LETTER **Communicated by Pankaj Mehta**

The Development of an Information Criterion for Change-Point Analysis

Colin H. LaMont
lamontc@uw.edu
Department of Physics, University of Washington, Seattle, WA 98195, U.S.A.

Paul A. Wiggins
pwiggins@uw.edu
Departments of Physics, Bioengineering, and Microbiology, University of Washington, Seattle, WA 98195, U.S.A.

Change-point analysis is a flexible and computationally tractable tool for the analysis of times series data from systems that transition between discrete states and whose observables are corrupted by noise. The change point algorithm is used to identify the time indices (change points) at which the system transitions between these discrete states. We present a unified information-based approach to testing for the existence of change points. This new approach reconciles two previously disparate approaches to change-point analysis (frequentist and information based) for testing transitions between states. The resulting method is statistically principled, parameter and prior free, and widely applicable to a wide range of change-point problems.

1 Introduction

The problem of determining the true state of a system that transitions between discrete states and whose observables are corrupted by noise is a canonical problem in statistics with a long history (Little & Jones, 2011a). The approach we discuss in this letter, change-point analysis, was proposed by E. S. Page in the mid-1950s (Page, 1955, 1957). Since its inception, change-point analysis has been used in a great number of contexts and is regularly reinvented in fields ranging from geology to biophysics (Chen & Gupta, 2007; Little & Jones, 2011a, 2011b).

Change-point analysis is applied to a signal consisting of a series of observations generated by a stochastic process:[1]

[1]When X appears in capitals, it should be understood as a random variable, whereas it is a normal variable when it appears in lowercase. If we need a statistically independent set of variables of equal size, we will use the random variables Y^N, which have identical properties to the X^N.

Neural Computation **28, 594–612** (2016)
doi:10.1162/NECO_a_00809

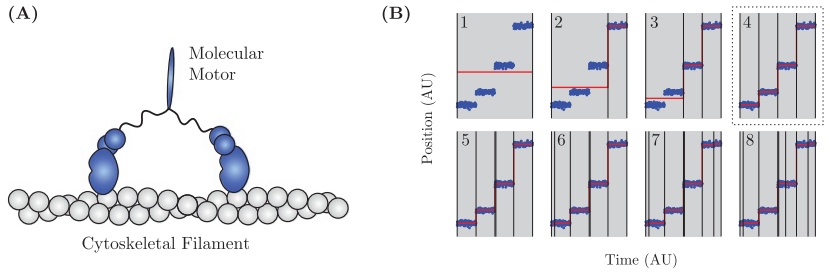

Figure 1: (A) Schematic of a biophysical system. One potential application of change-point analysis is to the characterization molecular motor stepping along a cytoskeletal filament. (B) Schematic of change-point analysis. A change-point model of motor stepping is shown for a series of position states. The blue dots represent measurements of motor position, corrupted by noise. The red line represents the change-point model for the true motor position. Each frame shows the optimal fit for $n = 1, \ldots, 8$ position states. From the figure, it is intuitively clear that $n = 4$ is the correct number of position states. Models with additional states improve the fit to the observed data but would result in information loss for an independent set of measurements of the same motor positions.

$$X^N \equiv (X_1, X_2, \ldots, X_N) \sim p(\cdot), \tag{1.1}$$

where the observation index is often, but not exclusively, temporal and the probability distribution for the stochastic process is represented as p.

1.1 The Change-Point Model. We define a model for the signal corresponding to a system transitioning between a set of discrete states. For example, a molecular motor transitions between position states as it steps along the cytoskelletal fillament. Each state generates a distinct distribution of measurements, as illustrated in Figure 1. We define the discrete time index corresponding to the start of the Ith state i_I. This index is called a change point. The model parameters describing the signal distribution in the Ith interval are θ_I. Together these two sets of parameters, i_I and θ_I, parameterize the model. The model parameterization for the signal (including multiple states) can then be written explicitly:

$$\Theta^n = \begin{pmatrix} 1 & i_2 & \ldots & i_n \\ \theta_1 & \theta_2 & \ldots & \theta_n \end{pmatrix}, \tag{1.2}$$

where n is the number of states or change points. The problem of change-point analysis is then to determine the number and location of change points with the parameter values describing the underlying states.

1.2 Model Selection and Predictivity. The central difficulty in change-point analysis is the problem of the bias–variance trade-off in selecting the dimension of the model: the determination of the number of states (or change points n). Adding states always improves the fit to the data, but overparameterization both reduces the model parsimony and results in a loss of model predictive performance. Akaike (1973) demonstrated that these two key principles of modeling, predictivity and parsimony, were in fact conceptually and mathematically linked. The addition of superfluous parameters to a model reduces predictivity (Burnham & Anderson, 1998). Under assumptions of model regularity (Watanabe, 2009), Akaike derived an unbiased estimator for the model predictivity, the Akaike information criterion (AIC), which proved to be exceptionally tractable and widely applicable.

Unfortunately, the change-point model is not regular; there exist singular points in parameter space for which the information matrix is not positive definite. As with nonanalytic points in complex analysis, the Taylor expansion of the information poorly approximates its behavior in the neighborhood of these singular points. The details of Akaike's derivation depend on the validity of this Taylor expansion, so AIC is not applicable to the change-point problem (Watanabe, 2009). Complicating matters, the data in a change-point problem are potentially structured and therefore are not necessarily independent and identically distributed for all observations X^N. These properties make the application of tools like naıve cross-validation and Watanabe's WAIC more difficult to apply (Gelman, Hwang, & Vehtari, 2014).

1.3 Proposed Approach. Our approach can be seen as a direct extension of AIC. In regular models, the expected information is quadratic about its minimum in parameter space. Realizations of the data generate maximum-likelihood estimators that fluctuate about this optimal value, in analogy with the thermal fluctuations of a particle confined to a harmonic potential. These fluctuations decrease the predictivity of models constructed using maximum likelihood procedure. AIC is derived through the consideration of these harmonic fluctuations. If a candidate change point I is supported by the data, then the continuous parameters θ_I are subject to harmonic confinement and their contribution to the model complexity is equal to their dimensionality, as Akaike predicted, while the change point i_I, as a highly constrained discrete variable, does not contribute to the complexity at all.

If a candidate change point is unsupported, the maximum likelihood change point is not constrained; it can be realized anywhere over a candidate interval. We have recently proposed a frequentist information criterion (FIC) applicable even in the context of singular models. Using FIC, we find that the information as a function of change-point location can then be approximated with the squared norm of a Brownian bridge and that expected

predictive loss can be estimated with a modified measure of the model complexity derived from this description. Consideration of these two distinct behaviors gives a piecewise information criterion that does not depend on the detailed form of the model for the individual states, only on the number of model parameters, in close analogy with AIC. Therefore, we expect this result to be widely applicable anywhere the change-point algorithm is applied.

1.4 Relation to Frequentist Methods. Frequentist statistical tests have been defined for a number of canonical change-point problems. It is interesting to examine the relation between this approach and our newly derived information-based approach. We find the approaches are fundamentally related. The information-based approach can be understood to provide a predictively optimal confidence level for a generalized ratio test. The Bayesian information criterion (BIC) has also been used in the context of change-point analysis. We find significant differences between our results and the BIC complexity that suggest that BIC is not suitable for application to change-point analysis.

2 Preliminaries

The essential notation is summarized in Table 1. We represent the probability distribution for a change-point model Θ^n as

$$q(X^N|\Theta^n). \tag{2.1}$$

2.1 Information and Cross-entropy. The information for signal X^N given model Θ is

$$h(X^N|\Theta^n) \equiv -\log q(X^N|\Theta^n), \tag{2.2}$$

and the cross-entropy for the signal (average information) is

$$H^N(\Theta^n) \equiv \mathop{\mathbb{E}_X}_{p(\cdot)} h(X^N|\Theta^n), \tag{2.3}$$

where the expectation over the signal X^N is understood to be taken over the true distribution p.

The state parameters, θ_l, and the change points, i_l, are fundamentally different parameters. We shall assume that the state model is regular: the parameters θ_l have nonzero Fisher information (LaMont & Wiggins, 2015). By contrast, the change-point indices i_l are discrete and typically nonharmonic parameters. For instance, consider a true model $p = q$ where $\theta_1 = \theta_2$. In this scenario, the cross-entropy will be independent of i_2 as long as $i_2 \in (i_1, i_3)$.

Table 1: Summary of Essential Notation for This Letter.

Data and observations	
$X^N,\ X^{[i,j]}$	All N observations/observations on interval $[i,\ j]$
$p(\cdot)$	True (unknown) distribution from which the data X^N were generated
$\mathbb{E}_{\substack{X \\ q(\cdot)}}$	Expectation over X taken with respect to q
Model parameterization	
i_I	Change point or first temporal index of state I
θ_I	Parameters describing state I
$\hat{\theta}_X$	The maximum likelihood estimator (MLE) of θ
Θ^n	Vector of θ_I and i_I describing n states
θ_0	True parameter values
Measures of information and entropy	
$h(X^N\|\Theta^n)$	Information for X^N (the negative of the log likelihood)
h_i	Information for the ith observation
$H^N(\Theta^n)$	N-observation cross entropy (expected information)
$\mathcal{K}(n)$	Complexity of a model with n states
IC	Information criterion or unbiased estimator of the cross-entropy
$k(n)$	Nesting complexity: $\mathcal{K}(n) - \mathcal{K}(n-1)$
Derivatives of information	
$\mathbf{x}_i$	Parameter gradient of information h_i
$\mathbf{X}$	Sum of the $\mathbf{x}_i$ (the negative of the score function)
$\mathbf{I}$	Fisher information (Hessian matrix of the information h_i)

The Fisher information corresponding to i_2 is therefore zero. These properties have important consequences for model selection (LaMont & Wiggins, 2015).

2.2 Determination of Model Parameters. Fitting the change-point model is performed in two coupled steps. Given a set of change-point indices $\boldsymbol{i}^n \equiv (i_1, \ldots, i_n)$, we hold the change points fixed and find the maximum likelihood estimators (MLE) of the state parameters $\boldsymbol{\theta}^n \equiv (\theta_1, \ldots, \theta_n)$. These are defined as

$$\hat{\boldsymbol{\theta}}^n_X = \arg\min_{\boldsymbol{\theta}^n} h(X^N|\Theta^n). \tag{2.4}$$

The determination of the change-point indices $\boldsymbol{i}^n$ is a nontrivial problem since not only are these unknown, but the number of transitions (n) is also unknown.

2.3 Binary Segmentation Algorithm. To determine the change-point indices, we will use a binary-segmentation algorithm that has been the subject of extensive study (see the references in Chen & Gupta, 2007). In the global algorithm, we initialize the algorithm with a single change point $i_1 = 1$. The data are sequentially divided into partitions by binary

segmentation. Every segmentation is greedy; that is, we choose the change point on the interval $(1, N)$ that minimizes the information 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 is said to be nested since successive change 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. In each step, after the optimum index for segmentation is identified, we statistically test the change in information (due to segmentation) to determine whether the new states are statistically supported. The n change-points determined by binary segmentation with their MLE state parameters compose $\hat{\Theta}^n$. We later distinguish between local and global segmentation: the local binary-segmentation algorithm differs from the global algorithm only in that we consider binary segmentation of each partition of the data independently. The algorithms are described explicitly in the online supplement.

2.4 Information-Based Model Selection. The model that minimizes the cross-entropy (see equation 2.3) is the most predictive model. Unfortunately, the cross-entropy cannot be computed: the expectation cannot be taken with respect to the true but unknown probability distribution p in equation 2.3. The natural estimator of the cross-entropy is the information (see equation 2.2), but this estimator is biased from below: due to overfitting, added model parameters always reduce the information, even as the predictivity of the model is reduced by the addition of superfluous parameters. To accurately estimate predictive performance, we construct an unbiased estimator of the cross entropy that we call the information criterion:

$$\mathrm{IC}(X^N, n) \equiv h(X^N|\hat{\Theta}^n_X) + \mathcal{K}(n), \tag{2.5}$$

where $\mathcal{K}$ is the complexity of the model, defined as the bias in the information as an estimator of cross-entropy:

$$\mathcal{K}(n) \equiv \mathbb{E}_{\substack{X,Y \\ p}}\{h(Y^N|\hat{\Theta}^n_X) - h(X^N|\hat{\Theta}^n_X)\}, \tag{2.6}$$

where the expectations are taken with respect to the true distribution p and X^N and Y^N are independent signals. Complexity is a measure of the flexibility of a family of models in fitting the observed data. A more complex model can be tuned to fit more features in the data, resulting in lower information than models with smaller complexity. However, the more complex model will be more prone to artificially decreasing the information relative to its optimally predictive parameter values and reducing the predictivity of the model by shifting the probability mass to accord with features not reproducible in different realizations of the data. The more flexible model

is expected to be more predictive only if the decrease in observed information is greater than the expected magnitude of these detrimental effects as measured by the complexity.

For a regular model in the asymptotic limit, the complexity is equal to the number of model parameters, and the information criterion is equal to AIC. In the context of singular models, a more generally applicable approach must be used to approximate the complexity.

2.5 Frequentist Information Criterion. The frequentist information criterion (FIC) uses a more general approximation to estimate the model complexity. Since the true distribution p is unknown, we make a frequentist approximation, computing the complexity for the model Θ^n as a function of the true parameterization,

$$\mathcal{K}_{\mathrm{FIC}}(\Theta^n, n) \equiv \mathop{\mathbb{E}_{X,Y}}_{q(\cdot|\Theta^n)} \{h(Y^N|\hat{\Theta}^n_X) - h(X^N|\hat{\Theta}^n_X)\}, \tag{2.7}$$

and the corresponding information criterion is defined,

$$\mathrm{FIC}(X^N, n) \equiv h(X^N|\hat{\Theta}^n_X) + \mathcal{K}_{\mathrm{FIC}}(\hat{\Theta}^n_X, n), \tag{2.8}$$

where the complexity is evaluated at the MLE parameters $\hat{\Theta}^n_X$. The model that minimizes FIC has the smallest expected cross-entropy.

2.6 Approximating the FIC Complexity. The direct computation of the FIC complexity (see equation 2.7) appears daunting, but a tractable approximation allows the complexity to be estimated. The complexity difference between the models is

$$\Bbbk(n) \equiv \mathcal{K}_{\mathrm{FIC}}(n) - \mathcal{K}_{\mathrm{FIC}}(n-1), \tag{2.9}$$

which is called the nesting complexity. An approximate piecewise expression can be computed as follows. Let the observed change in the MLE information for the addition of the nth change point be

$$\Delta h_n \equiv h(X^N|\hat{\Theta}^n_X) - h(X^N|\hat{\Theta}^{n-1}_X). \tag{2.10}$$

Consider two limiting cases. When the new parameters are identifiable, let the nesting complexity be given by $\Bbbk_+$, whereas when the new parameters are unidentifiable, let the nesting complexity be given by $\Bbbk_-$. When the new parameters are identifiable, the model is essentially regular; therefore

$$\Bbbk_+(n) = d, \tag{2.11}$$

where d is the number of harmonic parameters added to the model in the nesting procedure, as predicted by AIC.[2]

To compute k_-, we assume the unnested model is the true model and compute the complexity difference in equation 2.9. We then apply a piecewise approximation for evaluating the nesting complexity (LaMont & Wiggins, 2015):

$$k(n) \approx \begin{cases} k_-(n), & -\Delta h_n < k_-(n) \\ k_+(n), & \text{otherwise} \end{cases}. \tag{2.12}$$

Since the nesting complexity represents complexity differences, the complexity can be summed:

$$\mathcal{K}_{\text{FIC}}(n) \equiv \sum_{j=1}^{n} k(j), \tag{2.13}$$

where the first term in the series, $k(1)$, is computed using the AIC expression for the complexity. An exact analytic description of the complexity remains an open question.

3 Information Criterion for Change-Point Analysis

3.1 Complexity of a State Model. As a first step toward computing the complexity for the change-point algorithm, we will compute the complexity for a signal with only a single state. It will be useful to break the information into the information per observation. Assuming the process is Markovian, the information associated with the ith observation is

$$h_i(X^N|\theta) \equiv -\log q(X_i|X_{i-1};\theta). \tag{3.1}$$

For a stationary process, the average information per observation is constant $\overline{h} \equiv \mathbb{E}\, h$. The fluctuation in the information $\delta h_i \equiv h_i - \overline{h}$ has the property that it is independent for each observation:

$$\mathbb{E}\, \delta h_i\, \delta h_j = C_0 \delta_{ij}, \tag{3.2}$$

where C_0 is a constant and δ_{ij} is the Kronecker delta due to the Markovian property. In close analogy to the derivation of AIC, we will Taylor-expand

[2] Harmonic parameters are parameters with sufficiently large Fisher information that they are not unidentifiable.

the information in terms of the model parameterization θ around the true parameterization θ_0. We make the following standard definitions,

$$\delta\theta \equiv \theta - \theta_0, \tag{3.3}$$

$$\hat{\boldsymbol{I}}_i \equiv \nabla_\theta \nabla_\theta^T h_i(X^N|\theta_0), \tag{3.4}$$

$$\boldsymbol{I} \equiv \mathbb{E}_X \nabla_\theta \nabla_\theta^T h_i(X^N|\theta_0), \tag{3.5}$$

$$\boldsymbol{x}_i \equiv \nabla_\theta h_i(X^N|\theta_0), \tag{3.6}$$

$$\boldsymbol{X} \equiv \sum_i \boldsymbol{x}_i, \tag{3.7}$$

where $\delta\theta$ is the perturbation in the parameters and $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}_i$ are the Fisher information and its estimator, respectively. We make the cannonical approximation that the estimator is well approximated by the true value: $\hat{\boldsymbol{I}}_i \to \boldsymbol{I}$. The subscript i refers to the ith observation. Note that since the true parameterization minimizes the information by definition, $\mathbb{E}\,\boldsymbol{x}_i = 0$. Furthermore, equation 3.2 implies that

$$\mathbb{E}\,\boldsymbol{x}_i \boldsymbol{x}_j^T = \boldsymbol{I}\delta_{ij}, \tag{3.8}$$

where $\boldsymbol{I}$ is the Fisher information. The Taylor expansion of the information can then be written as

$$h(X^N|\theta) = h(X^N|\theta_0) + \delta\theta^T \boldsymbol{X} + \frac{1}{2}\delta\theta^T N\boldsymbol{I}\delta\theta + \mathcal{O}(\delta\theta^3), \tag{3.9}$$

to quadratic order in $\delta\theta$.

It is convenient to transform the random variables $\boldsymbol{x}_i$ to a new basis in which the Fisher information is the identity. This is accomplished by the transformation

$$\boldsymbol{x}_i' \equiv \boldsymbol{I}^{-1/2}\boldsymbol{x}_i, \tag{3.10}$$

$$\theta' \equiv \boldsymbol{I}^{1/2}\theta, \tag{3.11}$$

which results in the following expression for the information:

$$h(\theta|X_I) = h(X^N|\theta_0) + \delta\theta'^T \boldsymbol{X}' + \frac{1}{2}N\delta\theta'^T\delta\theta' + \mathcal{O}(\delta\theta^3). \tag{3.12}$$

In our rescaled coordinate system, $\boldsymbol{X}'$ can be interpreted as an unbiased random walk of N steps with unit variance in each dimension.

We determine the MLE parameter values:

$$\delta\hat{\theta}'_X = -\frac{1}{N}\boldsymbol{X}'. \tag{3.13}$$

To compute the complexity, we need the following expectations of the information:

$$\mathbb{E}_{X,Y}\, h(Y^N|\hat{\theta}_X) = \mathbb{E}_{X,Y}\, \left\{ h(Y^N|\theta_0) - \frac{1}{N}\boldsymbol{X}'^T\boldsymbol{Y}' + \frac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\theta^3) \right\}, \tag{3.14}$$

$$\mathbb{E}_{X}\, h(X^N|\hat{\theta}_X) = \mathbb{E}_{X,Y}\, \left\{ h(X^N|\theta_0) - \frac{1}{2N}\boldsymbol{X}'^2 + \mathcal{O}(\delta\theta^3) \right\}. \tag{3.15}$$

Since the signals X^N and Y^N are independent, the second term on the right-hand side of equation 3.14 is exactly zero. It is straightforward to demonstrate that

$$\mathbb{E}_X \boldsymbol{X}_I'^2 = N\mathcal{d}, \tag{3.16}$$

where $\mathcal{d}$ is the dimension of the parameter θ, which has an intuitive interpretation as the mean squared displacement ($\boldsymbol{X}'^2$) of an unbiased random walk of N steps in $\mathcal{d}$ dimensions. The complexity is therefore

$$\mathcal{K} \equiv \mathbb{E}_{X,Y}\, \left\{ h(Y^N|\hat{\theta}_X) - h(X^N|\hat{\theta}_X) \right\} = \mathcal{d}, \tag{3.17}$$

which is the AIC complexity.

This derivation of the AIC complexity through an expectation of a random walk in the score function $\boldsymbol{X}$ can now be extended to include the effects when the change point is not supported. When i_l is not fixed by the data, it is another a free parameter that can be chosen to maximize the decrease in information. The nesting complexity will then be the maximum mean squared displacement of many (correlated) random walks.

The first unsupported change point in a single state system is the first segmentation. We compute the nesting complexity $\mathcal{k}(2)$ of this first segmentation using equation 2.12. We will therefore generate the observations X^N and Y^N using the unsegmented model Θ^1. Remember that by convention, we assign the first change-point index to the first observation $i_1 = 1$. The optimal but fictitious change-point index for binary segmentation is

$$\hat{\imath}_2(X) = \arg\min_{1<i\leq N} \{\, h(X^{[1,i-1]}|\hat{\theta}_{X^{[1,i-1]}}) + h(X^{[i,N]}|\hat{\theta}_{X^{[i,N]}}) \,\}, \tag{3.18}$$

where the $X^{[j,k]}$ represent the respective partitions of the signal X^N made by the change point i. (Note that in the case of an autoregressive process, it is

possible to write overlapping partitions to account for the system memory.) The MLE model for two states is defined as

$$\hat{\Theta}^2_X \equiv \begin{pmatrix} 1 & i \\ \hat{\theta}_{X^{[1,i_2-1]}} & \hat{\theta}_{X^{[i_2,N]}} \end{pmatrix}. \tag{3.19}$$

To compute the nesting complexity, we compute the difference in the information between the two-state and one-state MLE models:

$$h(X^N|\hat{\Theta}^2_X) - h(X^N|\hat{\Theta}^1_X) = \min_{1<i\le N} \Bigg\{ \cancel{h(X^{[1,i-1]}|\theta_0)} + \cancel{h(X^{[i,N]}|\theta_0)} - \cancel{h(X^{[1,N]}|\theta_0)} - \frac{1}{2(i-1)} \boldsymbol{X}'^2_{[1,i-1]} - \frac{1}{2(N+1-i)} \boldsymbol{X}'^2_{[i,N]} + \frac{1}{2N} \boldsymbol{X}'^2_{[1,N]} \Bigg\}, \tag{3.20}$$

where $\boldsymbol{X}'_{[i,j]}$ are the $\boldsymbol{X}'$ computed in the two partitions of the data. The terms that are zeroth order in the perturbation cancel since the model is nested. (This equation is analogous to equation 3.15.) It is straightforward to compute the analogous expression for information difference for signal Y^N. The nesting penalty can then be written as

$$\hbar_-(2) \equiv \mathbb{E}_{\substack{X,Y \\ q(\cdot|\Theta^1_0)}} \left\{ h(Y^N|\hat{\Theta}^2_X) - h(X^N|\hat{\Theta}^2_X) - h(Y^N|\hat{\Theta}^1_X) + h(X^N|\hat{\Theta}^1_X) \right\} \tag{3.21}$$

$$= \mathbb{E}_{\substack{X \\ q(\cdot|\Theta^1_0)}} \max_{1<i\le N} \left\{ \frac{1}{i-1} \boldsymbol{X}'^2_{[1,i-1]} + \frac{1}{N+1-i} \boldsymbol{X}'^2_{[i,N]} - \frac{1}{N} \boldsymbol{X}'^2_{[1,N]} \right\}, \tag{3.22}$$

where the cross-terms between signals X^N and Y^N are zero since the signals are independent. It is now convenient to introduce a d-dimensional discrete Brownian bridge,

$$\boldsymbol{B}'_j \equiv \boldsymbol{X}'_{[1,j]} - \frac{j}{N} \boldsymbol{X}'_{[1,N]}, \tag{3.23}$$

by using the well-known relation between Brownian walks and bridges (Revuz & Yor, 1999). The Brownian bridge has the property that $\boldsymbol{B}'_0 = \boldsymbol{B}'_N = 0$, where each step has unit variance per dimension and mean zero. After some algebra, the nesting complexity can be written as

$$\hbar_-(2) = \mathbb{E}_{\substack{X \\ q(\cdot|\Theta^1_0)}} \max_{1\le j<N} \left\{ \frac{N}{j(N-j)} \boldsymbol{B}'^2_j \right\}. \tag{3.24}$$

It is not surprising that the nesting complexity should be well modeled by the square of a Brownian bridge. At the end points, the addition of a change point does nothing; it is indistinguishable from a change point

already in place. The complexity almost certainly increases: the smaller model is nested in the larger model. These observations are captured in the facts that $\boldsymbol{B}'_0 = \boldsymbol{B}'_N = 0$ and $\boldsymbol{B}^2 \geq 0$, respectively.

The details of the state model will determine the distribution function for the discrete steps in the Brownian bridge, but the central limit theorem implies that the distribution will approach the normal distribution. Therefore, it is convenient to approximate the discrete Brownian bridge $\boldsymbol{B}'_n$ as an idealized Brownian bridge with normally distributed steps,

$$\boldsymbol{B}'_j \to \boldsymbol{B}_j \equiv \sum_{i=1}^{n} \boldsymbol{b}_i, \quad \text{such that } \boldsymbol{B}_N = 0, \tag{3.25}$$

where the $\boldsymbol{b}_i$ are steps that are normally distributed with variance one per dimension d and mean zero. We now introduce a new random variable $U(N, d)$, the d-dimensional change-point statistic (Revuz & Yor, 1999),

$$U(N, d) \equiv \frac{1}{2} \max_{1 \leq j < N} \frac{N}{j(N-j)} \boldsymbol{B}_j^2, \tag{3.26}$$

which is a d-dimensional generalization of the change-point statistic computed by Hawkins (1977). In terms of the statistic U, the nesting penalty is

$$k_-(2) = 2\,\mathbb{E}_U\, U(N, d) = 2\overline{U}(N, d). \tag{3.27}$$

We will discuss the connection to the frequentist likelihood ratio test shortly.

3.2 Nesting Complexity for n States. The generalization of the analysis to n states is intuitive and straightforward. In the local binary-segmentation algorithm, segmentation is tested locally. The relevant complexity is computed with respect to the length of the Jth partition. It is convenient to work with the approximation that all partitions are of equal length since the complexity is slowly varying in N. We therefore define the local nesting complexity,

$$k_{L-}(n) = 2\,\mathbb{E}_U\, U\left(\frac{N}{n-1}, d\right) = 2\overline{U}\left(\frac{N}{n-1}, d\right), \tag{3.28}$$

where $\frac{N}{n-1}$ is the mean partition length. The nesting complexity for the binary segmentation of a single state is shown in Figure 2 for several different dimensions d, and compared with the complexity predicted by AIC and BIC.

In the global binary-segmentation algorithm, the next change point is chosen by identifying the best position over all intervals. We therefore

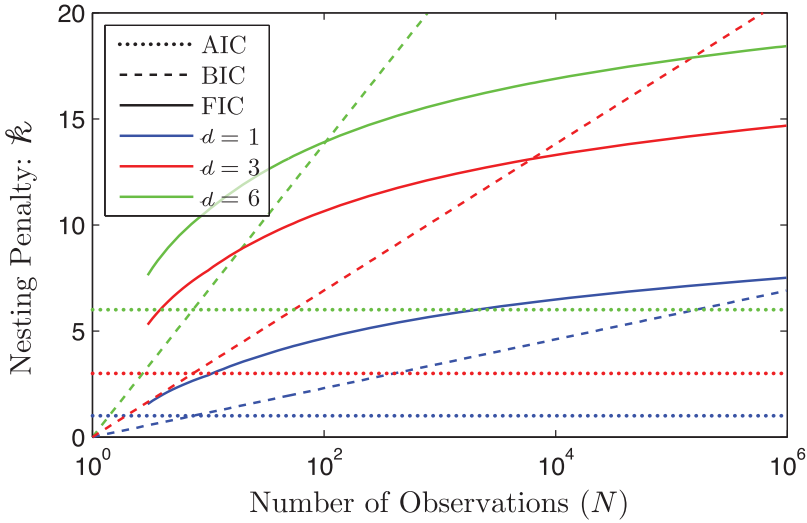

Figure 2: Nesting complexity for AIC, FIC, and BIC. The nesting complexity is plotted for three state dimensions: $d = \{1, 3, 6\}$ and $n = 2$. First, note that the AIC penalty is much smaller than the other two nesting complexities. BIC is empirically known to produce acceptable results under some circumstances. For sufficiently large samples (N), the $\kappa_{\mathrm{BIC}} > \kappa_{\mathrm{FIC}}$, resulting in overpenalization and the rejection of states that are supported statistically. This effect is more pronounced for large state dimension d, where the crossover occurs for small observation number N. κ_{BIC} is too small for a wide range of sample sizes, resulting in oversegmentation.

generalize all our expressions accordingly. We introduce a generalization of the change-point statistic where we replace N with a vector of the lengths of the constituent segment lengths $\boldsymbol{N}^n \equiv (N_1, \ldots N_n)$. We now define our new change-point statistic:

$$U_G(\boldsymbol{N}^n, d) \equiv \max_{1 \le i \le n} U(N_i, d). \tag{3.29}$$

Because it is computationally intensive to compute U_G for all possible segmentations $\boldsymbol{N}^n$, we assume that all the partitions are roughly the same size and consider n segments length $N/(n-1)$. Since the complexity is slowly varying in N, this does not in general lead to significant information loss.[3] We therefore introduce another change-point statistic,

[3]We empirically invesigated this equal-interval approximation, and it bounds the true complexity from above and is therefore conservative.

$$\mathcal{k}_{G-}(n) \equiv 2\,\mathbb{E}_U \max_{1\le i\le n}\left\{U_i\left(\frac{N}{n-1}, \mathcal{d}\right)\right\} \quad (\approx 2\,\mathbb{E}_U\, U_G(\boldsymbol{N}^n, \mathcal{d})), \tag{3.30}$$

that we will apply in the global binary-segmentation algorithm.

3.3 Asymptotic Expressions for the Nesting Complexity. It is straightforward to compute the asymptotic dependence of the nesting penalty on the number of observations N (Horváth, 1993; Horváth, Kokoszka, & Steinebach, 1999):

$$k_{G-}(n) \approx 2\log\log\frac{N}{n} + 2\log n + \mathcal{d}\log\log\log\frac{N}{n} + \ldots, \tag{3.31}$$

$$k_{L-}(n) \approx 2\log\log\frac{N}{n} + \mathcal{d}\log\log\log\frac{N}{n} + \ldots \tag{3.32}$$

These expression are slowly converging, and in practice, we advocate using Monte Carlo integration to determine the nesting penalty. If this is computationally cumbersome, equations 3.31 and 3.32 are useful in placing our approach in relation to existing theory.

Both the local and the global encoding have the same leading-order $2\log\log N$ dependence that has been advocated by Hannan and Quinn (1979), although interestingly not in this context. In contrast, this $2\log\log N$ dependence is in disagreement with the Bayesian information criterion, which has often been applied to change-point analysis. As illustrated by Figure 2, the BIC complexity,

$$\mathcal{K}_{\mathrm{BIC}} = \frac{\mathcal{d}}{2}\log N, \tag{3.33}$$

can be either too large or too small depending on the number of observations and the dimension of the model. It has long been appreciated that BIC can be only strictly justified in the large-observation-number limit. In this asymptotic limit, the BIC complexity is always larger than the FIC complexity due to the leading-order $\log N$ dependence, which will tend to lead to underfitting or undersegmentation. It is clear from Figure 2 that large ($N > 10^6$) may constitute much larger data sets than are produced in many applications.

3.4 Global versus Local Complexity. We proposed two possible parameter encoding algorithms that give rise to two distinct complexities: k_{L-} and k_{G-}. Which complexity should be applied in the typical problem? For most applications, we expect the number of states n to be proportional to the number of observations N. Doubling the length of the data set will result in the observation of twice as many change points on average. The application of the local nesting complexity clearly has this desired property

since it depends on the ratio of N/n. It is this complexity that we advocate under most circumstances.

In contrast, the global nesting complexity contains an extra contribution to the complexity $2\log n$. The reason is subtle. In the global binary-segmentation algorithm, one picks the best change point among n segments, and therefore complexity must reflect this added degree of choice. Consequently, a larger feature must be observed to be above the expected background. The use of the global nesting complexity makes a statement of statistical significance against the entire signal, not just against a local region. In the context of discussing the significance of the observation of a rare state that occurs just once in a data set, the global nesting complexity is the most natural metric of significance.

3.5 Computing the Complexity from the Nesting Complexity. To compute the FIC complexity, we sum the nesting complexities using equation 2.13. For data sets with identifiable change points, the FIC complexity is initially identical to AIC,

$$\mathcal{K}_{\mathrm{FIC}}(n) = n\mathcal{d}, \tag{3.34}$$

until the change in the information on nesting $\Delta h < \mathcal{k}_{\_}$, when FIC predicts a change in slope of the penalty. The FIC-, AIC-, and BIC-predicted complexities are compared with the true complexity for an explicit change-point analysis in Figure 3 C. It is immediately clear from this example that FIC quantitatively captures the true dependence of the penalty, including the change in slope at $n = 4$, exactly as predicted by the FIC complexity. As predicted, the AIC complexity is initially correct until the segmentation process must be terminated. At this point, the complexity increases significantly, with the result that the AIC complexity fails to terminate the segmentation process. In contrast, the BIC complexity is initially too large but fails to grow at a sufficient pace to match the true complexity for $n > 4$.

When a change point is supported by the data (i.e., its location is reproducible in multiple realizations of the observations), the complexity is approximated by the expectation of a single chi-squared variable (i.e., the AIC complexity). When a change point is unidentifiable (the location is determined by the noise and is not reproducibly positioned), the complexity is effectively equivalent to the expectation of the maximum of a number of independent chi-squared random variables and therefore is significantly larger than the AIC complexity (LaMont & Wiggins, 2015). These two distinct complexity behaviors are captured by our piecewise approximation.

4 The Relation between Frequentist and Information-Based Approach

Consider the likelihood-ratio test for the following problem. We propose the binary segmentation of a single partition. In the null hypothesis (H_0), the partition is described by a single state (unknown model parameters θ_0),

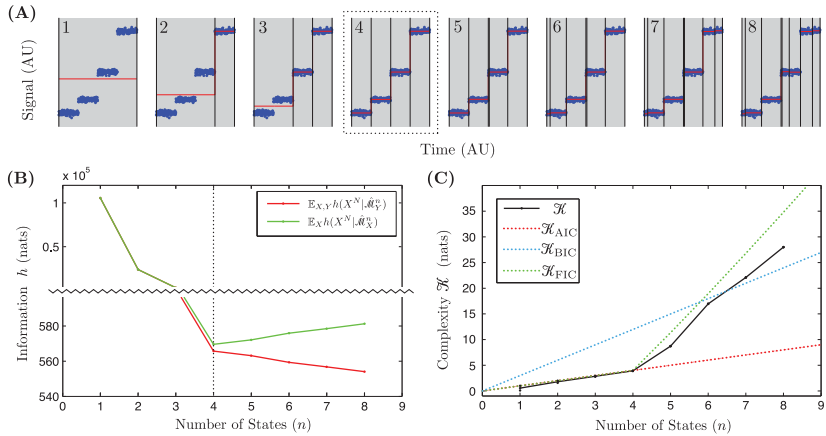

Figure 3: Information-based model selection. (A) Nested models generated by a change-point algorithm. Simulated data (blue points) generated by a true model with four states are fitted to a family of nested models (red lines) using a change-point algorithm. Models fit with $1 \leq n \leq 8$ states are plotted. The fit change points are represented as vertical black lines. The number of states (n) in each fit model is shown in the top-left corner of each panel. The true model has four states, and the fit model with four states is indicated with a dotted box. The models with five through eight states have superfluous states that are not present in the true model. (B) Four change points minimizes information loss. Both the expectation of the information (red) and the cross-entropy (green) are plotted as a function of the number of states n. The y-axis (h, information) is split to show the initial large changes in h, as well as the subsequent smaller changes for $4 \leq n \leq 8$. The cross-entropy (green) is minimized by the model that best approximates the truth ($n = 4$). The addition of parameters leads to an increase in cross-entropy (less predictive) as a consequence of the addition of superfluous parameters, as indicated by the increase of the cross-entropy (green) for $n \geq 4$. The information loss estimator (red) is biased and continues to decrease with the addition of states as a consequence of overfitting. In an experimental context, only the information can be computed since the true distribution is unknown. (C) Complexity of change-point analysis. The true complexity is computed for the model shown in panel A via Monte Carlo simulation for 10^6 realizations of the observations X^N and compared with three models for the complexity AIC, FIC, and BIC. For models with states numbering $1 \leq n \leq 4$, the true complexity (black) is correctly estimated by the AIC complexity (red dotted) and the FIC complexity (green). But for a larger number of states ($4 \leq n \leq 8$), only FIC accurately estimates the true complexity.

and the hypothesis to be tested (H_1) is that the partition is subdivided into two states: unknown change point and model parameters θ_1 and θ_2. We use the log-likelihood ratio as the test statistic:

$$V(X^N) \equiv \log \frac{q(X^N|\hat{\Theta}_X^2)}{q(X^N|\hat{\Theta}_X^1)} = h(X^N|\hat{\Theta}_X^1) - h(X^N|\hat{\Theta}_X^2). \tag{4.1}$$

In the Neyman-Pearson approach to hypothesis testing, we assume the null hypothesis (1 state) and compute the distribution in the test statistic V. As before, we will expand the information around the true parameter values θ_0. In exact analogy to equation 3.20, we find that V and our previously defined statistic U identically distributed,

$$V \sim U, \tag{4.2}$$

up to the approximations discussed in the derivation. Therefore, we will simply refer to V as U.

In the canonical frequentist approach, we specify a critical test statistic value u_γ above which the alternative hypothesis is accepted. u_γ is selected such that the alternative hypothesis H_1 is rejected given that the null hypothesis H_0 is true with a probability equal to the confidence level γ,

$$\gamma = F_U(u_\gamma), \tag{4.3}$$

where F_U is the cumulative distribution of U.

Therefore, we can interpret both the information-based approach and the frequentist approach as making use of the same statistic U. In the frequentist approach, a confidence level (γ) is specified to determine the critical value u_γ with which to accept the two-state hypothesis. The information-based approach also uses the statistic U, but the critical value of the statistic ($\hbar_-$) is computed from the distribution of the statistic itself $\hbar_- = 2\overline{U}$. The information-based approach chooses the confidence level that optimizes predictivity.

5 Applications

In the interest of brevity, we have not included analysis of either experimental or simulated data with a signal-model dimension larger than one, but we have tested the approach extensively. For instance, we have applied this technique to an experimental single-molecule biophysics application that is modeled by an Ornstein-Uhlenbeck process with a state-model dimension of four (Wiggins, 2015a). We also applied the approach in other biophysical contexts including the analysis of bleaching curves and cell and molecular-motor motility (Wiggins, 2015b).

6 Discussion

In this letter, we present an information-based approach to change-point analysis using the frequentist information criterion (FIC). The information-based approach to inference provides a powerful framework in which models with different parameterization, including different model dimension, can be compared to determine the most predictive model. The model with the smallest information criterion has the best expected predictive performance against a new data set.

Our approach has two advantages over existing frequentist-based ratio tests for change-point analysis. First, we derive an FIC complexity that depends on only the dimension of the state model (d), the number of states (n), and observations (N). Therefore, it may be unnecessary to develop and compute custom statistics for specific applications. Second, in the frequentist approach, one must specify an ad hoc confidence level to perform the analysis. In the information-based approach, the confidence level is chosen automatically based on the model complexity. The information-based approach is therefore parameter and prior free.

As the number of change-points increases, the model complexity is observed to transition between an AIC-like complexity $\mathcal{O}(N^0)$ and a Hannan-and-Quinn-like complexity $\mathcal{O}(\log\log N)$. We propose an approximate piecewise expression for this transition. The computation of this approximate model complexity can be interpreted as the expectation of the extremum of a d-dimensional Brownian bridge. We believe this information-based approach to change-point analysis will be widely applicable.

Acknowledgments

We thank K. Burnham, J. Wellner, L. Weihs and M. Drton for advice and discussions; D. Dunlap and L. Finzi for experimental data; and M. Lindén and N. Kuwada for advice on the manuscript. This work was supported by NSF MCB grant 1243492.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & E. Csaki (Eds.), *Proceedings of the 2nd International Symposium of Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and multimodel inference* (2nd ed.). New York: Springer-Verlag.

Chen, J., & Gupta, A. K. (2007). On change point detection and estimation. *Communications in Statistics: Simulation and Computation*, *30*(3), 665–697.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Hannan, E., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190–195.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, *72*(357), 180–186.

Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics*, *21*(2), 671–680.

Horváth, L., Kokoszka, P., & Steinebach, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, *68*, 96–119.

LaMont, C. H., & Wiggins, P. A. (2015). *The frequentist information criterion (FIC): The unification of information-based and frequentist inference*. Manuscript submitted for publication. arXiv:1506.05855.

Little, M. A., & Jones, N. S. (2011a). Generalized methods and solvers for noise removal from piecewise constant signals. I. background theory. *Proc. Math. Phys. Eng. Sci.*, *467*(2135), 3088–3114.

Little, M. A., & Jones, N. S. (2011b). Generalized methods and solvers for noise removal from piecewise constant signals. II. New methods. *Proc. Math. Phys. Eng. Sci.*, *467*(2135), 3115–3140.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, *42*, 523–527.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, *44*, 248–252.

Revuz, D., & Yor, M. (1999). *Continuous martingales and Brownian motion*. New York: Springer-Verlag.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge: Cambridge University Press.

Wiggins, P. A. (2015a). An information-based approach to change-point analysis with applications to biophysics and cell biology. *Biophys J.*, *109*, 346–354.

Wiggins, P. A. (2015b). *An information-based approach to change-point analysis with applications to biophysics and cell biology*. Unpublished manuscript.

Received June 26, 2015; accepted October 17, 2015.