Information-based inference in sloppy and singular models

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

A central problem in statistics is model selection: the choice between competing models of a stochastic process whose observables are corrupted by noise. In information-based inference, model selection is performed by maximizing the estimated predictive performance. We propose a frequentist information criterion (FIC) which extends the applicability of information-based inference to the analysis of singular and sloppy models. In these scenarios, the Akaike information criterion (AIC) can result in significant under or over-estimates of the predictive complexity. Two important mechanisms for this failure are examined: an implicit multiple testing problem and the presence of unidentifiable parameters. FIC rectifies this failure by applying a frequentist approximation to compute the complexity. For regular models in the large-sample-size limit, AIC and FIC are equal, but in general the complexity exhibits a sample-size dependent scaling. In the context of singular models, FIC can exhibit Bayesian information criterion-like or Hannan-Quinn-like scalings with sample size. FIC does not depend on ad hoc prior distributions or exogenous regularization and can be applied when structured data complicates the use of cross-validation.

PACS numbers: 02.50.Tt,02.50.Cw,02.50.Sk,07.05.Kf,05.30.Ch

I. INTRODUCTION

Over parameterization of a model results in a degradation of the predictive performance. The mechanism for this degradation is intuitive: predictive performance depends on the accurate determination of the model parameters and the uncertainty in each parameter contributes additively to the reduction in predictive performance [1]. Therefore there are two opposing imperatives in modeling: (i) to add parameters to the model in order to improve the fit while (ii) keeping the number of parameters to a minimum to avoid decreased predictive performance due to parameter uncertainty. The optimization of predictive performance with respect to these two competing qualitative goals is realized in an explicit quantitative mathematical form in the minimization of Akaike information criterion (AIC) [1, 2].

Although AIC succeeds in many contexts, it fails in some important modeling applications, e.g. [3]. Two important mechanisms of failure of AIC are the presence of unidentifiable parameters and model multiplicity. Models with unidentifiable parameters are called singular statistical models [4]. These models are of immense practical importance and include layered neural networks, normal mixtures, Bayes networks, Boltzmann machines, reduced rank regressions, hidden Markov models, time series prediction and biostatistics, among many others [4]. Recently, Watanabe and others have introduced methods addressing this problem in the Bayesian paradigm [4]. These methods require the introduction of a prior, Bayesian marginalization and, in the context of the Watanabe-Akaike information criterion (WAIC), unstructured data [5]. To avoid these requirements, we propose a new information criterion, the Frequentist Information Criterion (FIC), which is analogous to AIC, but more broadly applicable. We focus specifically on the application of FIC to models parameterized by maximum-likelihood estimators, but the FIC formulation can be generalized to evaluate the relative merits of models and learning algorithms of nearly any type. In the context of regular models in the large-sample size limit, FIC is equal to AIC, but for singular models FIC can exhibit both Bayesian Information Criterion (BIC)-like (log N) and Hannan-Quinn-like (log log N) scaling with sample size N. Although FIC can exhibit BIC-like scaling, FIC is a measure of model predictive performance, not an approximation of either the Bayesian evidence, like BIC, or the Minimum Description Length [6].

II. PRELIMINARIES

The true probability distribution $p$ is unknown. We approximate $p$ by training a model $M$ on a dataset of sample size $N$: $X = (X_1, ..., X_N)$. The model $M$ consists of a set of candidate probability distributions $q(x|\theta)$ for parameter $\theta \in \Theta$. If there exists a parameter $\theta$, such that $q(\cdot|\theta_t) = p(\cdot)$, the true distribution is realizable and the model $M$ is well-specified. In general we do not expect $p$ to be realizable for realistic applications in the context of experimental data. An important class of models is referred to as nested. Lower-dimension model $M_0$ is nested in higher-dimensional model $M_1$ if all candidate distributions in $M_0$ are realizable in $M_1$. 

A. Information and entropy

Model predictive performance is naturally measured by expectations of the information [1]. The Shannon information is defined [7]:

\[ h(X | \theta) \equiv -\log q(X | \theta). \]  

The expected information content of observations \( Y \sim p \) is the cross entropy, defined:

\[ H(\theta) \equiv E_{Y \sim p} \{ h(Y | \theta) \}, \]  

where \( E \) is the expectation of the random variable \( Y \sim p \) with the same sample size \((N)\) as the observed dataset \( X \). The maximum likelihood estimates (MLEs) \( \hat{\theta}_X \) minimize \( h \). The unknown realizable true parameters \( \theta_0 \) minimize the cross entropy \( H \).

B. Model selection

In information-based inference, model selection is performed by the selection of the model with the best expected predictive performance. The performance of the trained model \( \mathcal{M} \) is measured by the cross entropy \( \hat{H} \equiv H(\hat{\theta}_X) \) where \( H \) is unknown. A natural but negatively biased estimator of \( H \) is \( h(X | \hat{\theta}_X) \). The bias in this estimator results from the model being trained and validated against the same dataset \( X \). Therefore, the estimator \( h(X | \hat{\theta}_X) \) provides no protection against overfitting. The optimum model cannot be determined by the comparison of the estimator \( h(X | \hat{\theta}_X) \) without correction for the bias.

Aside. For readers unfamiliar with information-based inference, it is useful to consider the example of a family of nested models. As described below, a signal is expanded in Fourier modes and information-based model selection is applied to determine the number of Fourier modes \((K)\) to include in the model. In Panel C of Figure 2 both \( h(X | \hat{\theta}_X) \) and FIC (an unbiased estimator of \( H \)) are plotted as a function of model dimension. \( h(X | \hat{\theta}_X) \) decreases monotonically with model dimension, while FIC is minimized at finite model dimension as a consequence of balancing the competing imperatives of goodness of fit and loss due to parameter uncertainty.

C. The predictive complexity and information criterion

We define the predictive complexity as the negative bias in \( h \) as an estimator of \( H \) [1]:

\[ \mathcal{K}(p, \mathcal{M}) \equiv E_{Y \sim p} \{ H(\hat{\theta}_Y) - h(Y | \hat{\theta}_Y) \}, \]  

where the \( Y \sim p \) has the same sample size as observations \( X \). We can construct an unbiased estimator for the expected information \( \hat{H} \):

\[ \text{IC}(X, \mathcal{M}) \equiv h(X | \hat{\theta}_X) + \mathcal{K}, \]  

which is called an information criterion. The information criterion is unbiased since its expectation is equal to the expectation of \( H \) by construction. Therefore, models which minimize the information criterion are expected to be optimally predictive. The information criterion is extremely powerful in that it can be used to compare two distinct models, regardless of differences in the model parameterization [1, 2].

III. THE FREQUENTIST INFORMATION CRITERION

Although the definition of the information criterion appears promising, the computation of the complexity cannot be computed as defined since it depends on expectations taken with respect to the unknown true distribution \( p \). We propose a new approximation for its computation, inspired by frequentist statistics where an analogous problem arises. In a likelihood-ratio test for composite hypotheses, the true distribution is also unknown, but the distribution of the test statistic is computed by sampling from the candidate distributions parameterized by the MLEs \( \theta \) [8, 9].
We therefore propose the following two approximations: (i) Although we do not expect the true distribution to be realizable, we shall compute the complexity with respect to the candidate distributions \( q \):

\[
\mathcal{K}_{\text{FIC}}(\theta) \equiv \mathbb{E}_{Z,Y \sim q(\cdot|\theta)} \left\{ h(Y|\hat{\theta}_Z) - h(Z|\hat{\theta}_Z) \right\},
\]

(5)

where \( Y, Z \sim q(\cdot|\theta) \) are two samples with the same sample size as \( X \) and we have written the cross entropy as an the expectation of the information. The complexity depends on the parameter \( \theta \), identifying the data generating distribution. Ideally, we would compute the complexity using the true parameter \( \theta = \theta_0 \) which gives the best approximation of the true distribution \( p \), but \( \theta_0 \) it is unknown. (ii) We will therefore approximate \( \theta_0 \) with its MLE \( \hat{\theta}_X \). We call the resulting information criterion:

\[
\text{FIC}(X, \mathcal{M}) \equiv h(X|\hat{\theta}_X) + \mathcal{K}_{\text{FIC}}(\hat{\theta}_X),
\]

(6)

the frequentist information criterion. The FIC complexity is frequentist only in the sense that the complexity is sampled from the candidate distributions. No confidence level is specified. The FIC complexity can be understood as an exact complexity computed from the wrong true distribution. It is exact because Eqn. 5 does not depend on empirical expectations or a perturbative expansion but is computed for \( q(\cdot|\hat{\theta}_X) \) rather than the unknown true distribution \( p \).

IV. THE PROPERTIES OF THE COMPLEXITY AND THE FAILURE OF AIC

In this section we develop an approximate analytic treatment of the computation of the complexity for both regular and singular models. We have two main aims: (i) to describe mechanisms for the failure of AIC and (ii) to develop a computational tool which facilitates the estimation of the FIC complexity in some simple models.

A. Error statistic

The information and entropy measure absolute model performance. For this discussion it will be more convenient to work in terms of the relative performance compared to the true parameter \( \theta_0 \). We therefore define the Kullback-Leibler statistic:

\[
d(\theta_0||\theta) \equiv h(X|\theta) - h(X|\theta_0),
\]

(7)

which will be interpreted as a random variable due to its implicit dependence on \( X \). The realizable Kullback-Leibler divergence is the expectation of the Kullback-Leibler statistic:

\[
D(\theta_0||\theta) \equiv H(\theta) - H(\theta_0),
\]

(8)

where \( D \) is the realizable Kullback-Leibler divergence since \( p \) may not be equal to \( q(\cdot|\theta_0) \) [4]. \( d \) is the empirical estimator of \( D \). The parameters that minimize \( d \) and \( D \) also minimize \( h \) and \( H \) respectively. \( d \) and \( D \) are interpreted as the information loss and average information loss of encoding message \( X \) with parameters \( \theta \) relative to the true parameterization \( \theta_0 \). The estimators \( h \) and \( d \) have the same bias, as shown in the supplement. To compute the bias of \( h \), we define a error statistic \( \kappa \) as the minus error of \( d \), when interpreted as an estimator of \( D \):

\[
\kappa(\theta) \equiv D(\theta_0||\theta) - d(\theta_0||\theta).
\]

(9)

The function \( \kappa \) is implicitly dependent on \( X \) and will therefore be interpreted as a random variable. It is straightforward to demonstrate that the complexity is the expectation of the error statistic evaluated at the MLEs:

\[
\mathcal{K}(p, \mathcal{M}) \equiv \mathbb{E}_{X \sim p} \{ \hat{\kappa} \}.
\]

(10)

Understanding the statistical properties of the error statistic \( \hat{\kappa} \) will play a central role in understanding the failure of AIC.
B. The coordinate complexities

We will expand the error statistic to quadratic order in the difference between the MLE and true parameter \( \delta \theta \equiv \hat{\theta}_X - \theta_0 \):

\[
\kappa \approx \delta \theta^2 \equiv \delta \hat{\theta}^i I_{ij} \delta \hat{\theta}^j,
\]

where the inner product is computed using the metric \( I \equiv \partial \theta \otimes \partial \theta H(\theta_0) \), which can also be interpreted as the realizable Fisher information matrix for \( N \) observations. \( I \) is called a realizable Fisher Information since \( p \neq q(\cdot|\theta_0) \). The complexity can then be written as the sum of the coordinate complexities \( k_i \):

\[
\mathcal{K} \approx \sum_{i=1}^{K} k_i \quad \text{where} \quad k_i \equiv \mathbb{E}_{X \sim p} \left\{ (\delta \hat{\theta}^i)^2 \right\},
\]

and we transform to a hatted basis corresponding to a local coordinate system on parameter space where \( I_{ij} \) is the Kronecker delta.

C. The Akaike Information Criterion

Although the true distribution \( p \) might be unknown, the distribution of the error statistic is approximately independent of \( p \) in many important applications. In the large-sample-size limit of a regular \( K \)-dimensional model, the central limit theorem ensures that the MLEs are normally distributed about the true values: \( \delta \hat{\theta} \sim \mathcal{N}(0, I^{-1}) \) where we will drop the small-sample-size corrections \( \theta(N^{-1/2}) \). Therefore, the error statistic is distributed like a \( K \)-dimensional \( \chi^2 \) distribution:

\[
\kappa \sim \chi^2_K,
\]

\( k_i = 1 \) for all parameters, and the complexity is equal to the dimension of the model: \( \mathcal{K}_{AIC} = K \) \([1, 2]\). The resulting Akaike Information Criterion is:

\[
\text{AIC}(X, \mathcal{M}) \equiv h(X|\hat{\theta}_X) + K.
\]

One might have expected small, but non-zero, eigenvalues of the Fisher Information to contribute less to the complexity, but in a regular model, the increased variation in the MLEs exactly cancels the magnitude of the eigenvalues of \( I \). Though AIC is an excellent approximation of the true complexity in regular models, especially at large sample size \([1]\), AIC fails in the presence of model singularity or multiplicity (e.g. \([4, 10, 11]\)).

The AIC complexity has a close analogy in statistical physics: the equipartition theorem states that the average thermal energy in a system is \( \frac{1}{2} kT \) per coordinate (i.e. degree of freedom) for a Hamiltonian with quadratic dependence on coordinates. Equipartition is known to fail at both low and high temperature. At low temperature, the energy levels become discrete and degrees of freedom freeze out. At high temperature, Hamiltonians are often anharmonic (i.e. non-quadratic). We will describe analogies of both these modes of failure for AIC.

D. Unidentifiable parameters

Non-regular or singular models are the result of unidentifiable parameters. Structural unidentifiability at \( \theta_0 \) is defined as the existence of parameters \( \theta_1 \) such that \( q(\cdot|\theta_0) = q(\cdot|\theta_1) \) for \( \theta_0 \neq \theta_1 \). In a singular model, there is no coordinate transformation on the parameter space that can remove the unidentifiability, which implies \( k_i = 0 \) at \( \theta_0 \) in the large \( N \) limit since \( I \) becomes singular \([4]\). For finite sample size \( N \), we also define practical unidentifiability as \( k_i \approx 0 \).

For models where \( \theta^i \) is unidentifiable at \( \theta_0 \), \( D(\theta_0||\theta) \) has an approximately degenerate minimum for a region of parameter space \( \Theta_{min} \), as shown for model \( \mathcal{M}_1 \) in Fig. 1; whereas for a regular model there is a well-defined point-like minimum as shown for model \( \mathcal{M}_2 \) in Fig. 1. One mechanism for unidentifiability of \( \theta^i \) at \( \theta_0 \) is the proper width of the manifold \( \Theta \) in direction \( d\theta^i \) becomes vanishingly small: \( \int d\theta^i \sqrt{T_{ii}} \approx 0 \) \([12, 13]\), implying \( k_i \approx 0 \). This is a consequence of both small \( I_{ii} \) and the compactness of the manifold in direction \( d\theta^i \) in proper distance. This geometry has been referred to as a hyper-ribbon \([12, 13]\). Since the proper width itself is expected to be proportional to \( \sqrt{N} \), \( \theta^i \) is expected to transition from unidentifiable to regular for large enough \( N \) unless the parameter is structurally unidentifiable. We shall discuss a second mechanism for local unidentifiability in the development of the application below.
If the number of unidentifiable parameters is large, AIC can fail dramatically ($\mathcal{K} \ll K$) [4]. Such models would be significantly more predictive than estimated by AIC and therefore AIC would spuriously select smaller, but less predictive models. Models that do not require accurate values for some combinations of parameters are called sloppy models, and arise in many areas of science, from renormalizable theories in physics, to kinetic models in systems biology [14].

\section*{E. Multiplicity}

A second important phenomena leads to the increase of $k_i$ above unity. Unidentifiable parameters present a particularly acute problem when $\Theta_{\text{min}}$ includes many independent distributions. We define the number of independent distributions $m$ as the number independent regions of parameter space such that the error statistics $\kappa(\theta_i)$ and $\kappa(\theta_j)$ are not dependent in the vicinity of region $\Theta_{\text{min}}$:

$$\text{corr}[\kappa(\theta_i), \kappa(\theta_j)] \approx 0,$$  \hspace{1cm} (15)

where corr is the correlation between the error statistics. This is an approximate condition; in most non-trivial models where multiplicity arises, the error statistic $\kappa$ is the sum of both correlated and uncorrelated random variables, which modifies the condition in Eqn. 15, as described in the supplement.

As shown in Fig. 1, the presence of independent distributions is characterized by roughness in the information landscape of $d$ in model $\mathcal{M}_1$, generating many local minima. In this case we have a multiple testing problem or the look-elsewhere effect [15]: We have $m$ independent error statistics $\kappa_i$, one from each local minimum of $d$. These independent minima are shown for model $\mathcal{M}_1$ in Fig. 1, Panel B. Each independent error statistic is distributed like a $\chi^2$ random variable with $\nu$ degrees of freedom. The maximum likelihood procedure selects the largest of the $m$ error statistics and therefore the complexity can be interpreted as the expectation of the extremum of $m$ random variables:

$$\mathcal{K} \approx \mathbb{E}_{X \sim p} \left\{ \max_{i=1...m} \hat{\kappa}_i \right\},$$

$$\approx 2 \log m + (\nu - 2) \log \log m + ..., \hspace{1cm} (16)$$

where the last approximation assumes independent $\hat{\kappa} \sim \chi^2$ in the large $m$ limit [16]. The $m$ dependent contribution is analogous to the Bonferroni correction for the multiple testing problem in frequentist statistics [17].
For large \( m \), such as when \( m \approx N \), this complexity can be very much greater than the AIC estimate: \( \mathcal{K} \gg K \). This is shown qualitatively in Panel B of Fig. 1 where \( \kappa \) is larger for the singular model with multiplicity \( \mathcal{M}_1 \) than for the regular model \( \mathcal{M}_2 \). In the presence of multiplicity, AIC will show preference for models with reduced predictive performance due to overfitting.

**F. Classification of parameters by coordinates complexities**

We can summarize these behaviors in a classification system for the parameter coordinate \( \theta_i \) based on the coordinate complexity \( k_i \):

<table>
<thead>
<tr>
<th>Condition</th>
<th>Classification of ( \theta_i )</th>
<th>Consistent IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_i = 1 )</td>
<td>Regular, ( N \to \infty )</td>
<td>AIC &amp; FIC</td>
</tr>
<tr>
<td>( k_i \approx 0 )</td>
<td>Singular/unidentifiable</td>
<td>FIC</td>
</tr>
<tr>
<td>( k_i \approx 1 )</td>
<td>Regular, finite ( N )</td>
<td>FIC</td>
</tr>
<tr>
<td>( k_i \gg 1 )</td>
<td>Regular with multiplicity</td>
<td>FIC</td>
</tr>
</tbody>
</table>

AIC can fail by two important mechanisms: (i) parameter unidentifiability and (ii) an implicit multiple testing problem in the selection of the maximum likelihood parameter values. Both mechanisms can lead to very large mismatch between the AIC estimates and the true complexity. As a non-perturbative calculation, FIC allows the parameter manifold to constrain the variance of unidentifiable coordinates, correctly reducing their complexity contribution in the context of sloppy models. In models with multiplicity, FIC simulates the implicit multiple testing problem in the estimate of the model performance, and increases the model complexity to compensate (the Bonferroni correction). Even in the absence of sloppiness or singularity, FIC can realize small samples size corrections to AIC which can be of significant size [1].

**V. APPLICATION: SELECTION OF FOURIER MODES FOR TIME SERIES REGRESSION**

In this section we have two principal aims: (i) To present a model selection analysis using AIC, BIC and FIC and (ii) To demonstrate the dependence of the FIC complexity on the model fitting algorithm. We present a model of simulated data inspired by the measurements of the seasonal dependence of the neutrino intensity detected at Super-Kamiokande using Fourier regression. This will be a toy model in the sense that we will idealize and simplify the analysis in the interest of clarity and to facilitate analytic approximations, whilst providing examples of both sloppy (\( \mathcal{K} < K \)) and singular behavior (\( \mathcal{K} \gg K \)).

**A. Simulated data**

We simulate normally distributed intensities (units of AU) with unit variance: \( X_j \sim N(\mu_j, 1) \), where the true mean intensity \( \mu_j \) depends on the discrete-time index \( j \):

\[
\mu_j = \sqrt{120 + 100 \sin(2\pi j/N + \pi/6)} \text{ AU}
\]

and the sample size is \( N = 100 \). This true distribution is therefore unrealizable for a finite number of Fourier modes. The generating model, simulated data and two model fits are shown in Figure 2, Panel A.

**B. Analysis**

We expand the model mean (\( \mu_i \)) and observed intensity (\( X_i \)) in Fourier coefficients \( \tilde{\mu}_i \) and \( \tilde{X}_i \) respectively. A detailed description is provided in the supplement. The MLE that minimize the information is \( \hat{\mu}_i = \tilde{X}_i \). We now introduce two different approaches to encoding our low-level model parameters \( \{\tilde{\mu}_i\}_{i=-N/2...N/2} \): the sequential and greedy algorithms. In both cases, the models will be built by selecting a subset of the same underlying model parameters, the Fourier coefficients (\( \tilde{\mu}_i \)).
C. Sequential-algorithm model

In the sequential algorithm we will represent our nested-parameter vector as follows:

\[ \theta_{(n)} = \begin{pmatrix} \hat{\mu}_{-1} & \ldots & \hat{\mu}_{-n} \\ \hat{\mu}_0 & \hat{\mu}_1 & \ldots & \hat{\mu}_n \end{pmatrix}, \]  

(19)

where all selected \( \hat{\mu}_i \) are set to their respective maximum likelihood values and all other \( \hat{\mu}_i \) are identically zero. We initialize the algorithm by encoding the data with parameters \( \theta_{(0)} \). We then execute a sequential nesting procedure, increasing temporal resolution by adding the Fourier coefficients \( \mu_{\pm i} \) corresponding to the next smallest integer frequency index \( i \), in sequential order. (Recall there are two Fourier coefficients at every frequency, labeled \( \pm i \), except at \( i = 0 \).) The cutoff frequency is indexed by \( n \) and is determined by the model selection criterion.

From the AIC perspective, the complexity is simply a matter of counting the continuous parameters fit for each model as a function of the nesting index. Counting the parameters in \( K_{AIC} = 2n + 1 \), since both an \( \hat{\mu}_i \) and an \( \hat{\mu}_{-i} \) are added at every level. Since the parameters are regular, FIC predicts the same complexity as AIC. In the Bayesian analysis, the complexity is: \( K_{BIC} = \frac{1}{2}(2n + 1) \log N \), where \( N = 100 \), which is clearly significantly larger than the AIC and FIC. Panel B of Figure 2 shows FIC model selection for the sequential algorithm. The \( n = 2 \) nesting level minimizes FIC and this model \( (n = 2) \) is shown in Panel A. The true and FIC complexity are compared in Panel D for a sample size of \( N = 1000 \). Both AIC and FIC are excellent approximations of the true complexity.

D. Greedy-algorithm model

Instead of starting with the lowest frequency and sequentially adding terms, an alternative approach would be to consider all the Fourier coefficients and select the largest magnitude coefficients to construct the model. In the greedy algorithm we will represent the Fourier coefficients as

\[ \theta_{(n)} = \begin{pmatrix} 0 & i_1 & \ldots & i_n \\ \hat{\mu}_0 & \hat{\mu}_{i_1} & \ldots & \hat{\mu}_{i_n} \end{pmatrix}, \]  

(20)

where the first row represents the Fourier index and the second row is the corresponding Fourier coefficient. As before, all unspecified coefficients are set to zero. We initialize the algorithm by encoding the data with parameters \( \theta_{(0)} \) and then we execute a sequential nesting procedure: At each step in the nesting process, we choose the Fourier coefficient with the largest magnitude (not already included in \( \theta_{(n-1)} \)). The optimal nesting cutoff will be determined by model selection.

By naïve arguments of parameter counting, the AIC and BIC complexities remain unchanged. There are still two parameters at every nesting level \( n \). For the FIC complexity the distinction between the sequential and greedy algorithms has profound consequences. After the algorithm is initialized, each nesting step \( n \) chooses the largest Fourier coefficient \( \hat{\mu}_i \). If the coefficient is large, the coordinate complexity corresponding to the amplitude is \( k_{\mu} = 1 \) since it is regular. Since there is no significant statistical uncertainty in the value of \( i_n \), \( E_X (\delta i_n)^2 \approx 0 \) due to the discreteness of \( i_n \). Therefore \( k_i = 0 \), even though \( i_n \) identifiable in a global sense since \( \delta i_n \approx 0 \). We therefore call \( i_n \) locally unidentifiable. The change in the FIC complexity on the addition of a globally identifiable Fourier mode is therefore \( \Delta K_{FIC} = 1 \), or half the naïve AIC contribution. In analogy with quantum statistical mechanics, the \( i_n \) degree of freedom is frozen out; the spacing between parameter states is greater than the thermal uncertainty (here characterized by the information matrix eigenvalues).

When the remaining Fourier modes are too small to be globally identifiable, we have multiplicity. The largest-magnitude of the remaining \( N - m \) Fourier coefficients is selected, implying that \( i_n \) is globally unidentifiable and therefore \( k_i = 0 \) since the effective Fisher Information for the Fourier index shrinks to zero. But, in this case the Fourier coefficient have large multiplicity. There are \( m \approx N \) independent distributions since each Fourier coefficient is statistically independent. Therefore, when Fourier index is globally unidentifiable, we can apply Eqn. 16 with \( \nu = 1 \). The coordinate complexity for the Fourier coefficient is \( k_{\mu} \approx 2 \log N \). The change in the complexity due to the addition of a Fourier mode is \( \Delta K_{FIC} = 2 \log N \). The change in the FIC complexity is much larger than the change in the AIC complexity and twice the change in the BIC complexity. This discontinuous change in slope is roughly analogous to a phase transition as identifiable parameters (bound states) transition to unidentifiable parameters (free states). We give an explicit piecewise expression for the complexity in the supplement.

Panel B of Figure 2 shows FIC model selection for the greedy algorithm. The \( n = 2 \) nesting level minimizes FIC and this model \( (n = 2) \) is shown in Panel A. The true and FIC complexity are compared in Panel D for a sample size
FIG. 2: Panel A: Truth, data and models. (Simulated for \( N = 100 \).) The true mean intensity is plotted (solid green) as a function of season, along with the simulated observations (green points) and models fitted using two different algorithms, sequential (red) and greedy (blue). The sequential algorithm results in a significantly better fit to the observed data. Panel B: Fourier coefficient magnitudes. (Simulated for \( N = 100 \).) The magnitude of the Fourier coefficients \( \hat{\mu}_j \) is plotted as a function of frequency index \( j \) for the sequential-algorithm model. Below the cutoff (dotted red line), there is qualitative agreement between the true values (green points) and the fit coefficients (red points). The model selection criterion correctly identifies the transition from average information loss to gain, as illustrated by the widely divergent true and fit coefficients for \( j \geq 3 \). Panel C: Information as a function of model dimension. (Simulated for \( N = 100 \).) The information is plotted as a function of the nesting index \( n \). The true information is compared with the information for sequential (red) and greedy (blue) algorithm models. The dashed curves represent the information as a function of nesting index and both are monotonically decreasing. The solid curves (red and blue) represents the estimated average information (FIC), which is equivalent to estimated model predictivity. The model selection criterion chooses the model size (nesting index) that is a minimum of FIC. Panel D: The true complexity matches FIC estimates. (Simulated for \( N = 1000 \).) In the sequential-algorithm model, the true complexity (red dots) is AIC-like (solid red). In the greedy-algorithm model, the true complexity (blue dots) transitions from AIC-like (slope = 1) to BIC-like (slope \( \propto \log N \)) at \( n = 4 \). In both cases, the true complexity is correctly predicted by FIC (solid curve). The BIC complexity is a poor approximation of the predictive complexity in all models.

of \( N = 1000 \). This large sample size emphasizes the difference between the slopes. In the greedy algorithm, only FIC provides an accurate approximation of the true complexity. For large nesting index, the piecewise approximation made to compute the FIC complexity fails due to order statistics. This is of little consequence since the complexity in this regime is not relevant to model selection. AIC initially overestimates the complexity for \( n \leq 2 \) due to unidentifiability and then underestimates the complexity for \( n > 2 \) due to multiplicity. The use of AIC model selection in this context leads to significant over fitting by the erroneous inclusion of noise-dominated Fourier modes.

VI. DISCUSSION

Information-based approaches to model selection approximate the predictive performance of the model by estimating the predictive complexity. The true complexity cannot be computed exactly since the true probability distribution \( p \) is generally unknown. FIC and AIC are both understood as different approximations of the unbiased
estimator for the expected information \(\hat{H}\). The AIC derivation usually assumes two different approximations: (i) a perturbative expansion of the complexity (Eqn. 11) and (ii) approximating \(p\) with \(q\). Previous proposed refinements of AIC have often focused on eliminating approximation (ii) by using self-consistent empirical estimates of the expectations with respect to \(p\) using the observed data \(X\). (See TIC for instance [1, 18] or bootstrap approaches [1].) But typically these empirical expectations have high variance, which are themselves problematic [1]. In our experience with singular models, it is approximation (i) and not (ii) that leads to failure. FIC succeeds since it circumvents approximation (i) by computing the exact complexity of a distribution in the vicinity of \(p\). The failure of approximation (i) arising from parameter unidentifiability and multiplicity are a feature of the model not the data and are therefore captured in \(\mathcal{H}_{\text{FIC}}\). FIC can be understood to subsume some previously proposed generalizations of AIC. For instance, FIC is equal to the previously proposed AIC \(C\) [19] for linear regression models with unknown variance, as shown in Appendix B 1.

Although the FIC complexity does recover BIC-like scaling with sample size \(N\) for some singular models, FIC is fundamentally different from the Bayesian evidence or the MDL message length. The similarity between the BIC and FIC complexity is incidental and not realized generically. For instance, we have recently demonstrated that FIC results in Hannan-Quinn-like scaling \((\log \log N)\) in change-point models [10, 11]. In the context of singular models, the approximations assumed in the derivation of BIC fail for the same reason AIC does, model singularity, and therefore the use of BIC is not justified. A modified version of BIC must be applied which accommodates the singular nature of the model [20, 21]. Even when correctly approximated, inefficiency of Bayesian inference (and minimum description length methods) at finite sample size, emphasized in the Lindley Paradox, has long been appreciated [22–24].

VII. CONCLUSION

We have presented a detailed analysis of information-based model selection. We have identified two distinct mechanisms that lead to significant failures of AIC: parameter unidentifiability and multiplicity. We show that unidentifiability can result in AIC overestimating the true complexity, a phenomenon that occurs in sloppy models, whereas multiplicity can result in AIC significantly underestimating the true complexity due to an implicit multiple testing problem in singular models. We define the coordinate complexity \(k\) and present new information-based definition of the sloppiness phenomenon \((\mathcal{H} \ll K)\) and practical unidentifiability of a parameter coordinate \((k_\theta \approx 0)\). To extend the applicability of information-based inference to sloppy and singular models, we introduce a new frequentist information criterion which is widely tractable and applicable. FIC is also superior to AIC in regular models when the sample size is finite. Our proposed approach is objective and free from ad hoc prior probability distributions or regularizations and therefore offers a promising alternative to Bayesian model selection approaches in contexts where the existing information-based approach fails.

Acknowledgement

P.A.W. and C.H.L. would like to thank J. Kinney, 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-1243492.

Appendix A: Generality of FIC

1. FIC Applied to General Learning Algorithms

A model \(\mathcal{M}\) can be considered to be any algorithm or procedure, which when trained on observations \(X\) of sample size \(N\), generates a probability density \(q(\cdot | \mathcal{M}_X)\) meant to approximate the true distribution \(p(\cdot)\) of the data.

The general form of the complexity in terms of this extended conception of models is

\[
\mathcal{H}(p, \mathcal{M}) = \mathbb{E}_{Y, X \sim p} \left\{ h(Y|\mathcal{M}_Y) - h(Y|\mathcal{M}_X) \right\}.
\]  

(A1)

The FIC complexity is then calculated with the fitted model \(\mathcal{H}_{\text{FIC}} = \mathcal{H}(q(\cdot | \mathcal{M}_X), \mathcal{M})\). Because the algorithm is exactly applied to simulated data, FIC has an an implicit algorithmic sensitivity which makes it successful in the example studied in Sec. V.
This definition can be applied to any learning procedure with an output distribution. The major limitation is in the computational requirements needed to generate and fit enough data to assure convergence in the estimation procedure.

Appendix B: Relationship of FIC to other methods

1. FIC and parameter invariant complexities for well specified models.

The proposed criterion, FIC, provides a unifying context for an important class of previously proposed complexities. These complexities have two properties. First they can be derived using the assumption the the true distribution is realizable under the candidate model. Second this class of complexities is parameter invariant. A complexity is parameter invariant if for all $\theta, \theta' \in \Theta$,

$$K(q(\cdot|\theta), M) = K(q(\cdot|\theta'), M).$$

(B1)

If the true distribution is realizable under the model, the FIC procedure amounts to a change of parameter values $q(\cdot|\theta_t) \rightarrow q(\cdot|\hat{\theta}_X)$. If the complexity is parameter invariant, this change does not affect the complexity. Therefore we can state, FIC complexities are identical to parameter-invariant complexities derivable from the assumption of model realizability.

This class of complexities includes refinements on AIC which are important in contexts when the variance is unknown and the sample size is not very large compared to the number of parameters. The canonical example is $\text{AIC}_C$ where the relevant model is linear least-squares regression $[19]$. The complexity can then be shown to be

$$K_C(\text{linear regression model}) = K \frac{N}{N - K - 1},$$

(B2)

and thus parameter invariant $[1, 25]$. Another exact result is for the one parameter exponential distribution, $q(x|\lambda) = \lambda e^{-\lambda x}$, where the complexity can be shown to be

$$K_C(\text{univariate exponential model}) = \frac{N}{N - 1}. \quad \text{(B3)}$$

Also in this class is the AIC complexity itself, the parameter invariant complexity derivable in the large sample size limit of regular, well-specified models. The FIC complexity can be understood to generalize this class of complexities to situations which are not parameter invariant.

2. Cross-Validation, Empirical Bootstrap Methods, and FIC

Cross validation and related methods are, like FIC, non-pertubative estimates for model performance. These methods rely on the use of the observed data for validation. For example, in leave-one-out cross validation, the minimized criterion is

$$\text{LOOCV} = \sum_i h(X_i|\hat{\theta}_{X \neq i}). \quad \text{(B4)}$$

This estimates the cross-entropy with little bias, but often unacceptable variance $[26]$. There is a significant literature examining variants of the cross-validation scheme with different sized validation sets and different sampling procedures $[27]$ to reduce the variance of the estimator.

Our method is most similar to the empirical bootstrap methods reviewed by $[28]$, which tends to have worse bias than cross validation, but lesser variance $[26]$. Bootstrap methods rely on resampling the empirical distribution

$$p_X(x) = \frac{1}{N} \sum_i \delta(x - X_i). \quad \text{(B5)}$$

The complexity is then estimated using the actual data and resampled data for validation. One such complexity of this form is

$$K_{\text{boot}}(X, M) = \mathbb{E}_{Y \sim p_X} \left\{ 2h(X|Y, M) - 2h(X|M, X) \right\}. \quad \text{(B6)}$$
complexities for various models can be calculated once and compiled in look-up tables [11].

When analytic expressions are not available, FIC can still be more computationally attractive because the methods, ii.) direct evaluation of the expectation integrals makes analytic expressions and approximations feasible. iii.) an important addition to the catalogue of bootstrap methods (Sec. B 2)

We can characterize the variance of FIC by expanding the FIC complexity about the optimal parameters \( \theta_0 \),

\[
\mathcal{K}_{\text{FIC}}(\hat{\theta}_X) = \mathcal{K}_{\text{FIC}}(\theta_0) + \delta \theta \cdot \partial_{\theta} \mathcal{K}_{\text{FIC}}(\theta_0) + O(\delta \theta^2); \\
\text{var}(\mathcal{K}_{\text{FIC}}(\hat{\theta}_X)) \approx \partial_{\theta} \mathcal{K}_{\text{FIC}}(\theta_0) \cdot I^{-1}(\theta_0) \cdot \partial_{\theta} \mathcal{K}_{\text{FIC}}(\theta_0).
\]

Since the information matrix \( I \) scales proportional to \( N \), the variance is suppressed in non-singular regions of parameter space. But even in a singular region, variance can be low. In the singular case of \( M_1 \) in Fig. 1, the variance of \( \delta \theta^1 \) is large because of singularity and \( \mathcal{K}_{\text{FIC}}(\theta) \) is strongly coordinate dependent. But \( \mathcal{K}_{\text{FIC}}(\theta) \) is only a function of \( \theta^2 \) and the FIC complexity variance remains small. Of course, for parameter invariant complexities of the common type described in Appendix B 1, \( \partial_{\theta} \mathcal{K}_{\text{FIC}} = 0 \) everywhere, and the complexity variance will be zero.

In contrast to the empirical bootstrap, FIC is immediately applicable to structured data. Strongly structured data, such as time series, are not amenable to k-fold division for cross validation or resampling for bootstrap estimation, although more complicated methods, which often rely on stationarity assumptions, have been proposed [29, 30].

3. Parametric Bootstrap and FIC

Since our method samples from the fitted distribution, FIC has more in common with the parametric bootstrap. The five empirical bootstrap variants studied in [28] each have corresponding parametric forms where \( q(\cdot|\hat{\theta}_X) \) plays the role of \( p^* \). Two of these forms are examined in [31], in the context of mixed effect linear regression and, in particular, the parametric form of Eq. B6 was found to behave well. Like FIC, these parametric bootstrap methods are applicable to structured data. Unlike FIC, Eq. B6 relies on the symmetry in the training and generalization error, hence the factor of 2. This symmetry is only approximate for finite sample size.

The explicit sample (\( X \)) validation in Eq. B6 gives bootstrap methods a potential advantage in that they require only one expectation with respect to simulated data, whereas FIC requires two. This can reduce the computational cost relative to the proposed method.

We consider the methods reviewed in [31] to be complementary to FIC, and advocate their use when the double expectation in FIC proves to be too computationally demanding. However, the lack of direct data dependence in the FIC complexity has several advantages: i.) the proposed method will often have less variance than bootstrap methods, ii.) direct evaluation of the expectation integrals makes analytic expressions and approximations feasible. iii.) when analytic expressions are not available, FIC can still be more computationally attractive because the complexities for various models can be calculated once and compiled in look-up tables [11].

4. Summary

FIC has a practical role in extending AIC to singular and sloppy modeling situations and can be viewed in relation to several existing statistical methods: i.) the direct application of frequentist principles toward information-based inference (Sec. III), ii.) a unifying framework for a class of corrections to AIC, generalizing these corrections to cases which are parameter dependent (Sec. B 1) iii.) an important addition to the catalogue of bootstrap methods (Sec. B 2) iv.) an extension of AIC to learning algorithms beyond that of the method of maximum likelihood (Sec. A 1).