

## Novel Tools and Methods

# Navigating the Statistical Minefield of Model Selection and Clustering in Neuroscience

Bálint Király,<sup>1,2</sup> and Balázs Hangya<sup>1</sup><https://doi.org/10.1523/ENEURO.0066-22.2022>

<sup>1</sup>Lendület Laboratory of Systems Neuroscience, Institute of Experimental Medicine, H-1083, Budapest, Hungary and  
<sup>2</sup>Department of Biological Physics, Eötvös Loránd University, H-1083, Budapest, Hungary

## Abstract

Model selection is often implicit: when performing an ANOVA, one assumes that the normal distribution is a good model of the data; fitting a tuning curve implies that an additive and a multiplicative scaler describes the behavior of the neuron; even calculating an average implicitly assumes that the data were sampled from a distribution that has a finite first statistical moment: the mean. Model selection may be explicit, when the aim is to test whether one model provides a better description of the data than a competing one. As a special case, clustering algorithms identify groups with similar properties within the data. They are widely used from spike sorting to cell type identification to gene expression analysis. We discuss model selection and clustering techniques from a statistician's point of view, revealing the assumptions behind, and the logic that governs the various approaches. We also showcase important neuroscience applications and provide suggestions how neuroscientists could put model selection algorithms to best use as well as what mistakes should be avoided.

**Key words:** Bayes; bootstrap; clustering; cross-validation; information criterion; resampling

## Significance Statement

As neuroscience is becoming increasingly quantitative with “big data” approaches gaining a firm foothold in neurophysiology, neurogenetics and animal behavior, proper statistics for neuroscience is critically important. Nevertheless, probability theory and statistics is a dynamically evolving branch of mathematics; therefore, frequent cross-fertilization is required between probability theory and neuroscience. Statistical model selection, either implicit or explicit, is an integral part of data analysis in many neuroscience studies. Here, we review available and upcoming methods for statistical model selection, with an enhanced focus on cluster analysis as a special case and provide neuroscience examples for their application.

## Introduction

“All models are wrong, but some are useful (George Box, 1979).”

Broadly speaking, model selection encompasses all knowledge and assumptions about the underlying distributions for an observed sample. These assumptions have strong relevance to subsequent statistical decisions, like how to estimate central tendencies (e.g., mean, median), which statistical test to employ, how to describe and visualize the data. Arguably, the model selection trap most commonly walked into is performing it without noticing it. For instance, pitfalls of machine

learning-inspired dimensionality reduction approaches have been demonstrated recently (Chari et al., 2021), suggesting that it may be useful to treat the choice among such algorithms as a model selection problem, asking which approach represents the important aspects of the data most faithfully (mine #1).

In simple cases, beaten paths are of help. For instance, limit theorems explain the laws that determine asymptotic distributions of the sums of independent random variables; as a result, normal distribution may be assumed for many types of data on the basis of the central limit theorem, or the Poisson process is a good model for “random” spiking of neurons, because of the Poisson limit theorem.

Received February 9, 2022; accepted June 22, 2022.

The authors declare no competing financial interests.

Author contributions: B.K. and B.H. wrote the paper.

Nevertheless, we suggest that conscious evaluation of such assumptions is not a futile exercise even in seemingly simple situations.

In less straightforward scenarios, one might have to resort to a statistical model selection approach, that is, to choose from possible models underlying a set of observations based on statistical principles (Konishi and Kitagawa, 2008). However, this is an inherently hard mathematical problem: the “true” model may or may not be among the investigated choices, and it is far from trivial to assess whether a model is better than a competing one. The probability of the observations is evaluated in face of a given underlying model, called the “likelihood,” providing a goodness-of-fit (GOF) measure with firm probability theory basis (for recent review on Bayesian model comparison, see Keysers et al., 2020). However, models of increasing complexity typically fit better (mine #2). This is easily illustrated by polynomial fitting: if we intend to fit  $n$  data points, a polynomial of degree  $n-1$  (or higher) will provide a perfect fit. However, this is a poor argument to propose it as the “true” or even the “best” model. More likely, the data may be generated from a distribution with some level of stochasticity, better captured by a lower degree model.

Problems of this sort are frequently encountered in neuroscience. Generally, the GOF is discounted by measures of complexity; however, there is no single recipe for this equation, and as often happens, the multitude of proposed solutions demonstrates that none of them is perfect (Konishi and Kitagawa, 2008; James et al., 2013). While covering all important aspects of model selection would fill volumes, we focus on model selection problems especially relevant for neuroscience, emphasizing dos and don'ts, in the following (for the list of common “mines” of model selection discussed, see Table 1).

## Model Selection Based on Akaike Information Criterion

Is it one bump or two bumps in my plot? More formally, does the mixture of two Gaussians provide a better model than a single one? Is the dependence between the measurements linear, logarithmic, exponential or best described by a quadratic equation? At this point,

This work was supported by the “Lendület” Program of the Hungarian Academy of Sciences (LP2015-2/2015), the Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Grant K135561, the European Research Council Starting Grant no. 715043 and the Artificial Intelligence National Laboratory Programme of the Ministry for Innovation and Technology (to B.H.), and the ÚNKP-21-3 New National Excellence Program of the Ministry for Innovation and Technology (to B.K.).

Acknowledgments: We thank Dr. Sergio Martínez-Bellver, Dr. Tamás Móri, Dr. Gergő Orbán, Dr. Zoltán Somogyvári, and Dr. Balázs Ujfalussy for helpful comments and discussion on this manuscript.

Correspondence should be addressed to Balázs Hangya at [hangya.balazs@koki.hu](mailto:hangya.balazs@koki.hu).

<https://doi.org/10.1523/ENEURO.0066-22.2022>

Copyright © 2022 Király and Hangya

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

we hit a roadblock: how to arbitrate between models of disparate complexity?

The most often-used statistical tools for model selection in neuroscience are so-called information criteria, stemming from the maximum likelihood concept (Akaike, 1969, 1973; Banks and Joyner, 2017). Generalized from autoregressive (AR) models, Akaike introduced “an information criterion” (AIC; known today as Akaike information criterion) to compare statistical models of different complexity (Akaike, 1974). The AIC stands on solid statistical basis, rooted in the Kullback–Leibler divergence (KL) of information theory (Seghouane and Amari, 2007; Seghouane, 2010). The KL divergence quantifies the difference of the true distribution of the data compared with that derived from the tested model. In other words, it can be thought of as the “information loss” (in bits) or the “coding penalty” associated with the imperfect approximation of the true distribution. If we can measure an unbiased empirical distribution, like the frequency of heads and tails when tossing a coin, in the limit of infinite coin flipping, its KL divergence from the true distribution, 0.5 probability for each outcome for a fair coin, will tend to zero (Shlens, 2014; Cavanaugh and Neath, 2019). Formally, the KL divergence of the distribution  $P$  from the distribution (model)  $Q$  is defined by

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

Important to the derivation of AIC, the KL divergence can be decomposed into entropy (information content) of  $P$ , denoted by  $H(P)$  and cross-entropy of  $P$  and  $Q$ , denoted by  $H(P, Q)$ :

$$\begin{aligned} D_{KL}(P \parallel Q) &= \int_{-\infty}^{\infty} p(x) \log p(x) dx - \int_{-\infty}^{\infty} p(x) \log q(x) dx \\ &= H(P, Q) - H(P). \end{aligned}$$

The model that minimizes this quantity (minimal AIC estimator; MAICE) is accepted (Akaike, 1974, 1978). Thus, AIC relies on comparing competing models by the difference of their KL divergencies with respect to the “true model” (here denoted by  $P$ ). It is easy to see that this optimization problem only depends on the cross-entropy term, as the entropy of  $P$  is cancelled in the difference. The cross-entropy is mathematically tractable and relatively easy to estimate (Rao and Nayak, 1985); however, the commonly used maximum likelihood estimation (MLE) is not unbiased in this case, hence it has to be corrected by subtracting the estimated bias. The core of deriving AIC is the bias estimation procedure by approximating the cross-entropy with its Taylor series to the quadratic term, relying also on the central limit theorem and the strong law of large numbers (Akaike, 1973; Cavanaugh, 1997; Konishi and Kitagawa, 2008). This results in the formal definition of AIC as follows:

**Table 1: List of common “mines” of model selection and clustering discussed in the paper**

	Issue	Suggestion	Example
Mine #1	Selecting models without noticing it	Be aware of the assumptions behind analysis methods; treat the choice among different algorithms as a model selection problem	Chari et al. (2021)
Mine #2	Overfitting with overly complex models	Use statistical model selection tools which penalize too many parameters	Polynomial fitting
Mine #3	Selecting from a pool of poorly fitting models might lead to false confidence	Simulate data from each of the tested models multiple times and test whether the real data are sufficient to distinguish across the competing models	Figure 1b
Mine #4	Different information criteria might favor different models	Consider the strengths and limitations of the different approaches (Table 2); simulated data can be used to test which model selection method is the most reliable for the given problem	Figure 1c (AIC favors overfitting), e (BIC chooses an oversimplified model), f; Evans (2019)
Mine #5	Model selection might be sensitive to parameters ignored by the tested models	Avoid model classes that are too restrictive to account for data heterogeneity	Chandrasekaran et al. (2018)
Mine #6	Cross-validation techniques are prone to overfitting	A data splitting approach was proposed by Genkin and Engel in which optimal model complexity is determined by calculating KL divergence	Genkin and Engel (2020)
Mine #7	Agglomerative hierarchical clustering is sensitive to outliers	Consider divisive methods	Figure 2c; Varshavsky et al. (2008)
Mine #8	K-means clustering might converge to local minima	Repeat several times from different starting centroid locations	Figure 2e, right
Mine #9	Number of clusters not known	Use the elbow method, gap statistics, or model selection approaches	Figure 2e, left

$$AIC = -2 \ln(\hat{L}) + 2k,$$

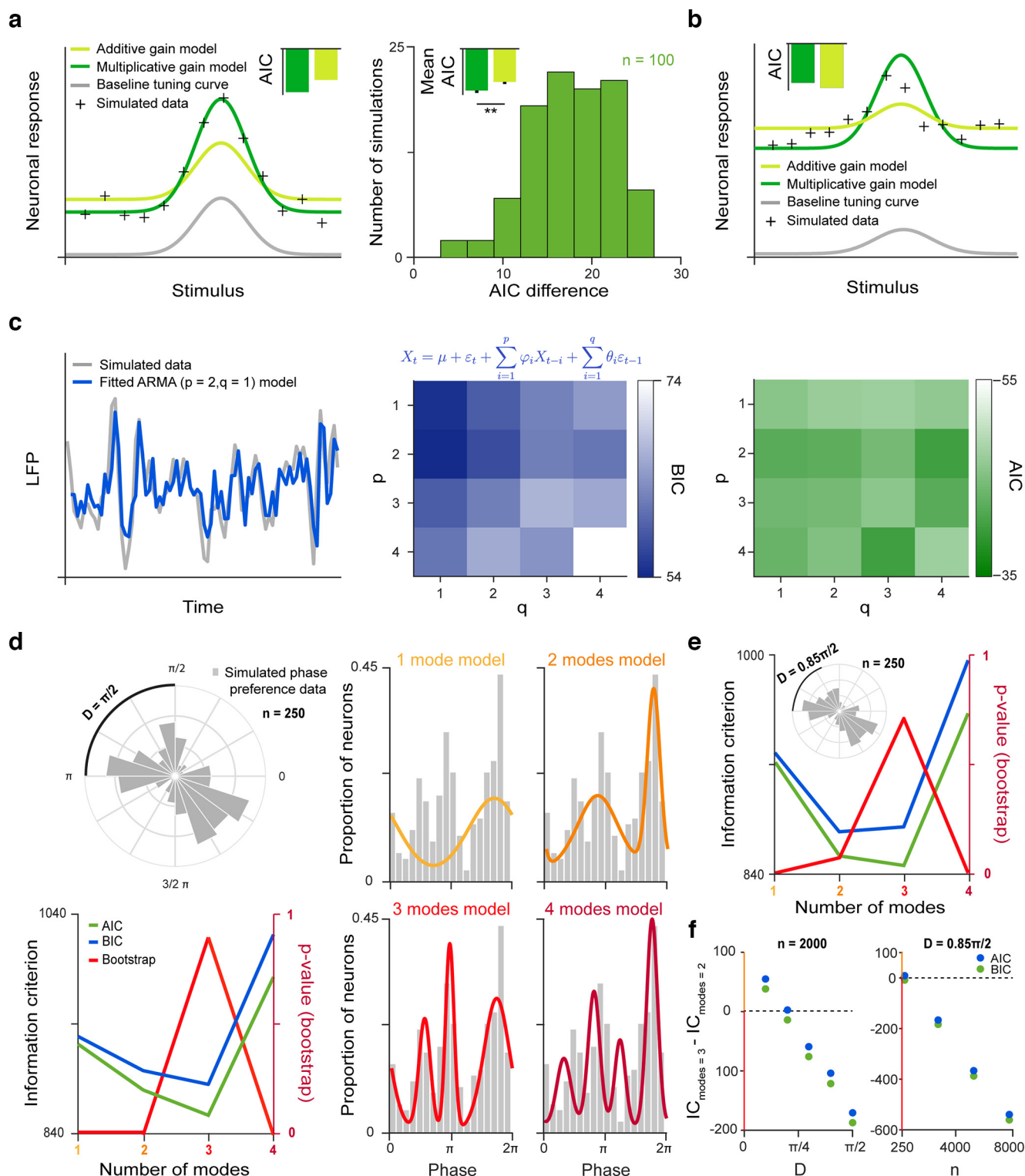
where  $\hat{L}$  is the maximum likelihood of the model and  $k$  is the number of free parameters (Akaike, 1973, 1974, 1978).

We demonstrate a use case of MAICE in Figure 1a by fitting different models to capture the essence of a tuning curve change. In sensory systems, neurons are often characterized by their tuning properties, which describe their responsiveness to external stimuli along different dimensions (Butts and Goldman, 2006). In the visual system, neurons of the primary visual cortex or the visual thalamus are investigated in terms of their orientation tuning, describing the angles of visual stimuli they prefer (Hubel and Wiesel, 1959; Atallah et al., 2012; Hillier et al., 2017). Auditory neurons can be categorized by their frequency-intensity tuning, revealing tonotopical organization of auditory cortices (Kilgard and Merzenich, 1998; Froemke et al., 2007; Hromádka et al., 2008). More broadly, “tuning” refers to the external or internal variables that drive neuronal firing, e.g., the tuning of hippocampal pyramidal neurons can be defined in terms of physical Euclidean coordinates (Muller et al., 1987; O’Keefe, 1993; Dupret et al., 2010) and the tuning within face patches of the inferotemporal cortex is defined by shape and appearance features of standardized images of faces as stimuli (Chang and Tsao, 2017). Tuning curves change during learning revealing rules of plasticity (Kilgard and Merzenich, 1998; Froemke et al., 2007), as a function of attentional modulation (Maunsell and Treue, 2006; Disney et al., 2007; Lee and Maunsell, 2009; Krueger and Disney, 2019) or in response to optogenetic modulation of different neuron types (Adesnik et al., 2012; Atallah et al., 2012; S.H.

Lee et al., 2012; Pi et al., 2013; Wilson et al., 2012). Therefore, they have become important tools to help understand cortical information processing (Seriès et al., 2004; Butts and Goldman, 2006; Lee and Maunsell, 2009). Tuning curve changes are often captured by additive or multiplicative gain modulation models, in which selecting the best linear gain model occurs as a typical model selection problem (Atallah et al., 2012; Pi et al., 2013; Hangya et al., 2014). In our example, we simulated a gain change by multiplying a baseline tuning curve with a scalar factor. The MAICE method correctly indicated that the simulated data were better explained by a multiplicative gain model. Moreover, repeating the simulation a hundred times revealed statistical superiority of the correct model.

### Limitations of AIC

Although MAICE is a strong tool, some caution should be raised. First, when the true model is not among the tested ones, comparing poorly fitting models with AIC may lead to a false confidence in a model that is marginally better than some others. In Figure 1b, we simulated a tuning curve by combining an additive and a multiplicative gain; both the purely additive and purely multiplicative model showed poor fit, still, AIC chose a “winner” (mine #3). A related caveat is that AIC, in itself, does not provide error bars; in other words, one might not have enough data to draw conclusions on two similar models, despite AIC being nominally smaller for one of them. When this might be the case, it is recommended to statistically test whether the data are sufficient to distinguish across the competing models. This can be done by simulating data points from each of the tested models with a sample size



**Figure 1.** Examples of model selection problems in neuroscience. **a**, Using MAICE to choose between competing models. Left, We used a bell curve to simulate neural responses as a function of stimulus features, generally referred to as a “tuning curve” (gray) and used a multiplicative gain model ( $y = ax(s) + \varepsilon_i$ , where  $y = x(s)$  is the baseline tuning curve,  $a$  is a scalar and  $\varepsilon_i$  is a Gaussian noise term) to simulate a tuning curve change (black crosses). Next, an additive (light green) and a multiplicative model (dark green) was fitted on the simulated data. Smaller AIC value indicated that the multiplicative model fitted better (inset), as expected based on the simulation. Right, We performed  $n = 100$  data simulations and calculated the difference between the AIC values of the competing models. The histogram (and the mean values in the inset) demonstrates that in every case the multiplicative model outperformed the additive one. Error bars show standard deviation from the mean;  $**p < 0.01$ ; two-sided bootstrap test. **b**, A gain model with both



*continued*

additive and multiplicative components ( $y = ax(s) + b + \varepsilon_i$ , where  $y = x(s)$  is the baseline tuning curve,  $a$  and  $b$  are scalars and  $\varepsilon_i$  is a Gaussian noise term) was used to simulate a tuning curve change (black crosses) relative to a baseline tuning curve (gray). Next, an additive (light green) and a multiplicative model (dark green) was fitted on the simulated data. While the best fit curves visible deviated from the simulated tuning curve, a smaller AIC value indicated that the additive model fitted somewhat better. **c**, Using BIC to choose the best fitting ARMA model. Left, An ARMA ( $p = 2, q = 1$ ) process was used to simulate LFP time series data (gray), where  $p$  denotes the order of the AR and  $q$  the order of the moving average component. Next, we fitted ARMA models on the data with different  $p$  and  $q$  values in the range 1–4. Blue trace shows the predicted data based on the best fitting  $p = 2, q = 1$  model. Middle, BIC was calculated for each model, and the model with the lowest value ( $p = 2, q = 1$ ) was chosen. Top, ARMA ( $p, q$ ) model, where  $\phi_i$  are AR parameters,  $\theta_i$  are moving average parameters,  $\varepsilon_i$  are Gaussian noise terms, and  $\mu$  is constant. Right, AIC was calculated for each model. AIC favored the more complex  $p = 2, q = 4$  model over the expected  $p = 2, q = 1$  model. **d**, Demonstration of the use of information criteria and the parametric bootstrap technique for choosing the number of modes in a distribution. We simulated phase preference data of neuronal firing ( $n = 250$ ) referenced to an LFP oscillation as the combination of three wrapped normal distributions (top left,  $D = \pi/2$  refers to the phase difference between the mean of the two closest wrapped normal distributions). Next, mixture models of 1–4 von Mises distributions (circular analog of the normal distribution that closely approximates the wrapped normal distribution) was fitted on the distributions (right) using an expectation maximization algorithm for circular data (Czurkó et al., 2011). Minimal AIC and BIC, as well as maximal bootstrap  $p$  value correctly suggest that the model of three modes is the best fitting one. BIC is penalizing the higher mode models more than AIC (bottom left). **e**, Same as panel **d** (left), but with  $D = 0.85 \pi/2$  phase difference parameter. While minimal AIC and maximal bootstrap  $p$  value still suggest that the model of three modes is the best fitting one, BIC favored a simpler model with two modes. **f**, Information criterion difference between the two-mode and three-mode models as a function of the phase difference parameter  $D$  (left,  $n = 2000$ ) and the sample size ( $n$ , at  $D = 0.85 \pi/2$ ). Color coding of the  $y$ -axis reflects the favored model (orange: 2 modes; red: 3 modes). If the modes are well-separated and the sample size is sufficient, both AIC and BIC choose the correct three-mode model, while they can fail for low sample size or less distinguishable modes. At parameters where the information criterion differences are closer to zero, BIC might favor the two-mode model while AIC might correctly identify three modes.

of the original dataset multiple times, and evaluating whether the difference in AIC is consistent across those simulations (Fig. 1a).

Second, as noted above, the heart of the derivation of AIC is a correction that eliminates the bias introduced by the specific mode of MLE employed for the formula. However, this bias estimation only works asymptotically; it is therefore not recommended to use AIC on small sample sizes. A “corrected” AIC (AICc) was proposed for these cases; it is worth noting, however, that AICc relies on more specific assumptions on the distribution of the underlying data (Cavanaugh, 1997; Brewer et al., 2016).

Third, AIC is based on MLE, which is widely used and has strongly established, favorable statistical properties. Nevertheless, it may yield unstable parameter estimates for complex models, and therefore, a range of “regularized” (or “penalized”) MLE methods are available (Konishi and Kitagawa, 2008; Jang et al., 2016; Chamroukhi and Huynh, 2019). Since AIC formulation does not provide a straightforward way to incorporate these more robust MLE techniques, the generalized information criterion (GIC) has been introduced, which provides a recipe for constructing novel information criteria (Konishi and Kitagawa, 2008). However, it will take further theoretical work to determine how to derive suitable information criteria for neuroscience based on GIC.

## Bayesian Model Selection

The Bayesian approach to model selection is rooted in calculating the posterior probability of a candidate model, that is, how probable a given model is, provided the observed data (Kass and Raftery, 1995; Wasserman, 2000). Then, competing models are compared by their posterior probabilities, and the one that maximizes this quantity is selected.

In practical applications, the Bayes factor is often used, comparing the relative strength of evidence for two models  $M_1$  and  $M_2$ :

$$B = \frac{\Pr(\text{Data}|M_1)}{\Pr(\text{Data}|M_2)}.$$

Of note, the Bayes factor becomes the ratio of posterior probabilities of the models in case their prior probabilities are the same (the case of “uniform priors”), used for instance in hypothesis testing, where an equally plausible null and alternate hypothesis are compared (Keysers et al., 2020). Although the interpretation of actual Bayes factor values remains somewhat subjective, Jeffrey suggested in his classic work that a value over 10 should be regarded as “strong evidence” (Jeffrey, 1961), which notion is still generally accepted (Keysers et al., 2020).

In 1978, Schwarz introduced the Bayesian information criterion (BIC; also called Schwarz–Bayesian information criterion; Schwarz, 1978), formally defined by the following:

$$BIC = -2 \ln(\hat{L}) + k \ln(n),$$

where  $\hat{L}$  is the maximum likelihood of the model (like in AIC),  $n$  is the sample size, and  $k$  is the number of free parameters. It has been shown that the BIC is an approximation of the logarithm of the Bayes factor; therefore, the BIC provides an easy-to-use tool for Bayesian model selection (Kass and Raftery, 1995).

To demonstrate a good use of BIC in a neuroscience context, we first go back to our example of polynomial fitting. MLE estimates trivially indicate better fits at higher degrees, because the class of  $n$  degree polynomials include all  $<n$  degree ones. As a counterbalance, information criteria penalize high number of parameters. AIC has

a penalty of  $2k$  (where  $k$  is the number of parameters), whereas BIC has a term  $\ln(n)k$  (where  $n$  is the sample size), thus BIC implies a stronger penalty and hence tends to select simpler models (Brewer et al., 2016; Fig. 1c–e). Relatedly, AIC is often pictured as a method for selecting models for good “predictive accuracy” by penalizing sensitivity to spurious features of the data also known as overfitting, whereas BIC attempts to provide a good description of the fitted data as it penalizes for model complexity more broadly (Kass and Raftery, 1995; Evans, 2019; but also see Rouder and Morey, 2019). Mathematically, AIC is “asymptotically efficient,” minimizing prediction error as sample size tends to infinity, while BIC is “asymptotically consistent,” selecting the correct model, if it is in the tested pool, as sample size increases. They can also be seen as representing two different world views: in a complex world where the true generating model is unknown and may be unknowable, as suggested by the quote by Box we cited, one must resort to efficiency, where AIC wins. When one believes that a relatively simple model, included in the set of tested candidates, generates the data, BIC can pick the true model, while AIC does not come with such guarantees (Aho et al., 2014; Chandrasekaran et al., 2018).

A good use case of BIC is determining the optimal model order, for instance, when fitting AR models. The order of AR models determines the time scale at which previous information influences the “present” of a time series signal. In this case, competing models tend to be simple and we have an a priori bias toward smaller, more parsimonious models (Fig. 1c). Such analyses are typical in predictive time series analysis of EEG and fMRI traces (Muthuswamy and Thakor, 1998; Baajour et al., 2020). For instance, pathologic synchrony during epileptic activity can be detected and measured by analyzing the residual covariance matrix of an AR model fitted on multichannel scalp EEG recordings (Franaszczuk and Bergey, 1999). Similarly, cross-area interactions can also be quantified by multivariate AR models in fMRI recordings (Harrison et al., 2003; Stephan et al., 2004; Ting et al., 2015). Epileptic patients also show interictal spikes in EEG recordings, which can be efficiently detected using AR models estimated with a Kalman filter (Oikonomou et al., 2007). Figure 1c demonstrates the application of BIC to choose the best fitting AR-moving-average (ARMA) model for simulated local field potential (LFP) data. As indicated above, in such problems AIC may not pick the true model; indeed, in our example AIC favored a more complex model (mine #4; Fig. 1c, right).

## More Information Criteria

We may view the information criteria as methods for estimating the correct number of model parameters by finding their minimum. The BIC has the advantage over AIC that in the infinite limit of sample size, it yields a parameter estimate that converges to the true number of parameters with a probability of 1, called a “strongly consistent” estimate in statistics. Another strongly consistent information criterion was introduced by Hannan and Quinn (Hannan and Quinn, 1979), inheriting its favorable properties from the law of the iterated logarithm (Erdős, 1942):

$$HQC = -2 \ln(\hat{L}) + 2k \ln(\ln(n)).$$

The penalty term grows very slowly as a function of sample size, which was suggested to lend the Hannan–Quinn Information Criterion better convergence properties compared with the BIC (Miche and Lendasse, 2009). Therefore, it is often used for determining the order of AR models (Sin and White, 1996; Miche and Lendasse, 2009), suggesting that it may have a yet unexploited place in the EEG and fMRI data analysis armament.

The deviance information criterion (DIC) is an extension of AIC, penalizing similarly for model parameters, but applying a different GOF measure, defined as the likelihood of the data averaged over the entire posterior distribution (Spiegelhalter et al., 2002; Celeux et al., 2006; Evans, 2019). By departing from the MLE-based GOF approach, it gained popularity in Bayesian model selection, when dealing with cases where maximum likelihood estimation is difficult (Celeux et al., 2006; Evans, 2019; Y. Li et al., 2020). While DIC assumes approximate normality of the posterior distribution (Spiegelhalter et al., 2014), Watanabe proposed a “widely applicable” (or Watanabe–Akaike) information criterion (WAIC) that does not rely on such assumptions (Watanabe, 2010a,b, 2013). Newest in this family, the “leave one out” information criterion (LOOIC) is similar to WAIC (Watanabe, 2010b; Gelman et al., 2014; Vehtari et al., 2017; Yong, 2018), but it has been proposed to yield more robust results in the finite case with weak priors or influential observations (Vehtari et al., 2017). Although these measures incorporate Bayesian notions, they can still be interpreted in terms of predictive accuracy, thus being advanced alternatives of AIC (Evans, 2019). Watanabe has made an attempt to generalize BIC as well, which resulted in the “widely applicable Bayesian” information criterion (WBIC) that seeks for the true model instead of minimizing predictive loss (Watanabe, 2013). Neural dynamics is usually best modelled by latent variable models, assuming a set of interacting hidden and observable variables (Sahani, 1999; Churchland et al., 2012; Vértes and Sahani, 2018; Pei et al., 2021), and doubly stochastic processes, where the dynamics is described by a random point process with varying intensity (Cunningham et al., 2008; Latimer et al., 2015). In these cases, one needs to apply the abovementioned information criteria relying on Bayesian GOF estimations (Latimer et al., 2015); therefore, we expect that these novel approaches will soon gain popularity in neuronal modeling and complex data analyses.

It is still debated among statisticians which information criterion is better and when (Konishi and Kitagawa, 2008; Mainassara and Kokonendji, 2016). Since choosing the penalizing term will remain somewhat arbitrary, there likely will not be “one information criterion to rule them all.” Indeed, it is possible that different information criteria will favor different models, without a clear argument on which particular criterion suits the statistical problem at hand best (mine #4). For instance, Evans conducted a systematic comparison of a number of information criteria for a specific class of evidence accumulation models of decision-making processes and found that while model selection approaches typically agreed when effect sizes were moderate to large, they could diverge in their

conclusions for small or nonexistent effects (Evans, 2019). He concluded that one should opt for “predictive accuracy” approaches like AIC when the primary goal is to avoid overfitting and thus select a model with strong predictive value, whereas BIC performs better if the goal is to provide the best account of the present data (Shmueli, 2010; Evans, 2019). Going one step further, one might adopt a simulation approach to test which model selection approach is the most reliable for the problem at hand, much the same as in the Evans study.

When complex systems, like those that determine the exact firing activity of neurons, are considered, it is unlikely that our models will capture all aspects of the true generating model. However, model selection approaches will always announce a winner, which raises a set of issues (Chandrasekaran et al., 2018). First, it is conceivable that all of the tested models fall far from the generating process, in which case model selection will yield a misleading conclusion about the data (mine #3). Second, model selection may be sensitive to parameters of the generating model not captured by the tested models. In such cases, model selection will suggest a model that is closer to the data in statistical or information theoretical terms, but not necessarily conceptually (mine #5). This is detailed in an elegant paper revealing model selection pitfalls when arbitrating between ramp-like and step-like changes in firing rates of single cortical neurons (Chandrasekaran et al., 2018). As a suggestion, one should take multiple close looks at the data, and avoid model classes that are too restrictive to account for data heterogeneity (Chandrasekaran et al., 2018; Genkin and Engel, 2020).

As a take-home message, information criteria are strong tools to contrast competing models, but researchers should always ask “Is my data sufficient and appropriate to discriminate these models?” (for details, see above, Limitations of AIC), and if the answer is yes, cautiously conclude that “this particular information criterion provides a statistical argument for model A describing the data better than model B.”

## Model Selection Using Resampling Techniques

Resampling techniques are strong tools of modern statistics with firm mathematical foundations while having minimal assumptions on the statistics of underlying data. On the flip side, they require substantial, sometimes prohibitively large, CPU-power.

In cross-validation methods, originating from machine learning, the data are split into a “training” and a “test” set. A model can be validated by fitting on the training set and obtaining GOF statistics on the test set. In neuroscience, the leave-one-out cross-validation algorithm is often applied: the model is fitted on  $n-1$  data points, and tested on the remaining one, repeated for all data points as the “test set” (Kohavi, 1995; Browne, 2000; Hastie et al., 2009; J.R. Cohen et al., 2010). A generalization is the leave-p-out; however, more CPU-intensive with increasing  $p$ . To reduce CPU-load, the  $k$ -fold cross-validation may be applied: instead of testing all combinations of

$p$ -sized subsamples, the data are split into  $k$  groups to obtain the GOF distribution. Of note, the law of total probability, characterizing the relationship of conditional and marginal probabilities, reveals a deep link between cross-validation techniques and the Bayesian concept of likelihood. Specifically, decomposing the marginalized log likelihood function by the chain rule provides a formula equivalent to the sum of leave-p-out Bayesian cross-validation errors (Sadler et al., 2014; Fong and Holmes, 2020).

It should nevertheless be noted that cross-validation techniques have some often overlooked, unfavorable statistical properties. Namely, they are prone to overfitting, giving undue credit to more complex models (Gronau and Wagenmakers, 2019). Indeed, when complex, flexible models are applied to broadly capture data heterogeneity, cross-validation techniques at realistic data amounts may not be able to prevent overfitting (mine #6). To overcome this, a novel data splitting approach was proposed recently, in which the data are divided into two halves, and the optimal model complexity is determined by calculating KL divergence between the distributions corresponding to models of the same complexity fitted on the two datasets (Genkin and Engel, 2020). A caveat of this proposal is that the KL divergence rises more or less monotonically with model complexity; thus, an empirical threshold is suggested to determine the “optimal point” the robustness of which has to be determined by future studies.

Neuroscience applications include a wide range of classification problems, from linking fMRI or fNIRS data to human behavior (J.R. Cohen et al., 2010; Jiang et al., 2012) to categorizing stimulus responses of cortical neurons (de Vries et al., 2020). Recently, cross-validation techniques have gained additional momentum as machine learning techniques revitalize many areas of neuroscience (Savage, 2019; Yang and Wang, 2020), since they are considered the first choice model selection tools when fitting artificial intelligence (AI) models. In this regard, it is important to highlight the attempts to automate classification of neurons based on their morphologic (m-types), electrophysiological (e-types) and transcriptomic (t-types) characteristics (Armañanzas and Ascoli, 2015; Saunders et al., 2018; Gouwens et al., 2019, 2020; Que et al., 2021).

Of the resampling approaches, parametric bootstrap is a particularly useful technique, often overlooked in neuroscience. When testing models that can be described with a relatively small number of parameters (e.g., a mixture of Gaussians), one can generate a bootstrap set of simulated data from these models, and use a GOF measure to describe the relationship of the model and the simulated data. A bootstrap distribution of such GOFs can then be used to estimate the probability of the original data violating the tested model (Fisher, 1993; Czurkó et al., 2011). Thus, unlike information criteria only providing a relative score, one obtains a  $p$  value. The model with the highest  $p$  value (least rejected model) wins. A practical application of parametric bootstrap is arbitrating the number of modes in an empirical distribution, using a mixture of unimodal distributions described by few parameters (e.g.,



**Table 2: Advantages and limitations of model selection and clustering algorithms**

Method	Advantages	Limitations	Suggestions
<b>Statistical model selection</b>			
Akaike information criterion (AIC)	Strong mathematical basis (KL-divergence)	May lead to false confidence in marginally better models	If critical, perform simulations to ascertain true differences among the tested models
	Easy to calculate	Difficult to test whether differences are significant	
Bayesian information criterion (BIC)	Suitable for comparing models of different complexity	Not suitable for low sample sizes	Consider AICc if its assumptions are met
	Strong mathematical basis (Bayesian statistics)	Asymptotic properties may not hold for complex (multi-parameter) models	BIC is more recommended for simpler models, especially when overfitting is a concern, e.g., deciding the order of an AR process
Resampling methods	Easy to calculate	Difficult to test whether differences are significant	Consider simulations, as for AIC
	No assumptions on data distributions	CPU-intensive	
Clustering	Provides a <i>p</i> value for each tested model	Does not always converge to the true model (statistically inconsistent in the M-closed case)	Parametric bootstrap and cross-validation are often the best choice for testing models with few parameters
Hierarchical clustering, agglomerative	Simple	CPU-intensive for large datasets	With careful consideration of choosing similarity measure, clustering rule and other parameters, the flexibility of hierarchical clustering can be used to its advantage; test the robustness of the results by exploring the parameter space
	Easy to interpret	Sensitive to outliers and choices of algorithms and parameters	
Hierarchical clustering, divisive	Includes more robust and CPU-efficient options	Sensitive to choices of algorithms and parameters	Ideal choice if number of expected clusters is known; explore robustness of results by starting the algorithm from different sets of centroids
	K-means clustering	Requires a priori estimate of number of clusters	
	Does not rely on many parameters	May converge to local minima and not find the global optimum	

The last column provides suggestions on how to use.

Gaussians). For instance, hippocampal neurons are often characterized by the systematic relationship between their action potentials and the dominant ongoing local population activity, the theta rhythm (Buzsáki, 2002; Klausberger and Somogyi, 2008; Czurkó et al., 2011; Buzsáki and Moser, 2013). Neurons are active at multiple phases of this oscillation (Klausberger and Somogyi, 2008; Czurkó et al., 2011), but is the observed circular phase histogram of hippocampal activity truly multimodal? We demonstrate the power of parametric bootstrap approaches on this example in Figure 1*d*. Of note, while AIC and BIC work well for sufficient sample sizes and well-separated modes, they can both fail for low sample sizes or if the modes of the generating model are less distinguishable (Fig. 1*e,f*).

Bootstrap techniques also differ from information criteria in their capabilities to evaluate standalone models, while AIC, BIC, and related methods can only perform comparison of competing models. As a middle ground, one may ask the question whether a given model is better than a minimalistic model that still captures selected features of the data. Bayesian decoding can be used to extract relevant features of the data in a model-free paradigm (Koyama et al., 2010; Kloosterman et al., 2014). In neuroscience, this approach may relate neuronal activity to external variables as a generalization of the

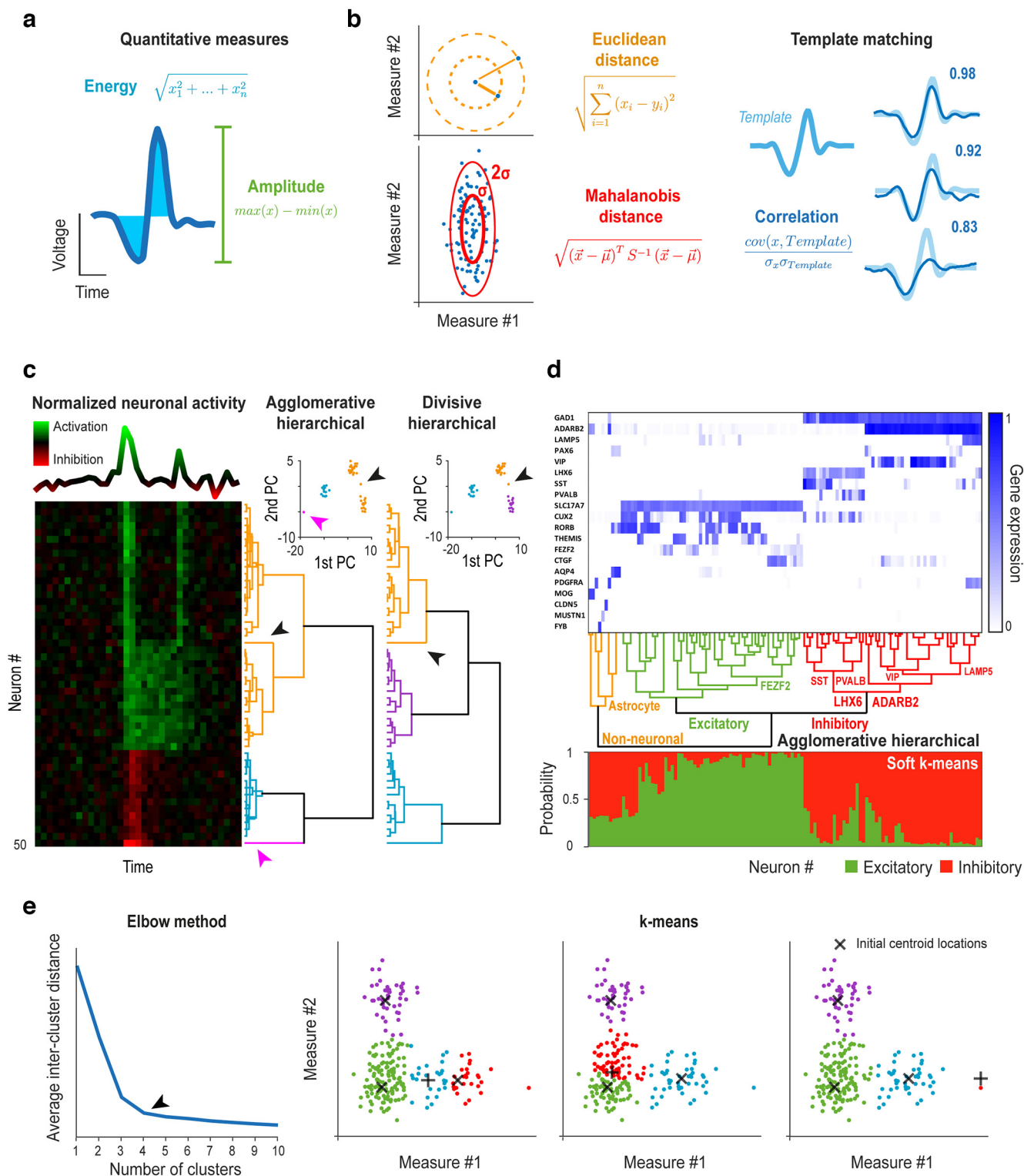
concept of tuning (see above; Okun et al., 2012; Kloosterman et al., 2014), or aim at understanding interdependences within populations of neurons (Okun et al., 2012; Bányai et al., 2019). Maximum entropy models (Tkačik et al., 2013) are powerful tools to generate “minimal” models with appropriate constraints, reviewed elsewhere (Savin and Tkačik, 2017). We provide an overview of the model selection techniques most commonly used in neuroscience in Table 2.

## Clustering Problems

Clustering problems form a special class of model selection that deserves attention because of its broad usefulness in neuroscience. Neuroscientists often aim to identify groups with similar properties: action potentials that likely belong to the same neuron based on similar spike shape (spike sorting; Quiroga, 2012; Yger et al., 2018; Magland et al., 2020), neurons that likely belong to the same cell type based on similar gene expression profiles (Saunders et al., 2018; Gouwens et al., 2020) and cells that represent the same behavioral variable based on similar response patterns in task-performing animals (J.Y. Cohen et al., 2012; Hangya et al., 2014).

Clustering problems typically require solving a series of model selection problems (Konishi and Kitagawa, 2008; James et al., 2013). First, multidimensional data should





**Figure 2.** Examples of clustering problems in neuroscience. **a**, Examples of simplified quantitative measures of waveforms often used in spike sorting. **b**, Examples of distance measures used for quantifying similarity between data points and clusters often used in spike sorting. The Mahalanobis distance normalizes the standard deviation across dimensions, while template matching approaches are based on waveform correlations with predefined waveform templates. **c**, Hierarchical clustering of simulated neuronal activity (peri-event time histograms) of  $n = 50$  neurons with 3 (seemingly) very well separated groups (left). First, principal component analysis (PCA) was used to reduce dimensionality of the time series data. Second, agglomerative and divisive hierarchical clustering was performed in the space spanned by the first two principal components (right). Agglomerative clustering separated a single outlier cell (magenta arrow) in an earlier step (second) than the three main clusters, in contrast with divisive clustering. The two methods

*continued*

also differed in the clustering of a cell similar to more than one main groups (black arrow). **d**, Clustering of human cells from multiple cortical areas based on RNA-sequencing data ( $n=50,281$  genes, publicly available at <https://portal.brain-map.org/atlas-and-data/rnaseq/human-multiple-cortical-areas-smart-seq>). Top, Trimmed mean expression of  $n=20$  marker genes. Middle, Agglomerative hierarchical clustering was performed based on the first 20 principal components, revealing the hierarchy of cell types (branches of the dendrogram were identified based on marker gene expression). Bottom, Soft K-means ( $k=2$ ) clustering was performed to assign probabilities to all cells of belonging to each of two main cell types, identified as excitatory and inhibitory based on marker gene expression. **e**, Spike sorting of simulated action potentials ( $n=221$ ) using K-means clustering. Left, We applied the elbow method on the average intercluster squared Euclidean distance of all points to find the optimal number of clusters ( $k=4$ , black arrow), based on 100 repetitions of K-means clustering for each  $k$  in the range of 1–10. Right, K-means clustering ( $k=4$ ) was performed with three different initial centroid locations (black crosses, 3 of which were kept at fix positions while one was changed), leading to surprisingly different clusters.

be modeled by simplified quantitative measures that capture important variance from the neuroscience point of view, like action potential amplitude for spike sorting (Schmitzer-Torbert et al., 2005; Fig. 2a). Second, “similarity” of data points needs to be defined by an appropriate distance measure, most often the Euclidean distance between points in the space defined by this model (Schmitzer-Torbert et al., 2005; Fig. 2b, left). However, this may not be straightforward, for example in time series analysis dealing with time-resolved membrane potential, spike trains, LFP, EEG, ECoG, or fMRI, where linear or information theory-based correlation measures (Fig. 2b, right), potentially combined with dimensionality reduction techniques like principal component analysis, may be considered (Fig. 2c,d; Goutte et al., 1999; J.Y. Cohen et al., 2012; Thirion et al., 2014). Third, alternatives to “hard/complete” clustering, allowing cluster overlaps or probabilistic cluster assignments may be contemplated (Fig. 2d; Yang, 1993; Goutte et al., 1999).

There are two fundamental approaches to create the clusters (James et al., 2013). In hierarchical clustering, clusters are defined by merging data-points bottom-up in agglomerative, or splitting groups top-down in divisive clustering, either way creating a dendrogram of clusters (Fig. 2c). Agglomerative clustering is a popular choice, owing to its simplicity and ease of interpretation (Ward, 1963). However, agglomerative techniques are computationally heavy for large datasets, and particularly sensitive to outliers, since local properties determine their amalgamation rules (mine #7). These pitfalls can be overcome by special divisive methods that reduce such noise sensitivity by taking the global data distribution into account and have the computationally efficient option to stop once the required number of clusters is reached (Varshavsky et al., 2008). We demonstrate this on a simulated example of hierarchical clustering of peri-event time histograms of averaged neuronal responses (J.Y. Cohen et al., 2012; Hangya et al., 2015; Sadacca et al., 2016; Takahashi et al., 2016). We represented the peri-event time histograms by their first and second principal components to reduce data dimensionality and compared the results of agglomerative and divisive hierarchical clustering in Figure 2c. Another neuroscience application of hierarchical clustering considers groups or types of neurons based on transcriptomic information (C.L. Li et al., 2016; Faure et al., 2020). We demonstrate agglomerative hierarchical clustering of human cortical cells based on publicly available

RNA-sequencing data (<https://portal.brain-map.org/atlas-and-data/rnaseq/human-multiple-cortical-areas-smart-seq>) in Figure 2d. In course of analyzing single cell transcriptomic data, the recently developed t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton, 2008; Kobak and Berens, 2019) is often applied for dimensionality reduction (Harris et al., 2018; Saunders et al., 2018; Kobak and Berens, 2019). Although t-SNE faithfully reflects local structure, i.e., within-cluster distances, it does not preserve global structure (between-cluster distances), which renders it suboptimal for clustering. In 2018, McInnes proposed a novel dimension reduction algorithm he coined uniform manifold approximation and projection (UMAP), which, by a better choice for the cost function, based on KL divergence in t-SNE and cross-entropy in UMAP, results in better preservation of global data structure (McInnes et al., 2018). Therefore, while both t-SNE and UMAP are excellent visualization tools for high-dimension datasets, UMAP is more recommended for subsequent cluster analysis (McInnes et al., 2018; Diaz-Papkovich et al., 2021). The advantage of using nonlinear dimensionality reduction algorithms like UMAP for solving neuroscience problems is demonstrated by a novel UMAP-based spike sorting approach that can successfully sort cerebellar Purkinje cell recordings, a notoriously hard problem due the high degree of variability of simple and complex Purkinje cell spikes (Sedaghat-Nejad et al., 2021).

In contrast to hierarchical algorithms, a number of arbitrary cluster centers are assigned in K-means clustering, updated repeatedly based on proximity of data points to these centroids (Hastie et al., 2009). With a priori information on the number of groups and well-separated clusters, K-means is fast, efficient, and does not rely on a large number of potentially ambiguous choices. Nevertheless, these assumptions often remain unmet, when the flexibility of hierarchical clustering (choice of similarity measure and clustering rules), used wisely and cautiously, may provide better results. Furthermore, hierarchical clustering is deterministic, unlike K-means, which depends on the initial choice of centroids and might converge to local minima that can give rise to incorrect interpretations (mine #8). To avoid this, it is recommended to repeat K-means clustering several times using different initial centroid positions. We showcase K-means clustering on a spike sorting example (Fig. 2e). We simulated action potentials and sorted them into clusters corresponding to putative single

neurons, mimicking a typical spike sorting problem in extracellular electrophysiology (Schmitzer-Torbert et al., 2005; Quiroga, 2012). Starting the algorithm from different centroid locations shows the sensitivity of K-means clustering to initialization (mine #8).

More robust results may be achieved by a relatively new technique called “spectral clustering” that combines K-means clustering with dimensionality reduction, however, at the price of losing the appealing simplicity of K-means clustering lauded above (Shi and Malik, 2000; Hirokawa et al., 2019). Additionally, fitting a mixture of Gaussians by expectation maximization (Jung et al., 2014) or other algorithms (Kántor et al., 2015) can be considered as an alternative to K-means clustering that allows operating with likelihood by proposing a statistical model (Fig. 1d).

An interesting algorithm based on physical properties of an inhomogeneous ferromagnetic model, called superparamagnetic clustering gained considerable popularity in neuroscience owing to its unsupervised nature that does not pose assumptions on the underlying data (Blatt et al., 1996; Domany, 1999). It has been used in a wide range of applications from spike sorting combined with wavelet spectral decomposition (Quiroga et al., 2004; Townsend et al., 2011) to morphologic classification of neurons (Zawadzki et al., 2010) to analyzing visual stimulus processing (Jasa et al., 2008).

In graph-like data structures where data points (“nodes”) are connected with links (“edges”), graph theory-based methods can be applied to detect clusters (“communities”) of the network. In such methods, a “modularity” measure is optimized that compares the link density inside versus outside the communities (Blondel et al., 2008). Lee and colleagues applied the graph-based Louvain community detection on spike waveforms of the macaque premotor cortex after nonlinear UMAP embedding (see above) and demonstrated the usefulness of this approach in revealing functional cell type diversities (Blondel et al., 2008; E.K. Lee et al., 2021). Of note, while the Louvain approach was developed to deal with extremely large graphs in a computationally efficient manner, its two-phase algorithm of finding high modularity partitions leaves the question open whether the order of considering the nodes throughout the algorithm can have a substantial effect on the results (Blondel et al., 2008). Nevertheless, Lee et al., showed that their approach resulted in stable clusters and outperformed Gaussian mixture model clustering applied on specific waveform features (E.K. Lee et al., 2021).

There is an important model selection problem often at the heart of clustering: how many clusters are there (mine #9)? A number of tools have been developed to aid this decision. The ratio of the between-cluster variance to the total variance monotonically increases as a function of the number of clusters, but typically flattens significantly at a point, called the “elbow” (Fig. 2e). The location of this bend is generally considered as an indicator of the appropriate number of clusters. A statistical approach to formalize this heuristic is the gap statistic (Tibshirani et al., 2001), based on comparing the total within-cluster variation with

its expected value under the null hypothesis of no clusters present in the data. The optimal number of clusters is the one that maximizes this difference (the “gap”; Tibshirani et al., 2001). The gap statistics has been employed in spike sorting (Nguyen et al., 2014) and other clustering problems in neuroscience (Ito et al., 2018; Gwo et al., 2019), including fMRI-based connectivity analyses (Hahamy et al., 2015). As an alternative, most model selection approaches discussed above, including information criteria and parametric bootstrap for different number of clusters as competing models may be recruited for clustering problems.

## Conclusion

We showcased widely used model selection and clustering approaches especially relevant to neuroscience problems, also pointing to promising “up-and-coming” methods. Nevertheless, an exhaustive overview would stretch beyond the limits of this review. Most importantly, we would like to stress that model selection is a scientific field on its own right and urge neuroscientists to take conscious decisions about selecting the appropriate techniques and parameters, very much the same way as deciding on experimental design. The bad news is there is no free lunch or rules of thumbs that solve it all; however, the overwhelming good news is that a whole, exciting, and dynamically evolving world waits out there to be discovered and used to the full benefit of neuroscience.

## Data availability statement

No original data were generated.

## Code availability statement

We generated code in MATLAB 2016b (MathWorks) for the simulations presented in the figures, available at [https://github.com/kiralyb/model\\_selection\\_mines](https://github.com/kiralyb/model_selection_mines).

## References

- Adesnik H, Bruns W, Taniguchi H, Huang ZJ, Scanziani M (2012) A neural circuit for spatial summation in visual cortex. *Nature* 490:226–231.
- Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95:631–636.
- Akaike H (1969) Fitting autoregressive models for prediction. *Ann Inst Stat Math* 21:243–247.
- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: 2nd International symposium on information theory, pp 267–281. Budapest: Akadémiai Kiadó.
- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–723.
- Akaike H (1978) A Bayesian analysis of the minimum AIC procedure. *Ann Inst Stat Math* 30:9–14.
- Armañanzas R, Ascoli GA (2015) Towards the automatic classification of neurons. *Trends Neurosci* 38:307–318.
- Atallah BV, Bruns W, Carandini M, Scanziani M (2012) Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron* 73:159–170.
- Baajour SJ, Chowdury A, Thomas P, Rajan U, Khatib D, Zajac-Benitez C, Falco D, Haddad L, Amirsadri A, Bressler S, Stanley JA, Diwadkar VA (2020) Disordered directional brain network



- interactions during learning dynamics in schizophrenia revealed by multivariate autoregressive models. *Hum Brain Mapp* 41:3594–3607.
- Banks HT, Joyner ML (2017) AIC under the framework of least squares estimation. *Appl Math Lett* 74:33–45.
- Bányai M, Lazar A, Klein L, Klon-Lipok J, Stippinger M, Singer W, Orbán G (2019) Stimulus complexity shapes response correlations in primary visual cortex. *Proc Natl Acad Sci U S A* 116:2723–2732.
- Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. *Phys Rev Lett* 76:3251–3254.
- Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory* 2008:P10008.
- Brewer MJ, Butler A, Cooksley SL (2016) The relative performance of AIC, AIC C and BIC in the presence of unobserved heterogeneity. *Methods Ecol Evol* 7:679–692.
- Browne MW (2000) Cross-validation methods. *J Math Psychol* 44:108–132.
- Butts DA, Goldman MS (2006) Tuning curves, neuronal variability, and sensory coding. *PLoS Biol* 4:e92.
- Buzsáki G (2002) Theta oscillations in the hippocampus. *Neuron* 33:325–340.
- Buzsáki G, Moser EI (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat Neurosci* 16:130–138.
- Cavanaugh JE (1997) Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat Probab Lett* 33:201–208.
- Cavanaugh JE, Neath AA (2019) The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements. *Wiley Interdiscip Rev Comput Stat* 11:1–11.
- Celeux G, Forbes F, Robert CP, Titterton DM (2006) Deviance information criteria for missing data models. *Bayesian Anal* 1:651–674.
- Chamroukhi F, Huynh B (2019) Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In: 2018 International joint conference on neural networks, pp 1–8. New York: IEEE.
- Chandrasekaran C, Soldado-Magraner J, Peixoto D, Newsome WT, Shenoy KV, Sahani M (2018) Brittleness in model selection analysis of single neuron firing rates. *bioRxiv*. doi:10.1101/430710.
- Chang L, Tsao DY (2017) The code for facial identity in the primate brain. *Cell* 169:1013–1028.e14.
- Chari T, Banerjee J, Pachter L (2021) The specious art of single-cell genomics. *bioRxiv* 1–25.
- Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV (2012) Neural population dynamics during reaching. *Nature* 487:51–56.
- Cohen JR, Asarnow RF, Sabb FW, Bilder RM, Bookheimer SY, Knowlton BJ, Poldrack RA (2010) Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. *Front Hum Neurosci* 4:47.
- Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482:85–88.
- Cunningham JP, Shenoy K V., Sahani M (2008) Fast Gaussian process methods for point process intensity estimation. In: Proceedings of the 25th international conference on machine learning, pp 192–199. New York: ACM.
- Czurkó A, Huxter J, Li Y, Hangya B, Muller RU (2011) Theta phase classification of interneurons in the hippocampal formation of freely moving rats. *J Neurosci* 31:2938–2947.
- de Vries SEJ, Lecoq JA, Buice MA, Groblewski PA, Ocker GK, Oliver M, Feng D, Cain N, Ledochowitsch P, Millman D, Roll K, Garrett M, Keenan T, Kuan L, Mihalas S, Olsen S, Thompson C, Wakeman W, Waters J, Williams D, et al. (2020) A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat Neurosci* 23:138–151.
- Diaz-Papkovich A, Anderson-Trocmé L, Gravel S (2021) A review of UMAP in population genetics. *J Hum Genet* 66:85–91.
- Disney AA, Aoki C, Hawken MJ (2007) Gain modulation by nicotine in macaque v1. *Neuron* 56:701–713.
- Domany E (1999) Superparamagnetic clustering of data — The definitive solution of an ill-posed problem. *Phys A Stat Mech its Appl* 263:158–169.
- Dupret D, O’Neill J, Pleydell-Bouverie B, Csicsvari J (2010) The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nat Neurosci* 13:995–1002.
- Erdős P (1942) On the law of the iterated logarithm. *Ann Math* 43:419–436.
- Evans NJ (2019) Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychon Bull Rev* 26:1070–1098.
- Faure L, Wang Y, Kastri ME, Fontanet P, Cheung KKY, Petitpré C, Wu H, Sun LL, Runge K, Croci L, Landy MA, Lai HC, Consalez GG, de Chevigny A, Lallemand F, Adameyko I, Hadjab S (2020) Single cell RNA sequencing identifies early diversity of sensory neurons forming via bi-potential intermediates. *Nat Commun* 11:4175.
- Fisher NI (1993) Statistical analysis of circular data. Cambridge: Cambridge University Press.
- Fong E, Holmes CC (2020) On the marginal likelihood and cross-validation. *Biometrika* 107:489–496.
- Franaszczuk PJ, Bergey GK (1999) An autoregressive method for the measurement of synchronization of interictal and ictal EEG signals. *Biol Cybern* 81:3–9.
- Froemke RC, Merzenich MM, Schreiner CE (2007) A synaptic memory trace for cortical receptive field plasticity. *Nature* 450:425–429.
- Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Stat Comput* 24:997–1016.
- Genkin M, Engel TA (2020) Moving beyond generalization to accurate interpretation of flexible models. *Nat Mach Intell* 2:674–683.
- Goutte C, Toft P, Rostrup E, Nielsen FÅ, Hansen LK (1999) On clustering fMRI time series. *Neuroimage* 9:298–310.
- Gouwens NW, et al. (2019) Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat Neurosci* 22:1182–1195.
- Gouwens NW, et al. (2020) Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. *Cell* 183:935–953.e19.
- Gronau QF, Wagenmakers EJ (2019) Limitations of Bayesian leave-one-out cross-validation for model selection. *Comput Brain Behav* 2:1–11.
- Gwo CY, Zhu DC, Zhang R (2019) Brain white matter hyperintensity lesion characterization in T2 fluid-attenuated inversion recovery magnetic resonance images: shape, texture, and potential growth. *Front Neurosci* 13:1–15.
- Hahamy A, Behrmann M, Malach R (2015) The idiosyncratic brain: distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nat Neurosci* 18:302–309.
- Hangya B, Pi HJ, Kvitsiani D, Ranade SP, Kepecs A (2014) From circuit motifs to computations: mapping the behavioral repertoire of cortical interneurons. *Curr Opin Neurobiol* 26:117–124.
- Hangya B, Ranade SP, Lorenc M, Kepecs A (2015) Central cholinergic neurons are rapidly recruited by reinforcement feedback. *Cell* 162:1155–1168.
- Hannan E, Quinn B (1979) The Determination of the order of an autoregression. *J R Stat Soc Ser B Stat Methodol* 41:190–195.
- Harris KD, Hochgerner H, Skene NG, Magno L, Katona L, Bengtsson Gonzales C, Somogyi P, Kessaris N, Linnarsson S, Hjerling-Leffler J (2018) Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biol* 16:e2006387.
- Harrison L, Penny WD, Friston K (2003) Multivariate autoregressive modeling of fMRI time series. *Neuroimage* 19:1477–1491.
- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, Springer series in statistics. New York: Springer New York.
- Hillier D, Fiscella M, Drinnenberg A, Trenholm S, Rompani SB, Raics Z, Katona G, Juettner J, Hierlemann A, Rozsa B, Roska B (2017) Causal evidence for retina-dependent and -independent visual motion computations in mouse cortex. *Nat Neurosci* 20:960–968.



- Hirokawa J, Vaughan A, Masset P, Ott T, Kepecs A (2019) Frontal cortex neuron types categorically encode single decision variables. *Nature* 576:446–451.
- Hromádka T, Deweese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol* 6:e16.
- Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574–591.
- Ito M, Aswendt M, Lee AG, Ishizaka S, Cao Z, Wang EH, Levy SL, Smerin DL, McNab JA, Zeineh M, Leuze C, Goubran M, Cheng MY, Steinberg GK (2018) RNA-sequencing analysis revealed a distinct motor cortex transcriptome in spontaneously recovered mice after stroke. *Stroke* 49:2191–2199.
- James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to statistical learning, Springer texts in statistics. New York: Springer New York.
- Jang H, Kim K-KK, Braatz RD, Gopaluni RB, Lee JH (2016) Regularized maximum likelihood estimation of sparse stochastic monomolecular biochemical reaction networks. *Comput Chem Eng* 90:111–120.
- Jasa T, Lanz T, Ott R, Stoop R (2008) Sequential superparamagnetic clustering as a predictor of visual fixations In: NOLTA (nonlinear theory and applications), pp 120–123. IECE: Tokyo.
- Jeffrey H (1961) The theory of probability, Ed 3. Oxford: Oxford University Press.
- Jiang J, Dai B, Peng D, Zhu C, Liu L, Lu C (2012) Neural synchronization during face-to-face communication. *J Neurosci* 32:16064–16069.
- Jung YG, Kang MS, Heo J (2014) Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnol Biotechnol Equip* 28:S44–S48.
- Kántor O, Varga A, Tóth R, Énzsöly A, Pálfi E, Kovács-Öller T, Nitschke R, Szél Á, Székely A, Völgyi B, Négyessy L, Somogyvári Z, Lukáts Á (2015) Stratified organization and disorganization of inner plexiform layer revealed by TNAP activity in healthy and diabetic rat retina. *Cell Tissue Res* 359:409–421.
- Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.
- Keyser C, Gazzola V, Wagenmakers E-J (2020) Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat Neurosci* 23:788–799.
- Kilgard MP, Merzenich MM (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* 279:1714–1718.
- Klausberger T, Somogyi P (2008) Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. *Science* 321:53–57.
- Kloosterman F, Layton SP, Chen Z, Wilson MA (2014) Bayesian decoding using unsorted spikes in the rat hippocampus. *J Neurophysiol* 111:217–227.
- Kobak D, Berens P (2019) The art of using t-SNE for single-cell transcriptomics. *Nat Commun* 10:5416.
- Kohavi R (1995) A study of crossvalidation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence, pp. 1137–1143. San Francisco: Morgan Kaufmann Publishers.
- Konishi S, Kitagawa G (2008) Information criteria and statistical modeling, Springer series in statistics. New York: Springer New York.
- Koyama S, Eden UT, Brown EN, Kass RE (2010) Bayesian decoding of neural spike trains. *Ann Inst Stat Math* 62:37–59.
- Krueger J, Disney AA (2019) Structure and function of dual-source cholinergic modulation in early vision. *J Comp Neurol* 527:738–750.
- Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW (2015) Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349:184–187.
- Lee J, Maunsell JHR (2009) A normalization model of attentional modulation of single unit responses. *PLoS One* 4:e4651.
- Lee EK, Balasubramanian H, Tsolias A, Anakwe SU, Medalla M, Shenoy KV, Chandrasekaran C (2021) Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex. *Elife* 10:e67490.
- Lee SH, Kwan AC, Zhang S, Phoumthippavong V, Flannery JG, Masmanidis SC, Taniguchi H, Huang ZJ, Zhang F, Boyden ES, Deisseroth K, Dan Y (2012) Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature* 488:379–383.
- Li CL, Li KC, Wu D, Chen Y, Luo H, Zhao JR, Wang SS, Sun MM, Lu YJ, Zhong YQ, Hu XY, Hou R, Zhou BB, Bao L, Xiao HS, Zhang X (2016) Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Res* 26:83–102.
- Li Y, Yu J, Zeng T (2020) Deviance information criterion for latent variable models and misspecified models. *J Econom* 216:450–493.
- Magland J, Jun JJ, Lovero E, Morley AJ, Hurwitz CL, Buccino AP, Garcia S, Barnett AH (2020) SpikeForest, reproducible web-facing ground-truth validation of automated neural spike sorters. *Elife* 9:e55167.
- Mainassara YB, Kokonendji CC (2016) Modified Schwarz and Hannan–Quinn information criteria for weak VARMA models. *Stat Inference Stoch Process* 19:199–217.
- Maunsell JHR, Treue S (2006) Feature-based attention in visual cortex. *Trends Neurosci* 29:317–322.
- McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv* 1–63.
- Miche Y, Lendasse A (2009) A faster model selection criterion for OP-ELM and OP-KNN: Hannan–Quinn criterion. *Proc 17th Eur Symp Artif Neural Netw* 1:177–182.
- Muller R, Kubie J, Ranck J (1987) Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *J Neurosci* 7:1935–1950.
- Muthuswamy J, Thakor NV (1998) Spectral analysis methods for neurological signals. *J Neurosci Methods* 83:1–14.
- Nguyen T, Khosravi A, Creighton D, Nahavandi S (2014) Spike sorting using locality preserving projection with gap statistics and landmark-based spectral clustering. *J Neurosci Methods* 238:43–53.
- Oikonomou VP, Tzallas AT, Fotiadis DI (2007) A Kalman filter based methodology for EEG spike enhancement. *Comput Methods Programs Biomed* 85:101–108.
- O'Keefe J (1993) Hippocampus, theta, and spatial memory. *Curr Opin Neurobiol* 3:917–924.
- Okun M, Yger P, Marguet SL, Gerard-Mercier F, Benucci A, Katzner S, Busse L, Carandini M, Harris KD (2012) Population rate dynamics and multineuron firing patterns in sensory cortex. *J Neurosci* 32:17108–17119.
- Pei F, Ye J, Zoltowski D, Wu A, Chowdhury RH, Sohn H, O'Doherty JE, Shenoy K V., Kaufman MT, Churchland M, Jazayeri M, Miller LE, Pillow J, Park IM, Dyer EL, Pandarinath C (2021) Neural Latents Benchmark '21: evaluating latent variable models of neural population activity. In: 35th Conference on neural information processing systems (NeurIPS 2021), pp 1–37. NeurIPS: New York.
- Pi HJ, Hangya B, Kvitsiani D, Sanders JL, Huang ZJ, Kepecs A (2013) Cortical interneurons that specialize in disinhibitory control. *Nature* 503:521–524.
- Que L, Lukacsovich D, Luo W, Földy C (2021) Transcriptional and morphological profiling of parvalbumin interneuron subpopulations in the mouse hippocampus. *Nat Commun* 12:108.
- Quiroga RQ (2012) Spike sorting. *Curr Biol* 22:R45–R46.
- Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput* 16:1661–1687.
- Rao CR, Nayak TK (1985) Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Trans Inform Theory* 31:589–593.
- Rouder JN, Morey RD (2019) Teaching Bayes' theorem: strength of evidence as predictive accuracy. *Am Stat* 73:186–190.
- Sadacca BF, Jones JL, Schoenbaum G (2016) Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife* 5:e13665.

- Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-kabara EC, Yu BM, Batista AP (2014) Neural constraints on learning. *Nature* 512:423–426.
- Sahani M (1999) Latent variable models for neural data analysis. Dissertation (Ph.D.). California Institute of Technology. doi:10.7907/hwy6-ar88.
- Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA (2018) Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 174:1015–1030.e16.
- Savage N (2019) How AI and neuroscience drive each other forwards. *Nature* 571:S15–S17.
- Savin C, Tkačik G (2017) Maximum entropy models as a tool for building precise neural controls. *Curr Opin Neurobiol* 46:120–126.
- Schmitzer-Torbert N, Jackson J, Henze D, Harris K, Redish D (2005) Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* 131:1–11.
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Sedaghat-Nejad E, Fakharian MA, Pi J, Hage P, Kojima Y, Soetedjo R, Ohmae S, Medina JF, Shadmehr R (2021) P-sort: an open-source software for cerebellar neurophysiology. *J Neurophysiol* 126:1055–1075.
- Seghouane AK (2010) Asymptotic bootstrap corrections of AIC for linear regression models. *Signal Processing* 90:217–224.
- Seghouane AK, Amari SI (2007) The AIC criterion and symmetrizing the Kullback-Leibler divergence. *IEEE Trans Neural Netw* 18:97–106.
- Seriès P, Latham PE, Pouget A (2004) Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat Neurosci* 7:1129–1135.
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22:888–905.
- Shlens J (2014) Notes on Kullback-Leibler divergence and likelihood. arXiv:1404.
- Shmueli G (2010) To explain or to predict? *Stat Sci* 25:289–310.
- Sin CY, White H (1996) Information criteria for selecting possibly misspecified parametric models. *J Econom* 71:207–225.
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J R Stat Soc Series B Stat Methodol* 64:583–616.
- Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2014) The deviance information criterion: 12 years on. *J R Stat Soc Series B Stat Methodol* 76:485–493.
- Stephan KE, Harrison LM, Penny WD, Friston KJ (2004) Biophysical models of fMRI responses. *Curr Opin Neurobiol* 14:629–635.
- Takahashi YK, Langdon AJ, Niv Y, Schoenbaum G (2016) Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron* 91:182–193.
- Thirion B, Varoquaux G, Dohmatob E, Poline JB (2014) Which fMRI clustering gives good brain parcellations? *Front Neurosci* 8:1–13.
- Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Series B Stat Methodol* 63:411–423.
- Ting CM, Seghouane AK, Salleh SH, Noor AM (2015) Estimating effective connectivity from fMRI data using factor-based subspace autoregressive models. *IEEE Signal Process Lett* 22:757–761.
- Tkačik G, Marre O, Mora T, Amodei D, Berry MJ, Bialek W (2013) The simplest maximum entropy model for collective behavior in a neural network. *J Stat Mech Theory Exp* 2013:1–5.
- Townsend BR, Subasi E, Scherberger H (2011) Grasp movement decoding from premotor and parietal cortex. *J Neurosci* 31:14386–14398.
- van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605.
- Varshavsky R, Horn D, Linal M (2008) Global considerations in hierarchical clustering reveal meaningful patterns in data. *PLoS One* 3:e2247.
- Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* 27:1413–1432.
- Vértes E, Sahani M (2018) Learning doubly intractable latent variable models via score matching. In: Proceedings of 32nd conference on neural information processing systems (*NeurIPS 2018*), pp 1–4. NeurIPS: New York.
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58:236–244.
- Wasserman L (2000) Bayesian model selection and model averaging. *J Math Psychol* 44:92–107.
- Watanabe S (2010a) Equations of states in singular statistical estimation. *Neural Netw* 23:20–34.
- Watanabe S (2010b) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 11:3571–3594.
- Watanabe S (2013) A widely applicable Bayesian information criterion. *J Mach Learn Res* 14:867–897.
- Wilson NR, Runyan CA, Wang FL, Sur M (2012) Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature* 488:343–348.
- Yang GR, Wang X (2020) Artificial neural networks for neuroscientists: a primer. *Neuron* 107:1048–1070.
- Yang MS (1993) A survey of fuzzy clustering. *Math Comput Model* 18:1–16.
- Yger P, Spampinato GLB, Esposito E, Lefebvre B, Deny S, Gardella C, Stimberg M, Jetter F, Zeck G, Picaud S, Duebel J, Marre O (2018) A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *Elife* 7:e34518.
- Yong L (2018) LOO and WAIC as model selection methods for polytomous items. arXiv:1806.09996.
- Zawadzki K, Miazaki M, Costa LDF (2010) Investigating the morphological categories in the NeuroMorpho database by using superparamagnetic clustering. arXiv:1003.3036.