\theta}}$ for a segment of the observations X_I. For the sake of brevity, we will drop the subscript I which will be implied (unless otherwise noted). Furthermore we shall number the indices starting at $i = 0$ for the boundary variable and ending at N, the number of variables in the segment. These results are derived in detail elsewhere [1].

The model probability density is

$$q(\vec{x}_i|\vec{x}_{i-1};\boldsymbol{\theta}) = \left(\frac{k}{2\pi}\right)^{D/2} \exp\left[-\tfrac{k}{2}\left(\Delta\vec{x}_i - \varepsilon\Delta\vec{x}_{i-1}\right)^2\right], \tag{7}$$
$$\Delta\vec{x}_i \equiv \vec{x}_i - \vec{\mu} - \vec{\alpha}\Delta t_i^*, \tag{8}$$
$$\Delta t_i^* \equiv t_i - t^*(\varepsilon), \tag{9}$$

where $\boldsymbol{\theta}$ is the vector of model parameters:

$$\boldsymbol{\theta} \equiv (k, \varepsilon, \vec{\mu}, \vec{\alpha}). \tag{10}$$

The information for the N observations is

$$h(\boldsymbol{\theta}|X) = \frac{ND}{2}\log\frac{2\pi}{k} + \frac{kV(\boldsymbol{\theta}|X)}{2}, \tag{11}$$

where V the summed variance to be defined below.

B. Evaluation by cumulative sum

For computational purposes, it is convenient to define quantities in terms of cumulative sums. These sums can be evaluated when the minimization is initiated and used throughout the calculation without the need for repeated

evaluation throughout the minimization process. We define the following cumulative sums:

$$\vec{X}_j \equiv \sum_{i=0}^{j} \vec{x}_i, \tag{12}$$

$$\vec{X}_{jk} \equiv \sum_{i=k}^{j} \vec{x}_i = \vec{X}_j - \vec{X}_{k-1}, \tag{13}$$

$$\vec{X}^{(\varepsilon)}_{jk} \equiv \vec{x}_j - \vec{x}_{k-1}, \tag{14}$$

$$C^{(0)}_j \equiv \sum_{i=0}^{j} \vec{x}_i \cdot \vec{x}_i, \tag{15}$$

$$C^{(0)}_{jk} \equiv \sum_{i=k}^{j} \vec{x}_i \cdot \vec{x}_i = C^{(0)}_j - C^{(0)}_{k-1}, \tag{16}$$

$$C^{(1)}_j \equiv \sum_{i=1}^{j} \vec{x}_i \cdot \vec{x}_{i-1}, \tag{17}$$

$$C^{(1)}_{jk} \equiv \sum_{i=k}^{j} \vec{x}_i \cdot \vec{x}_{i-1} = C^{(1)}_j - C^{(1)}_{k-1}, \tag{18}$$

$$\vec{P}^{(0)}_j \equiv \sum_{i=0}^{j} t_i \vec{x}_i, \tag{19}$$

$$\vec{P}^{(0)}_{jk} \equiv \sum_{i=k}^{j} t_i \vec{x}_i = \vec{P}^{(0)}_j - \vec{P}^{(0)}_{k-1} \tag{20}$$

$$\vec{P}_{jk} \equiv \vec{P}^{(0)}_{jk} - \bar{t}_{jk} \vec{X}_{jk}, \tag{21}$$

$$\vec{P}^{(\varepsilon)}_{jk} \equiv \tfrac{1}{2} N \left(\vec{x}_j + \vec{x}_{k-1}\right) - \tfrac{1}{2} \left(\vec{x}_j - \vec{x}_{k-1}\right) - \vec{X}_{j-1,k-1}. \tag{22}$$

Note that each of these sums depends only on the observations (random variables X).

C. Level mean MLE

It is convenient to introduce the effective stiffnesses for the level mean:

$$k_\mu \equiv N, \tag{23}$$

and the factors

$$A^{(\mu)} = (1-\varepsilon)^2 k_\mu, \tag{24}$$

$$\vec{B}^{(\mu)} = (1-\varepsilon)^2 \left(\vec{X}_{jk} + \frac{\varepsilon}{1-\varepsilon} \vec{X}^{(\varepsilon)}_{jk}\right), \tag{25}$$

in terms of which, the MLE level mean $\hat{\mu}$ is:

$$\hat{\vec{\mu}}(\varepsilon; X) = \frac{\vec{B}^{(\mu)}(X)}{A^{(\mu)}(X)}. \tag{26}$$

Note that the MLE level mean depends on ε.

D. Level slope MLE

It is convenient to introduce the effective stiffnesses for the level slope:

$$k_\alpha \equiv \frac{N(N-1)(N+1)}{12}, \tag{27}$$

and the factors

$$A^{(\alpha)} \equiv (1-\varepsilon)^2 k_\alpha \tag{28}$$

$$\vec{B}^{(\alpha)} \equiv (1-\varepsilon)^2 \left(\vec{P}_{jk} + \frac{\varepsilon}{1-\varepsilon} \vec{P}^{(\varepsilon)}_{jk} \right), \tag{29}$$

in terms of which, the MLE level slope $\hat{\alpha}$ is:

$$\hat{\vec{\alpha}}(\varepsilon; X) = \frac{\vec{B}^{(\alpha)}}{A^{(\alpha)}}. \tag{30}$$

Note that the MLE level slope depends on ε.

E. Coupling MLE

We now substitute the MLE level mean and slope back into the summed variance:

$$\begin{aligned} V(\vec{\mu}, \vec{\alpha}, \varepsilon; X) &= C^{(0)}_{jk} - 2\varepsilon C^{(1)}_{jk} + \varepsilon^2 C^{(0)}_{j-1,k-1} \cdots \\ &+ (1-\varepsilon)^2 k_\mu \left[\left(\vec{\mu} - \hat{\vec{\mu}} \right)^2 - \hat{\vec{\mu}}^{\,2} \right] \cdots \\ &+ (1-\varepsilon)^2 k_\alpha \left[\left(\vec{\alpha} - \hat{\vec{\alpha}} \right)^2 - \hat{\vec{\alpha}}^{\,2} \right], \end{aligned} \tag{31}$$

which is an expression for arbitrary level mean and slope but written more concisely in terms $\hat{\mu}$ and $\hat{\alpha}$.

Since $\hat{\mu}$ and $\hat{\alpha}$ depend implicitly on ε, we need to consider four possible cases. For both α and μ we consider the case where they are either set to an external value or they are set to the respective MLE value. Again, we now make the following convenient definitions analogous to those made for the level mean and slope:

$$A^{(\varepsilon)} \equiv C^{(0)}_{j-1,k-1} - R_\mu(\vec{\mu}) - R_\alpha(\vec{\alpha}) \tag{32}$$

$$B^{(\varepsilon)} \equiv C^{(1)}_{jk} - Q_\mu(\vec{\mu}) - Q_\alpha(\vec{\alpha}) \tag{33}$$

in terms of which, the MLE coupling $\hat{\varepsilon}$ is:

$$\hat{\varepsilon}(\vec{\mu}, \vec{\alpha}; X) = \frac{B^{(\varepsilon)}}{A^{(\varepsilon)}}. \tag{34}$$

The Qs and Rs are defined:

$$Q_\mu(\vec{\mu}) = \begin{cases} 0, & \vec{\mu} = 0 \\ k_\mu^{-1} \vec{X}_{jk} \cdot \left(\vec{X}_{jk} - \vec{X}^{\varepsilon}_{jk} \right), & \vec{\mu} = \hat{\vec{\mu}} \\ -k_\mu \vec{\mu}^{\,2} + 2\vec{\mu} \cdot \left(\vec{X}_{jk} - \frac{1}{2} \vec{X}^{\varepsilon}_{jk} \right), & \text{otherwise} \end{cases} \tag{35}$$

$$R_\mu(\vec{\mu}) = \begin{cases} 0, & \vec{\mu} = 0 \\ k_\mu^{-1} \left(\vec{X}_{jk} - \vec{X}^{\varepsilon}_{jk} \right)^2, & \vec{\mu} = \hat{\vec{\mu}} \\ -k_\mu \vec{\mu}^{\,2} + 2\vec{\mu} \cdot \left(\vec{X}_{jk} - \vec{X}^{\varepsilon}_{jk} \right), & \text{otherwise} \end{cases} \tag{36}$$

and the following relations for the level slopes:

$$Q_\alpha(\vec{\alpha}) = \begin{cases} 0, & \vec{\alpha} = 0 \\ k_\alpha^{-1} \vec{P}_{jk} \cdot \left(\vec{P}_{jk} - \vec{P}^{\varepsilon}_{jk} \right), & \vec{\alpha} = \hat{\vec{\alpha}} \\ -k_\alpha \vec{\alpha}^{\,2} + 2\vec{\alpha} \cdot \left(\vec{P}_{jk} - \frac{1}{2} \vec{P}^{\varepsilon}_{jk} \right), & \text{otherwise} \end{cases} \tag{37}$$

$$R_\alpha(\vec{\alpha}) = \begin{cases} 0, & \vec{\alpha} = 0 \\ k_\alpha^{-1} \left(\vec{P}_{jk} - \vec{P}^{\varepsilon}_{jk} \right)^2, & \vec{\alpha} = \hat{\vec{\alpha}} \\ -k_\alpha \vec{\alpha}^{\,2} + 2\vec{\alpha} \cdot \left(\vec{P}_{jk} - \vec{P}^{\varepsilon}_{jk} \right), & \text{otherwise} \end{cases} . \tag{38}$$

In combination, these results lead to algebraic equation that are uncoupled. For instance, if the level slope is set to zero by hand and μ and ε are both chosen to be their respective MLE values, we first compute $\hat{\varepsilon}$, then using $\hat{\varepsilon}$, we compute $\hat{\mu}$.

F. Information

After the summed variance has been computed it is straightforward to compute the information. The MLE stiffness is

$$\hat{k}(\vec{\mu},\vec{\alpha},\varepsilon;X) = \frac{ND}{V(\vec{\mu},\vec{\alpha},\varepsilon;X)}, \tag{39}$$

where D is the dimension of the space of the observations $\vec{x}_i$. In terms of the MLE stiffness k, the information can be written:

$$h(\boldsymbol{\theta}|X) = \frac{ND}{2}\left[\log\frac{2\pi}{k} + \frac{k}{\hat{k}(\vec{\mu},\vec{\alpha},\varepsilon;X)}\right], \tag{40}$$

which is a general and exact expression for the information for any parameter set $\boldsymbol{\theta}$ that can be computed rapidly and algebraically without the need for solving any coupled or transcendental equations.

III. GREEDY BINARY-SEGMENTATION ALGORITHM

In this section we introduce an algorithm for selecting the change-point indices $\boldsymbol{i} \equiv \{i_I\}_{I=1..n}$. This is a nontrivial problem since not only are the change-point indices unknown, but even the number of transitions (n) is unknown. The algorithm described here is called the Binary Segmentation Change-Point Algorithm and has been the subject of extensive study (e.g see the references in [2]). The Change-Point Algorithm is at its heart a data segmentation algorithm. The sequence of the observations is always maintained, but the data is divided into partitions, as specified by the change-points $\boldsymbol{i} \equiv \{i_I\}_{I=1..n}$. Every binary segmentation is *greedy*: i.e. we choose the change point that minimizes the information loss in that given step, without any guarantee that this is the optimum choice over multiple segmentations. The family of models generated by successive rounds of segmentation are said to be *nested* since successive changes points are added without altering the time indices of existing change points. Therefore, the previous model is always a special case of the new model. The binary segmentation process is shown schematically in Figure 1. In each step, after the optimum index for segmentation is identified, we statistically test the change in information loss (due to segmentation) to determine whether the new state is statistically supported. The algorithm is written explicitly in Table I in the main text.

In some situations the Change-Point Algorithm can suffer from non-convexity: Any possible segmentation leads to an increase in the unbiased estimator of information loss, but subsequent segmentation operations lead to reductions in the unbiased estimator of information loss. To avoid this problem, we typically segment the data using a complexity term half the true complexity. After the segmentation processes has been terminated, we reset the complexity term to its true value and merge neighboring regions using a greedy algorithm, choosing there merger that leads to the largest decrease in the unbiased estimator of information loss. The algorithm is written explicitly in Table I.

IV. STATISTICAL TESTS FOR CHANGE POINTS

There are two principle non-Bayesian classes of tests used to evaluate the existence of a change point[1]. These are (i) the Frequentist *Likelihood-Ratio Procedure Test* (LPT) and (ii) the *Informational-Based* approach. Chen and Gupta have compiled a summary of the literature which gives an extensive list of examples of each procedure as well as others [2]. The LPT test leverages detailed knowledge of a test statistic that has been specifically derived for a particular model. In an LPT test, the existence of a change point is tested against the null hypothesis that there is no change in a specified interval. A confidence level must be chosen by the investigator to test the null hypothesis. In some case there seems to be fairly general agreement about the correct statistic. For instance, to test for a level mean change for 1D observations x_i that are assumed to have equal and known variance, the U test statistic is used [2]. In the case of the rather general model that we have proposed, there are no existing test statistics to the author's knowledge. We therefore propose to take the information-based approach.

---

[1] We use the word *test* in an informal sense here since model selection criterion are not rigorously considered a statistical test.

Greedy Merge Algorithm

1. Merge state of model $\hat{\mathscr{M}}(\boldsymbol{i})$:
   (a) Compute the entropy change that results from all possible state mergers:
   $$\Delta h_I \leftarrow \hat{h}(\boldsymbol{i}|X) - \hat{h}(\{...,\not{i}_I,...\}|X), \tag{41}$$
   (b) Find the maximum entropy change Δh_I, and the corresponding index $I_{\max}$.
   (c) **If** the entropy change plus the nesting complexity is less than zero:
   $$\Delta h_I + \not{k}_- > 0 \tag{42}$$
   then then remove change-point i_I
      i. Add the new change-point to the change-point vector.
      $$\boldsymbol{i} \leftarrow \{...,\not{i}_I,...\} \tag{43}$$
      ii. Merge state of model $\hat{\mathscr{M}}(\boldsymbol{i})$.
   (d) **Else** terminate the merger process.

TABLE I: **The Greedy-Merge Algorithm.**

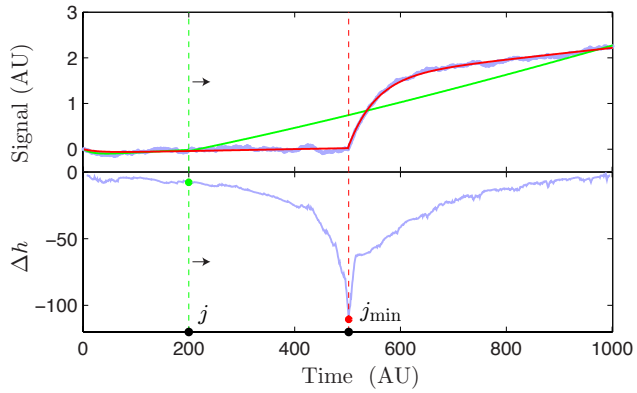

FIG. 1: **Schematic of binary segmentation.** To segment a partition, the information change due to placing a new boundary at each time point is computed. (The dashed red and green lines represent partition positions.) At each boundary position, a maximum likelihood fit is performed to the data (top panel, blue dots) in each of the two new partitions, resulting in the solid curves (top panel, red and green for the respective boundary positions). For each boundary position, an information change is computed (bottom panel). The partition is placed at the position that minimizes the information change (red dashed line), which maximizes the likelihood.

At the time Chen and Gupta wrote their review, the Bayesian Information Criterion (BIC) was essentially the only viable information-based approach. We recently proposed a new information criterion: the Frequentist Information Criterion (FIC). Note that we have proposed that this information-based approach and the LPT approach are essentially equivalent [1, 3].

V. ASYMPTOTIC SCALING OF THE COMPLEXITY FOR THREE INFORMATION CRITERIA.

We will not discuss the computation of the complexity here due to space limitations. A detailed description is given in Ref. [1].

The resolution of the Change-Point Algorithm is determined by the size of the complexity $\mathcal{K}$ in the information criterion. We now present a brief comparison of the FIC approximation for the complexity with two competing information criteria: the Akaike Information Criterion (AIC) and the Bayesisan Information Criterion[2] (BIC). Consider the change in the information (Δh) on the addition of a new state, parameterized by d local parameters describing the noise model. The additional state is considered predictive if it is larger than the nesting complexity, which is equal to the difference between the post and pre nesting complexities:

$$-\Delta h > \mathcal{k}_{-} \equiv \mathcal{K}_{n+1} - \mathcal{K}_n. \tag{44}$$

The asymptotic values for the nesting complexities for the three information criteria are:

$$-\Delta h > \mathcal{k}_{-} = \begin{cases} d, & \text{AIC (pathological)} \\ 2\log\log N + \mathcal{O}(\log\log\log N), & \text{FIC} \\ \frac{1}{2} d \log N, & \text{BIC} \end{cases}, \tag{45}$$

where we have assumed that the number of observations N is large and we have only preserved the leading order contribution to the complexity in the case of FIC. Note that the AIC complexity is known to be too small to terminate the segmentation process due to presence of unidentifiable parameters. On-the-other-hand, the Bayesian approach (as approximated by BIC) is known empirically to prevent over-fitting, at least in the asymptotic limit (large N, e.g. [5]). But, is the BIC complexity efficient in the sense that it balances the competing mechanisms of information loss of over and under-fitting to optimize model predictivity? Since FIC is defined to optimize the predictivity, BIC is only optimal when it is equal to FIC. In the asymptotic limit, the FIC complexity is expected to be smaller than the BIC complexity in two respects: In FIC (i) the leading order contribution to the complexity is independent of dimension of the noise model d and (ii) $\log\log N$ clearly increases more slowly with N than $\log N$. Therefore in the large N limit, the use BIC leads to under-fitting. We have also demonstrated that BIC also leads to overfitting at intermediate to small N values where its the justification for its applicability is somewhat ambiguous in any case [1].

Note that we *do not* recommend the direct application of the asymptotic FIC complexity since the complexity converges to the $2\log\log N$ limit very slowly. We therefore advocate a Monte Carlo computation of the complexity. Since the complexity is clearly very weakly dependent on N, we advocate the generation of a lookup-table from which complexity values can be interpolated on demand. See Ref. [1].

VI. APPLICATION TO SIMULATED DATA

We analyze simulated TPM data to test the performance of the Frequentist Information Criterion (FIC) and the Change-Point Algorithm under ideal circumstances.

Data simulation: The microscopic physics is a discrete Ornstein-Uhlenbeck Process which obeys Equation 6 with $k^{-1} = 1200\ \mathrm{nm}^{-2}$, $\vec{\mu} = 0\ \mathrm{nm}$, and ε alternating between the values of 0.92 and 0.70 by state. Naïvely this would appear to be a fairly good model for the TPM experiment: (i) Repressor-DNA binding induces DNA looping, reducing the effective tether length without changing the mean spatial position of the bead; therefore we expect the mean position $\vec{\mu}$ to be equal for all states. (ii) Similarly, diffusion is dominated by the movement of the bead; threfore we would expect the diffusion constant to be equal for all states[3]. The diffusion constant is parameterized by the stiffness k. (iii) Repressor binding shortens the effective DNA tether length by looping the tether. The DNA tether can be approximated as a linear spring with a tether-length-dependent spring constant. This spring constant is parameterized by the nearest-neighbor coupling ε and therefore we expect ε to be state dependent. The explicit mapping between the discrete-time stochastic parameters and the physical parameters is derived in Ref. [1]. In the interest of clarity we shall discuss results only in terms of the discrete parameters fit in the model, rather than the extrapolated physical parameters. We simulated a 2D trajectory with 5×10^4 frames and four transitions corresponding to five states. The simulated data are shown in Figure 2.

Analysis of simulated data. As a first step in the data analysis, we need to determine which family of nested change-point models to analyze. We set the level slope $\alpha = 0$ by hand. Now, we need to determine whether to fit the remaining parameters (k, ε and μ) as local (L) or global (G) parameters. We perform Change-Point Analysis for each

[2] Notes that despite its name, BIC is not an information criterion in the strict sense that it can be understood as an estimator of the cross entropy. See e.g. [4].

[3] We fluid coupling to the coverslip and the change in the effective diffusion constant as a result of the finite frame rate.

Model $\mathscr{M}$ (k, ε, μ)	Information Criterion ΔFIC (nats)	Number of States n	Parameters per State d
L 0 G	67,158	111	1
G G L	1,584	1	2
L G L	1,561	4	3
L G G	1,559	5	1
L L L	10	5	5
G L L	8	5	3
L L G	4	5	2
G L G	2	5	1
G L G Clustered	**0**	**5**	**1**

TABLE II: **Model selection for simulated TPM data.** We considered eight families of nested models where the parameters k, ε and μ are either optimized locally (L, a distinct value for each state I) or globally (G, identical values for all states). For each model family, the information criterion FIC is computed for the optimal change-point model (which minimizes FIC). The overall optimum model is chosen by minimizing FIC with respect to these minimal family models. In this case, the GLG model results in the minimum FIC, which corresponds to global (G) values for both μ and k and local state-specific values (L) for ε. A further reduction in FIC is achieved by clustering the states into two cluster corresponding to looped and unlooped states with common model parameters. We have grouped the models into three categories: unacceptable (top), acceptable (middle) and optimal (Bottom). All models except those with global values for ε lead to acceptable values for ΔFIC. Model "GLG Cluster" results in the smallest information loss since it contains the minimum number of parameters required to describe the data.

possibility: In each case, we minimize FIC for the family of nested models to find the optimal model in the model family. Once the family-optimum model has been determined, we record the FIC value for this model which is the unbiased estimate of information loss. We also consider a model that is constrained to be a gaussian process ($\varepsilon = 0$) rather than Ornstein-Uhlenbeck with a global mean (model L0G) since this was essentially the model employed by Finzi and coworkers [6]. A summary of the analysis of the simulated data is shown in Table II. In practice, the FIC values have a large constant offset and therefore only the FIC values relative to the minimum FIC value (ΔFIC) are shown in the table. The model with the lowest FIC is the model with the strongest statistical support. The relative difference between the FIC values for each model encodes the relative strength of the statistical support for the model. Four out of the initial eight models (black text) result in small information loss and are therefore expected to be close approximations to the truth. (We shall discuss the ninth, clustered model shortly.) As one might expect, the only models which fail are those in which ε is made global (gray text). Remember, this was the only parameter that changes between states in the simulated data and therefore we expect the analysis to fail when all states are constrained to take the same parameter value. By far the worst model is L0G (similar to that employed by Finzi and coworkers). It drastically overestimates the number of states, resulting in enormous information loss. As explained in the previous section, the correlations between successive observations in an Ornstein-Uhlenbeck Process are not accounted for by the microscopic model when $\varepsilon = 0$; therefore these correlations lead the Change-Point Algorithm to call states to explain these correlations. It should be noted that this is only a problem for states with ε close to one. In the experimental data from the Dunlap Lab discussed in the next section, ε is small enough to avoid this artifact. The remaining four models (black) are all acceptable, but the optimal model is the model where only ε is local: GLG (black). Remember that from the perspective of Maximum-Likelihood, more model parameters always results in a better fit and therefore choosing k and μ as global parameters is a non-trivial success.

We now turn our attention to a detailed look at the LLL model where all parameters are local. Although this is not the optimal model, the analysis of this model will be more pertinent to the analysis of the experimental data. The locations of the change points in both the GLG and LLL models were identical and the smaller GLG FIC value is due only to the reduction in information loss due to over-parameterization of the LLL model.

Since we are analyzing simulated data, we can check the fit model against the known true model. Every simulated data set generated (10) produced five states with accurately positioned change-points. The simulated data and the LLL model are shown in Figure 2. An experimental schematic is shown in Panel A, superimposed on the y-position trace of the bead. In the background of Panel A, the simulated trace of the y-position is shown, colored by state as identified by the Change-Point Algorithm with the state number plotted above the trace. To be clear, both x and y positions of the bead are simulated and analyzed but only the y positions are shown in the figure. The true change-point positions are known and are shown with dashed black lines. The determination of the change-points is extremely accurate, as is clearly observed in the zoomed region of the trace, shown as an inset in Panel A. The median distance between the estimated change point and the true change point was 10 frames in our simulations and analysis. This precision depends on the model parameters simulated. The qualitative features of Ornstein-Uhlenbeck Process are also clearly illustrated by the zoomed trace. On short times, the bead diffuses but on longer times the bead shows an autoregressive motion towards the mean position. ε parameterizes the lifetime of these

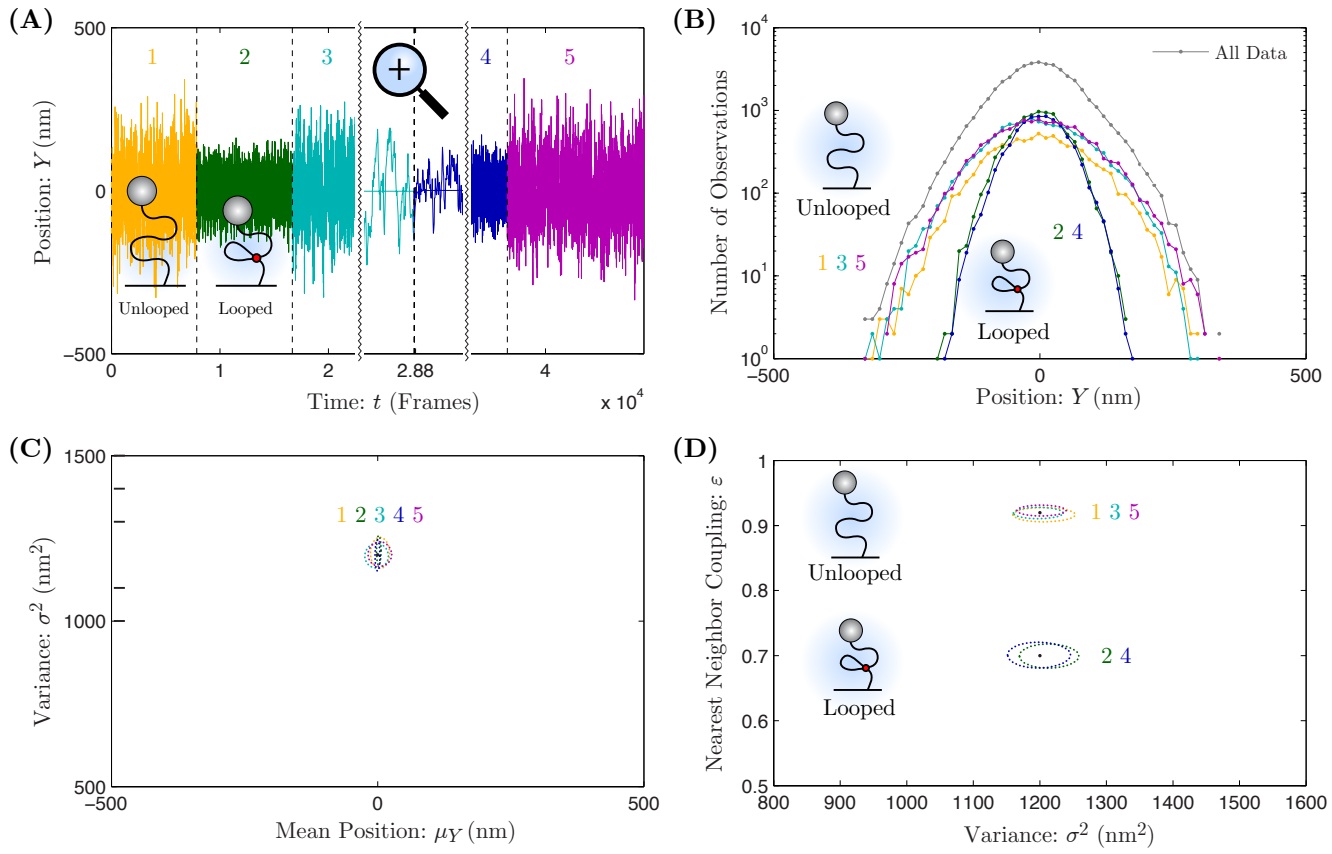

FIG. 2: **Analysis of simulated TPM data. Panel A: Experimental schematic and y-position trace.** The y-position trace is shown for 5×10^4 frames. Two classes of states are clearly identifiable: Unlooped and looped states with larger and smaller variance in bead position (respectively) resulting from changes in the effective tether length. Change-point analysis was employed to indentify the system state and the traces are colored by state with the state number plotted above the trace. The true change points are shown as dashed black lines. We have shown a high-time-resolution inset in the trace for the transition between states 3 and 4 to show the accuracy of the change-point detection. **Panel B: Histogram of y-position by state.** The histogram of all y-positions and positions by state are shown. Two types of states are observed: an unlooped state (1, 3, 5) and a looped state (2, 4). **Panels C & D: State model parameters.** Dotted curves represent 95% confidence regions for parameter values by state. Black points represent true parameter values. The clusters corresponding to the looped and unlooped states are clear in the plot of stiffness $k = \sigma^{-2}$ versus nearest-neighbor coupling ε. Since all states have the same values of μ and k, only a single cluster is observed in the mean versus variance plot.

fluctuations. In Panel B, the histogram of bead y-position is shown for each state. Unlike the experimental data, the probability distributions are gaussian. In Panels C and D the state model parameters are shown by state. The dotted curves represent the 95% confidence region for each, colored by state. As the reader can see, the model parameters are correctly estimated by the analysis.

State clustering. The tight clustering of states $\{1, 3, 5\}$ and $\{2, 4\}$ in panels C and D immediately suggest that these states are described by identical parameters. Again, it is straight forward to investigate such a model: We start with the optimal GLG model and constrain all parameters to be equal for the two state clusters respectively[4]. The resulting FIC value for the model "GLG Clustered" is smaller than the GLG model and therefore optimal. Again, this should come as no surprise since this was precisely the model that was simulated.

In summary, we have optimized families of nested models with different numbers of state parameters using the Change-Point Algorithm and the Frequentist Information Criterium (FIC). The resulting optimal models were then compared using the FIC information criterion to choose between different families of models. This technique enabled

[4] It should be note that automated clustering poses problem similar to the Change-Point Algorithm. Many techniques have been proposed. E.g. see references in [7].

us to identify the optimal model with the same number of model parameters as the true model that was simulated. The optimal model identified change points (state transitions) that were essentially identical to the simulated model and the resulting estimated parameters closely matched those simulated. Although this result is satisfying, it is not necessarily indicative of success in the context of experimental data. It is important to note that the existence of a true model with a finite number of model parameters and furthermore a true model exactly equal to one of the candidate models is an unrealistic convenience of the analysis of simulated data [4].

---

[1] Wiggins, P. A., 2015. The development of an information criterion for Change-Point Analysis with applications to biophysics and cell biology. *Submitted to PRE.* .

[2] Chen, J., and A. K. Gupta, 2007. On change point detection and estimation. *Communications in Statistics–Simulation and Computation* 30:665–697.

[3] Wiggins, P. A., 2015. The Frequentist Information Criterion (FIC): The unification of information-based and frequentist inference. *In preparation.* .

[4] Burnham, K. P., and D. R. Anderson, 1998. Model selection and multimodel inference. Springer-Verlag New York, Inc., 2nd. edition.

[5] Kalafut, B., and K. Visscher, 2008. An objective, model-independent method for detection of non-uniform steps in noisy signals. *Computer Physics Communications* 179:716–23.

[6] Manzo, C., and L. Finzi, 2010. Quantitative analysis of DNA-looping kinetics from tethered particle motion experiments. *Methods Enzymol* 475:199–220.

[7] Watkins, L. P., and H. Yang, 2005. Detection of intensity change points in time-resolved single-molecule measurements. *J Phys Chem B* 109:617–28.