# The problem of perfect predictors in statistical spike train models

**Sahand Farhoodi[1]** | **Uri Eden[1]**

[1]Department of Mathematics and Statistics, Boston University, Boston, MA, 02215, USA

**Correspondence**
Sahand Farhoodi, Department of Mathematics and Statistics, Boston University, Boston, MA, 02215, USA
Email: sahand@bu.edu

Generalized Linear Models (GLMs) have been used extensively in statistical models of spike train data. However, the IRLS algorithm, which is often used to fit such models, can fail to converge in situations where response and non-response can be separated by a single predictor or a linear combination of multiple predictors. Such situations are likely to arise in many neural systems due to properties such as refractoriness and incomplete sampling of the signals that influence spiking. In this paper, we describe multiple classes of approaches to address this problem: Standard IRLS with a fixed iteration limit, computing the maximum likelihood solution in the limit, Bayesian estimation, regularization, change of basis, and modifying the search parameters. We demonstrate a specific application of each of these methods to spiking data from rat somatosensory cortex and discuss the advantages and disadvantages of each. We also provide an example of a roadmap for selecting a method based on the problem's particular analysis issues and scientific goals.

**KEYWORDS**
*neural coding, generalized linear models, model convergence, perfect predictors, complete separation*

# 1 | INTRODUCTION

Generalized linear models (GLM) provide a powerful tool for relating observed neural spiking data to the biological, behavioral, and stimulus signals that influence them [1, 2, 3]. These models express the conditional intensity of spiking, which defines the probability of observing a spike in any small time interval, in terms of a design matrix whose columns represent the signals, or covariates, of interest. GLMs possess a number of properties that make them well suited to spike train modeling [1, 4, 5, 6]. One important property is that GLMs can be formulated to guarantee that the likelihood of the spiking data is convex with respect to the model parameters, allowing for rapid computation of their maximum likelihood estimates (MLE) [7, 8]. Most statistical software packages implement maximum likelihood estimation for GLMs using the computationally efficient Iteratively Reweighted Least Squares algorithm (IRLS algorithm) [9, 10, 11]. Another useful property of GLMs is that in most cases, the parameter estimates have an asymptotically multivariate normal distribution with a covariance matrix determined by the Fisher information, which is computed as part of the IRLS algorithm [9, 11]. This makes it easy to compute confidence intervals (CI) about individual parameters or about the firing intensity, as well as providing for simple hypothesis tests, e.g. maximum likelihood ratio tests, about whether firing rates are influenced by specific sets of covariates [9, 11]. Accurate estimation of the Fisher information also helps determine the extent to which one signal or set of signals is confounded with another in its influence on neural spiking [9, 11].

For some GLMs, the IRLS algorithm may fail to converge because the likelihood is maximized in the limit as one or multiple parameters tend to $\pm\infty$. This problem is sometimes referred to as the monotone likelihood or non-existence of the MLE problem [12]. Another perspective is that this situation arises when the response and non-response can be separated by a single predictor or a linear combination of multiple predictors. For this reason, this phenomenon is more commonly referred to as the complete separation problem [13]. This problem is well-studied in the case of logistic regression with small to medium-sized data [13, 14, 15, 16, 17, 18]. Heinze et. al. explore a number of potential solutions for this case including the omission of predictors that result in complete separation, fixing the parameter values for such predictors prior to fitting the model, changing the form of the model, and using an ad-hoc adjustment [19, 20]. They arrive at a preferred procedure using a modification proposed by Firth in order to reduce the bias of maximum likelihood estimates in GLMs [21, 22, 23].

The problem arises in the statistical literature on point processes such as Cox processes as well [24, 25], but is not discussed in detail in the neural modeling literature, despite the fact that it is likely to arise in many neural systems due to properties such as refractoriness and incomplete sampling of the signals that influence spikes. In these models, separation occurs when any covariate forces the system not to spike whenever it takes on a nonzero value. Such covariates are called perfect predictors, since one can perfectly predict no spiking when they are nonzero. For example, a model for a neuron with an absolute refractory period lasting longer than 1 ms will be perfectly predicted by an indicator for whether there has been a spike in the past ms. Perfect predictors can also appear in models when nonzero values of certain predictors are infrequently sampled, even in cases where a lot of data exists. For example, a model fit for a rat's hippocampal place field that includes an indicator for being in a corner of the environment that the rat tends to avoid may suggest that this variable is a perfect predictor, even though the neuron might occasionally fire in that region given enough time. We say that perfect predictors of the kind in the first example are structural, while ones of the kind in the second example are the result of sampling.

We posit that when these issues arise in statistical neural models, researchers adopt one of a set of ad-hoc methods to avoid dealing with it, and the issues do not make the discussion of the resulting papers. However, perfect prediction can lead to a variety of issues limiting the utility of inference from point process GLMs, which may not be alleviated by ad-hoc approaches. One issue is that the IRLS algorithm will not converge, and will typically only stop

when some fixed iteration limit has been reached. This often involves an order of magnitude more computational time than fitting a GLM without perfect predictors. In some cases, the IRLS algorithm reaches a region of the likelihood surface where the computations become numerically unstable, and subsequent iterations can actually lead to a reduction in likelihood. Another important issue is that the computed Fisher information in these fit models does not accurately reflect the variance-covariance structure of the parameter estimates. In particular, the computed Fisher information is often close to singular, and if inappropriately used to construct confidence bounds and perform hypothesis tests, typically leads to extremely large standard error estimates for the perfect predictors and inaccurate covariance between the perfect predictors and all other signals [19, 26, 27].

In this paper, we explore a range of different approaches for dealing with perfect predictors in point process GLMs. These approaches fall into multiple categories: Interpreting the IRLS output despite lack of convergence, fixing the parameter estimates for perfect predictors, Bayesian estimation methods, regularization methods, re-parameterizing the model to remove perfect predictors, and modifications to the parameter search domain or stopping criteria. In the Methods section, we introduce these different approaches and explain how they deal with the problem of separation. In the Results section, we select one specific method from each category and illustrate its application to modeling spiking data from a rat cortical neuron, in vitro. Furthermore, we compare these methods and illustrate how they deal with different problems associated with complete separation. Finally, in the Discussion section, we discuss the properties of these approaches, point out some advantages and disadvantages of each method, and introduce a potential road map to select the most appropriate method to use in various situations.

## 2 | METHODS

### 2.1 | Problem Formulation

Let $X$ be the design matrix with $p$ predictors of neural spiking (columns) and $n$ observations (rows) and $Y = (y_1, \ldots, y_n)^T$ be the response vector of spike counts where the superscript $T$ denotes the transpose of a matrix. We use $x_{ij}$, $x_{i\cdot}$ and $x_{\cdot j}$ respectively to denote the $ij$-th element, $i$-th row and $j$-th column of matrix $X$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ be the parameter vector, $\theta = X\beta$, and $\mu = \mathbb{E}(Y|X)$ be the conditional expectation of $Y$ given $X$. Let $g(\cdot)$ be the link function that relates the conditional expectation of $Y$ to the linear combination of predictors, i.e. $g(\mu) = \theta$. Here, we work with the canonical link function for Poisson point process which is the log function [9]. The Poisson point process GLM is described by

$$y_j | x_{\cdot j} \sim \text{Poiss}(\lambda_j) \tag{1a}$$

$$\lambda_j = \exp\left(\beta_0 + \sum_{i=1}^{p} x_{ji}\beta_i\right). \tag{1b}$$

where $\lambda_j$ is the firing rate for the $j$-th observation and $1 \leq j \leq n$. The IRLS algorithm is widely used to find the maximum likelihood solution for any generalized linear model. This algorithm starts with an initial vector $\beta^{(0)}$ and then computes $\beta^{(s)}$ recursively using the update equation

$$\beta^{(s+1)} = \beta^{(s)} + I(\beta^{(s)})^{-1} U(\beta^{(s)}) \tag{2}$$

until convergence where $U(\beta)$ and $I(\beta)$ respectively denote the score function and the Fisher information computed for $\beta$. For the process described in Eq. 1, the log-likelihood function $l(\beta)$, the score function $U(\beta)$, and the Fisher

information $I(\beta)$ are defined as

$$I(\beta) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i \tag{3a}$$

$$U(\beta) = \frac{\partial I(\beta)}{\partial \beta} = X^T (Y - \mu) \tag{3b}$$

$$I(\beta) = -\frac{\partial^2 I(\beta)}{\partial \beta^2} = X^T W X \tag{3c}$$

where $W$ is the matrix of weights defined by $W^{-1} = (\frac{d\theta}{d\mu})^2 V$ and $V$ is the variance function evaluated at $\mu$ [9]. For our choice of link function $g(\cdot)$, $W$ will be a diagonal matrix with $i$-th diagonal element equal to $\mu_i$.

The $i$-th column of $X$ is a perfect predictor if having a nonzero value at this column leads to zero spike count, i.e. for any $j \in \{1, \ldots, n\}$, $x_{ji} \neq 0$ implies $y_j = 0$. Additionally, a set of columns, $i_1, \ldots, i_m$, generate a perfect predictor if some linear combination of these are a perfect predictor, i.e. there exists a set of weights $a_1, \ldots, a_m$, such that whenever $a_1 x_{ji_1} + \cdots + a_m x_{ji_m} \neq 0$ then $y_j = 0$. We let $S$ be the set of row indices corresponding to perfect predictors or columns that generate perfect predictors, and define the set of perfect parameters as $\{\beta_i : i \in S\}$. These parameters diverge to $\pm\infty$ when the IRLS algorithm (described in Eq. 2) is used to fit the model. We say that $i$-th row is a perfect row if it has at least one nonzero value at a perfect column.

A naive way to deal with perfect predictors is to omit them from the model by removing all perfect columns and perfect rows from $X$. However, perfect predictors are extremely informative and removing them can harm the model substantially in terms of goodness-of-fit and statistical power. Putting this naive approach aside, in what follows we briefly explore different families of approaches that can be used to deal with perfect predictors.

## 2.2 | Standard IRLS with Fixed Iteration Limit

The first approach for dealing with perfect predictors is to use the standard IRLS procedure to estimate the MLE without adjusting the model or estimation procedure. As pointed out before, in the presence of perfect predictors the actual maximum likelihood solution is achieved only when perfect parameters are equal to $\pm\infty$, and therefore the parameter estimates will depend on when the algorithm terminates, typically after a fixed iteration limit is reached. We will call this approach "Standard IRLS" throughout the rest of this paper.

## 2.3 | Maximum Likelihood Limit

Eq. 1b can also be rewritten as

$$\lambda_j = \exp\left(\beta_0 + \sum_{i \notin S} x_{ji}\beta_i + \sum_{i \in S} |x_{ji}|\beta_i\right). \tag{4}$$

For perfect parameters ($i \in S$) taking the absolute value of $x_{ji}$ only affects the sign of $\beta_i$ and doesn't change the predicted values of our model. Thus, the maximum likelihood solution is achieved when for any $i \in S$, $\beta_i$ is equal to its limiting value $-\infty$. Therefore, one approach is to set perfect parameters equal to their limiting value manually, and then use the IRLS algorithm to estimate non-perfect parameters, after omitting perfect columns and rows from $X$ and $Y$. The final model then includes both the estimated parameters when perfect predictors are removed, and the perfect parameters set to $-\infty$. In practice, a perfect parameter cannot be set numerically equal to -$\infty$, so often a very

large negative number is used instead.

## 2.4 | Bayesian Estimation

The next approach to deal with the problem of separation is using Bayesian GLM. In this approach, we treat the parameters $\beta$ as random variables, assign them prior distributions $\pi$, and compute and maximize their posterior distributions given the observed data. For the case that $\pi$ is a zero-mean distribution with covariance matrix $\Sigma$, the posterior log-likelihood function, the posterior score function, and the posterior Fisher information are respectively given by

$$l_b(\beta) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i - \tfrac{1}{2}\beta^T \Sigma^{-1} \beta \tag{5a}$$

$$U_b(\beta) = X^T(Y - \mu) - \Sigma^{-1}\beta \tag{5b}$$

$$I_b(\beta) = X^T W X + \Sigma^{-1} \tag{5c}$$

where the subscript $b$ stands for Bayesian. In order to use the IRLS algorithm to fit Bayesian GLM, these equations can be used in Eq. 2 to maximize $l_b(\beta)$. An appropriate choice of $\Sigma$ can solve the perfect predictor problem in a couple of ways: the diagonal terms of $\Sigma$ prevent any individual parameters from diverging to $\pm\infty$, and the off-diagonal terms of $\Sigma$ can be used to impose smoothness between sets of parameters, so that sampling issues are less likely to cause perfect predictors.

## 2.5 | Regularization Methods

Another approach that is commonly used to address the problem of separation is regularization. In this approach, instead of maximizing the log-likelihood function, a weighted average of the log-likelihood function and a penalty function, $p(\beta)$, is maximized. This penalty function is intended to prevent parameters from diverging by penalizing larger values of the parameters. Lasso and Ridge penalty functions are two of the most used [28]. A penalty function that has been used specifically to deal with the problem of separation in logistic regression is Firth's penalty function [20]. In fitting a regularized GLM, the regularized log-likelihood function, the regularized score function and the regularized Fisher information are computed according to

$$l_r(\beta) = (1 - \Lambda)l(\beta) - \Lambda p(\beta) \tag{6a}$$

$$U_r(\beta) = (1 - \Lambda)X^T(Y - \mu) - 2\Lambda \frac{dp(\beta)}{d\beta} \tag{6b}$$

$$I_r(\beta) = (1 - \Lambda)X^T W X - 2\Lambda \frac{d^2 p(\beta)}{d\beta^2} \tag{6c}$$

where the subscript $r$ stands for regularization and $\Lambda$ is a variable determining the amount of penalization. These quantities then are used directly in the IRLS algorithm described by Eq. 2 to maximize the regularized log-likelihood function.

## 2.6 | Change of Basis

The next family of methods focuses on reparameterizing the model in order to impose smoothness between specific sets of parameters. This method can be used when we believe that the receptive field has some smoothness properties that are not explicitly captured by the model structure. For example, it is common to use estimates spike rates in spatial bins to model place fields, despite the fact that this does not capture the fact that rates in adjacent bins tend to be close. Similarly, models of refractoriness or bursting often focus on autocovariance in lagged bins, ignoring the tendency of neurons to have smooth history dependence structure. One way of imposing smoothness on predictors is using basis functions such as smoothing splines [29, 30] or raised cosines [31]. Employing this approach Eq. 1b is replaced by

$$\lambda_j = \exp\Big(\tilde{\beta}_0 + \sum_{k=1}^{q} g_k(x_{j.})\tilde{\beta}_k\Big) \tag{7}$$

where $g_k$ is the $k$-th basis function and $q$ denotes the number of basis functions. As before, the IRLS algorithm described in Eq. 2 can be used to fit this model. This approach is capable of alleviating issues related to both structural and sampling perfect predictors, by introducing a new smoother predictor, $g_k(x_{j.})$, that integrates information over a range of the receptive field that is better sampled.

## 2.7 | Modified Search Criteria

Another approach is to modify the search or stopping criteria for the estimation algorithm to prevent parameters from diverging. One explicit way of doing so is to restrict the search space of the IRLS algorithm and force parameters to remain in that restricted space. Another way of preventing parameters from diverging by modifying the search criterion is to update $\beta_i$ in equation 2 only if $U(\beta_i)$ is greater than a threshold. This prevents the IRLS algorithm from making adjustments to $\beta$ when the log-likelihood function is nearly flat, which is a characteristic of the likelihood in the presence of perfect predictors.

# 3 | RESULTS

## 3.1 | Data Set and the Specific Estimation Algorithms

In this section, we pick one specific method from each family of approaches introduced in the Methods section, and compare them for the problem of fitting spiking data collected from rat cortical neurons. This data has been previously used for challenge A of the Quantitative Single-Neuron Modeling Competition (2009) developed by researchers at Ecole Polytechnique Fédérale de Lausanne (EPFL) and is accessible on CRCNS.org [1]. It contains spiking activity from the somatosensory cortex of a Wistar rat in response to somatic current stimulation. This dataset contains 13 repetitions of a 60-second stimulation protocol, where for the first 39 seconds both the injected current waveform in pA and voltage responses are provided [32, 33]. We only use the first 39 seconds of these repetitions to construct our training set. Two repetitions are used to fit the model and the other repetitions are used for resampling to compute cross-validation statistics and to examine whether perfect predictors arise due to subsampling. Different pairs of repetitions are used to fit each model multiple times (10 times in total) and we report average statistics for memory usage and computational run times (See Table 1).

---

[1]See http://crcns.org/data-sets/challenges/ch-epfl-2009

We denote the firing rate of the neuron at time $t$ by $\lambda(t)$, the spiking count in the interval $(t, t + \Delta t)$ by $\Delta N_t$, the firing history of the neuron up to time $t$ by $H_t$, and let $p$ be a fixed value representing the length of history which we believe can influence the current spike intensity. We divide the whole range of current stimulus levels into $q$ disjoint intervals, and let $I_i(t)$ be an indicator function that takes value 1 if the input current at time $t$ falls into the $i$-th interval. We relate the firing rate at time $t$ to the past history of spike counts and the input current using the following point process GLM:

$$\lambda(t \mid H_t) = \exp\Big(\beta_0 + \sum_{i=1}^{p} \beta_i \Delta N_{t-i} + \sum_{i=1}^{q} \beta_{p+i} I_i(t)\Big). \tag{8}$$

where $p$ and $q$ are set to 200 ms and 6 respectively. In general, statistical model identification procedures, such as the step-wise regression, can be used to select the order of a model. Here, our purpose is not to identify the most parsimonious model, but to introduce a model that can capture specific features of the data, and compare multiple fitting procedures in their handling of perfect predictors.

When we fit this model for a single neuron using the IRLS algorithm directly, we obtain a fit with 31 perfect predictors. 20 of these perfect predictors are associated with the parameters that capture history dependence (first sum of Eq. 8) occurring at small lags due to the refractory period of the neuron. These likely reflect structural perfect predictors, in that we would not expect to see any spikes occur immediately after another, even if we collected much more data. There are also two perfect predictors corresponding to the two lowest current indicator functions, $I_1(t)$ and $I_2(t)$. It is not immediately clear if these are structural perfect predictors or sampling predictors (i.e. whether the neuron has a non-negligible probability of firing when the stimulus current is this low). In this case, when we increase the data size (by using the extra trials in the training set to fit the model), these perfect predictors remain so, suggesting they may be structural as well. The remaining perfect predictors occurred either in the history component of the model at specific lags that had neighbors that were not perfect predictors, or in the stimulus response component for stimulus levels that we expect could drive some amount of neural spiking. In order to analyze if these perfect predictors are structural or due to sampling, we expand the training data by including additional trials and fit the same model described in Eq. 8. We observed that all these perfect predictors vanish and hence categorize them as sampling perfect predictors, since they only existed in the first place because of limited sampling in the training data. Without using any of the methods described in the Methods section, all the perfect predictors in this example diverged to large negative numbers, limited only due to the finite iteration limit of the IRLS algorithm.

In order to compare the general approaches defined in the Methods section, we selected one specific method from each family that, in our estimation, is the most natural option or is most likely to be used by neural data analysts when encountering the problem of separation. However, we focus only on those advantages and disadvantages of each method that can be generalized to all methods of the family from which it was selected. Below, we specify the method selected from each family of approaches.

**Standard IRLS.** For this method, we set the iteration bound equal to 100, and run the IRLS algorithm without adjusting the model or estimation procedure.

**Maximum Likelihood Limit.** In this method, the perfect rows and columns are removed from $X$ and $Y$ prior to fitting the model. The final model includes the estimated parameters when perfect predictors are removed, combined with perfect parameters set to $-\infty$.

**Bayesian Estimation.** There are two types of parameters in the model described by Eq. 8, the ones corresponding history spiking counts ($\beta_{1:p}$) and the ones corresponding input current ($\beta_{p+1:p+q}$). We consider separate prior

distributions for these two groups, as we observed no correlation between their parameters. Each one of these prior distributions is a zero-mean multivariate Normal distribution with a covariance matrix $\Sigma$ with elements $\Sigma_{ij} = c^{|i-j|}$ where $c \in [0, 1]$. This choice of $\Sigma$ defines a temporal correlation on parameters that fall off geometrically as the distance between them increases, and therefore can deal with the problem of separation by imposing smoothness between nearby estimated parameters. We selected the parameter $c = 0.9$ using a leave-one-out cross-validation method. We note that more advanced methods for estimating this parameter are possible (e.g. treating it as a hyperparameter and building a hierarchical Bayesian model to estimate it).

**Regularization.** We used ridge regression, which uses an $L_2$ penalty term, to fit the model parameters because of its common usage in the neural data analysis literature. Therefore, the penalty function in Eq. 6 will be $p(\beta) = \beta^T \beta$. The choice of $\Lambda$ can affect the outcome of the regularization model significantly and thus, we use cross-validation to determine it. The value that gave the best model fit was $\Lambda = 0.1$. This specific method is called "Ridge GLM" throughout the rest of this paper.

**Change of Basis.** We reparameterized the model using a cardinal spline basis expansion [29, 30] in place of the spike count and indicator functions in Eq. 8. As shown in Appendix 1, this basis expansion method is capable of eliminating perfect predictors if the knots are chosen such that there is at least one non-perfect predictor between every two successive knots. However, by choosing knots too far from each other, a good amount of information in the data will be lost due to excessive imposed smoothness. This makes it challenging to determine the ideal number and location of the knots. One complicated but proven to work option is to use Bayesian curve-fitting with free-knot spline algorithms [34]. Another simpler option is letting knots be distributed regularly and then using cross-validation only to select the number of knots. In this case study, we used the latter approach, where we pick the numbers of knots that gives the best goodness-of-fit. This specific method is called "Cubic Smoothing Spline" throughout the rest of this paper.

**Modified Search Criterion.** As an example of a modified search criterion approach, we restrict the search space of the IRLS algorithm to $R = \{\beta : \sum_{i=1}^{p+q} \beta_i^2 \leq r\}$. We set $r = (p+q)d^2$, where $p+q$ is the total number of parameters (excluding the intercept) and $d = -5$ is the threshold for identifying if a parameter is perfect or not based on the output of the Standard IRLS method. We call this method "Bounded Search" in the rest of this paper.

## 3.2  |  Comparison of Specified Methods

First, we compare the different methods based on their goodness-of-fit. One approach for investigating goodness-of-fit is to examine Kolmogorov-Smirnov (KS) plots comparing the empirical and theoretical distributions of transformed spike times based on the times-rescaling theorem [35, 36]. Fig. 1 depicts these plots for different methods we specified in the previous section. The blue line represents the KS plot and the red lines represent global 95% bounds for a well-fit model. The larger the deviation outside of these bounds, the poorer the model fit to the data. The Cubic Smoothing Spline provides the best fit and Ridge GLM provides the poorest fit to the data from this neuron in terms of the KS analysis. The rest of the methods are fairly similar based on the KS plots.

Another approach for comparing goodness-of-fit between models is to examine the proportion of deviance in a null model that is explained by each, which is expressed by

$$R = \frac{D_N - D_M}{D_N}. \tag{9}$$

where $D_M$ and $D_N$ denote the deviance of the proposed model and the null model respectively. The value of $R$ shows the ratio of the deviance explained by the proposed model that the null model fails to explain. This measure does
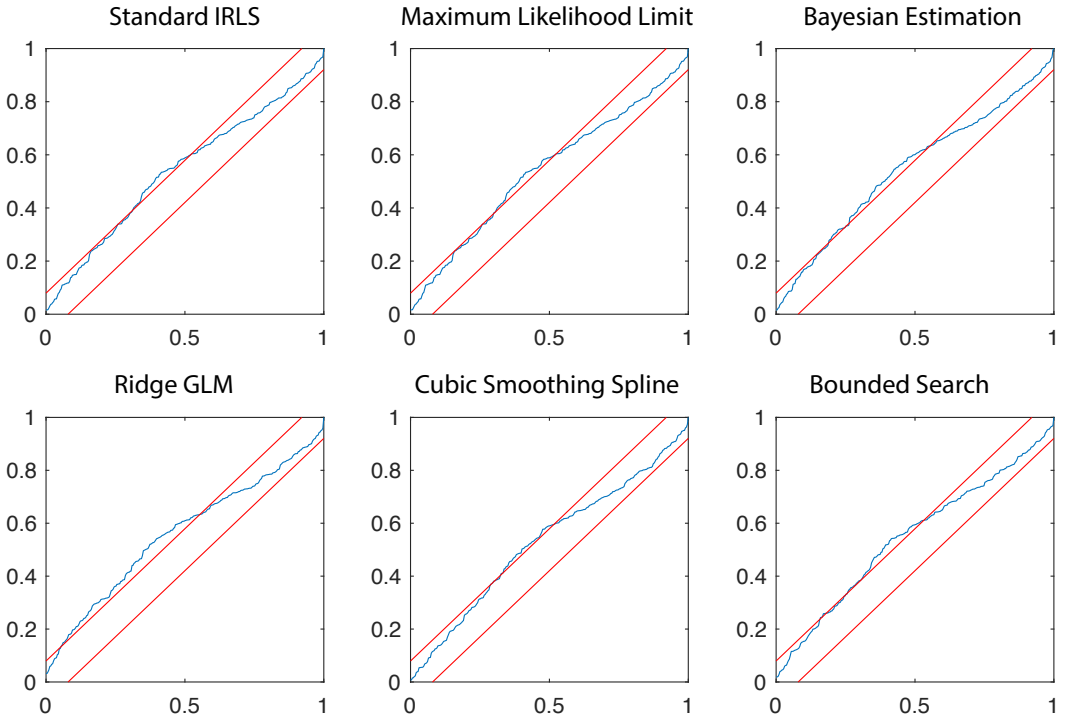
**FIGURE 1** KS plots for six different fitting approaches. The spike times were transformed via the time-rescaling theorem [35, 36] based on each fitted model. The blue line is the KS plot and the red lines represent global 95% confidence bounds for a well fit model. The Cubic Smoothing Spline method shows the smallest deviation outside of the bounds and the Ridge GLM has the largest, suggesting the best and the worst performance respectively.

not attempt to account for over-fitting, and tends to be higher for larger models. Therefore, we also consider a cross-validated version of $R$, denoted by $R_C$, which can directly be used to assess the prediction power of the fitted models. A statistical comparison between different models based on their deviances requires taking into account the effective degrees of freedom (d.o.f.) of each model as well. Among various measures introduced for the effective d.o.f., we use the approach that to our knowledge is the most prevalent. This approach defines the effective d.o.f to be the trace of the hat matrix $H$, which is satisfying $HY = \hat{Y}$ where $\hat{Y}$ denotes the estimate of the response vector [28]. In order to make the comparison easier, we divide the effective d.o.f. for each method by that for Standard IRLS, and call it effective d.o.f. ratio. Additionally, we compare the different methods based on the computational resources they require. To do so, we compute the relative run time and the relative peak memory of each method with respect to standard IRLS on the same system. All models have been fit 10 times, each using a different pair of trials from the training set, and mean values are computed.

Table 1 summarizes the comparisons between different methods, in terms of goodness-of-fit and consumed computational resources. Of all the methods, the Standard IRLS and Maximum Likelihood Limit methods have the largest $R$ values for the data used to fit the model, and the Cubic Smoothing Spline has the smallest. However, these methods do not generalize as well to other datasets, as they have the smallest $R_{CV}$ values. In fact, $R_{CV}$ for the standard IRLS approach is on average negative, indicating a poorer fit than a null, homogeneous Poisson model of the data (see Eq.

**TABLE 1** Summary of goodness-of-fit measures and computational resource consumption for different estimation methods. All models have been fit 10 times, each time using a different pair of trials of the training set, and mean values are computed.

| Method | $R$ | $R_{CV}$ | Number of Parameters | Effective d.o.f. Ratio | Relative Run Time | Relative Peak Memory |
|---|---|---|---|---|---|---|
| Standard IRLS | 0.3404 | < 0 | 206 | 1 | 1 | 1 |
| Maximum Likelihood Limit | 0.3404 | 0.1479 | 175 | 0.83 | 0.07 | 0.07 |
| Bayesian Estimation | 0.3227 | 0.2827 | 206 | 0.45 | 0.16 | 0.01 |
| Ridge GLM | 0.3236 | 0.2319 | 206 | 0.79 | 1.95 | 0.05 |
| Cubic Smoothing Spline | 0.2612 | 0.2418 | 40 | 0.18 | 0.09 | 0.18 |
| Bounded Search | 0.3397 | 0.2124 | 206 | 0.86 | 0.15 | 0.01 |

9). This suggests that the effect of leaving the perfect predictors in the model is to overfit the observed data, with large perfect parameter values that make a few observed non-spike intervals slightly more likely, but lead to extremely poor fits when future datasets include spikes associated with nonzero values of these predictors. On the other hand, all of the other methods impose some kind of constraint on the estimated parameters, e.g. via a prior distribution in Bayesian Estimation or via smoothing in the Cubic Smoothing Spline approach, which prevents the sampling perfect parameters from diverging. We observe that Cubic Smoothing Splines and Bayesian Estimation, which have noticeable smaller effective d.o.f ratios, have better predictive power (larger $R_{CV}$) than other methods. The Ridge GLM is the slowest method (largest relative run time), which is due to its exhaustive search to find an optimal value for the tuning parameter $\Lambda$. Cubic Smoothing Splines, despite the fact that it has the smallest effective d.o.f ratio, shows the largest relative peak memory after Standard IRLS. This is because this method needs to construct and store a new design matrix before running the IRLS algorithm.

Another issue that accompanies the complete separation or perfect prediction problem is that the observed Fisher information at the maximum likelihood solution typically does not accurately reflect the variance-covariance structure of the parameter estimates. In particular, the observed Fisher information is often close to singular, and if inappropriately used to construct confidence bounds and perform hypothesis tests, typically leads to extremely large standard error estimates for the perfect predictors and inaccurate covariance between the perfect predictors and all other signals [19, 26, 27]. Below, we examine how different each of the methods deals with this problem.

Bootstrapping represents one alternative to the Fisher information for computing standard error estimates for perfect predictors for the Standard IRLS estimator. In this approach, random subsets of the original data are used to fit the model multiple times, and then the empirical distribution of estimated values is used to compute standard error estimates for all parameters. Bootstrapping is far more computationally expensive than the other proposed methods, but results in more robust standard error estimates for IRLS, and therefore is considered as our benchmark here. The estimated values, $\hat{\beta}$, along with their confidence intervals using Standard IRLS, Bootstrapping, Bayesian Estimation, and the Cubic Smoothing Spline approaches are shown in Fig. 2. As expected, the bootstrap approach results in no standard error estimates for structural perfect parameters, as they always diverge to $-\infty$. The bootstrap intervals are tighter than those obtained by the Fisher Information for IRLS for nearly all parameters, suggesting that the existence of perfect predictors can influence uncertainty estimates in the IRLS procedure even for non-perfect predictor parameters. Both Bayesian Estimation and Cubic Smoothing Spline methods result in relatively narrow confidence intervals which provide smoother estimates than bootstrapping IRLS. We observed that Ridge GLM and Bounded Search obtain standard error estimates that are artificially larger than what was obtained by Bootstrapping,
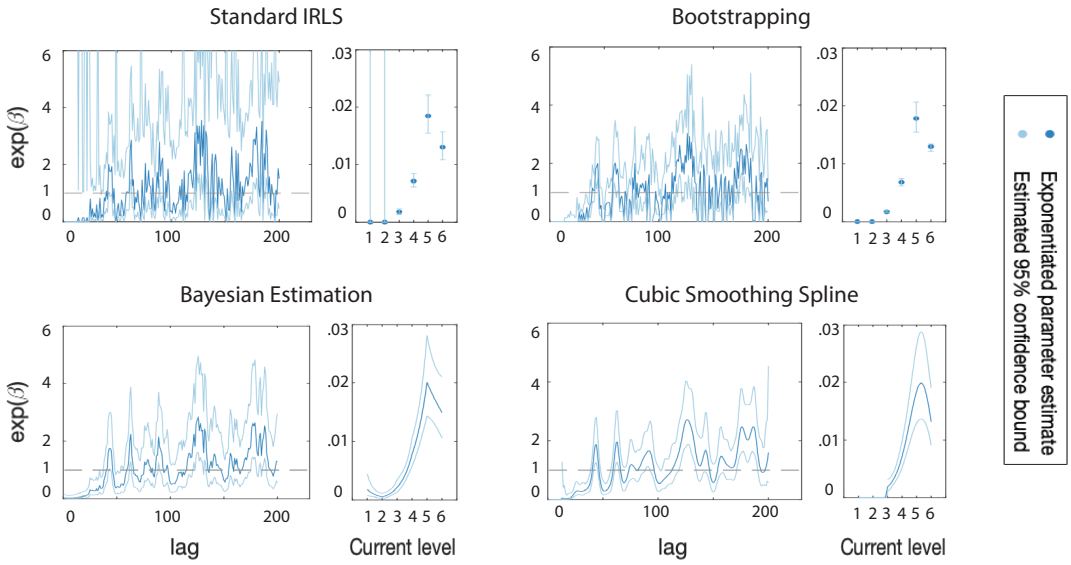
**FIGURE 2** Estimated parameter values, $\hat{\beta}$, and estimated confidence intervals obtained by four different approaches: Standard IRLS, Bootstrapping, Bayesian Estimation, and Cubic Smoothing Splines. For each approach, the estimated influence of past spiking is shown in the left panel and the estimated influence of the input current is shown in the narrower right panel. Thick blue lines or dots represent the exponentiated parameter estimates and thin blue lines and bars represent the estimated 95% confidence bounds.

especially for structural perfect parameters. The Maximum Likelihood Limit approach doesn't provide an estimation procedure for standard error of perfect parameters, as they are completely removed from the model prior to fitting.

In Fig. 3 in order to examine how different methods estimate the correlation between parameters, we concentrate on the first 100 predictors, which are associated with the influence of past spike counts at lags of 1-100 ms, and illustrate the estimated correlation between one fixed predictor and all other predictors. For the fixed predictor, we consider a sampling perfect predictor (left column), a non-perfect predictor (middle column), and a putative structural perfect predictor associated with the neuron's refractory period (right column). We observe that for non-perfect predictors, all methods perform more or less similarly, with Bayesian Estimation providing a smoother version of the correlation structure estimated by the Ridge GLM and Standard IRLS methods (middle column). In this case, we can see a local correlation structure for parameters around the fixed predictor (60-th predictor), which is expected; the effect of firing activity at lag $i$ and $i+1$ is likely to be correlated. For the sampling perfect predictor (left column), we see that Standard IRLS results in zero correlation for all non-perfect predictors. However, we know that this predictor is perfect only due to sampling, and hence we expect to see a temporal structure similar to what we observed between two non-perfect predictors in the middle column. Ridge GLM shows this local correlation structure weakly, but with substantially reduced effect size and statistical power. On the other hand, the correlation estimated by Bayesian GLM is much more aligned with what we expect to see, and clearly shows a temporal correlation structure similar to that between non-perfect predictors. For the putative structural perfect parameter, Standard IRLS again estimates all the correlations to non-perfect predictors to be zero. In this case, this might be the preferred inference; if it is impossible for the neuron to fire during its refractory period, this should be true regardless of the influence of a spike beyond the refractory period. In this case, Ridge GLM also gives very small correlations, but Bayesian GLM still estimates a
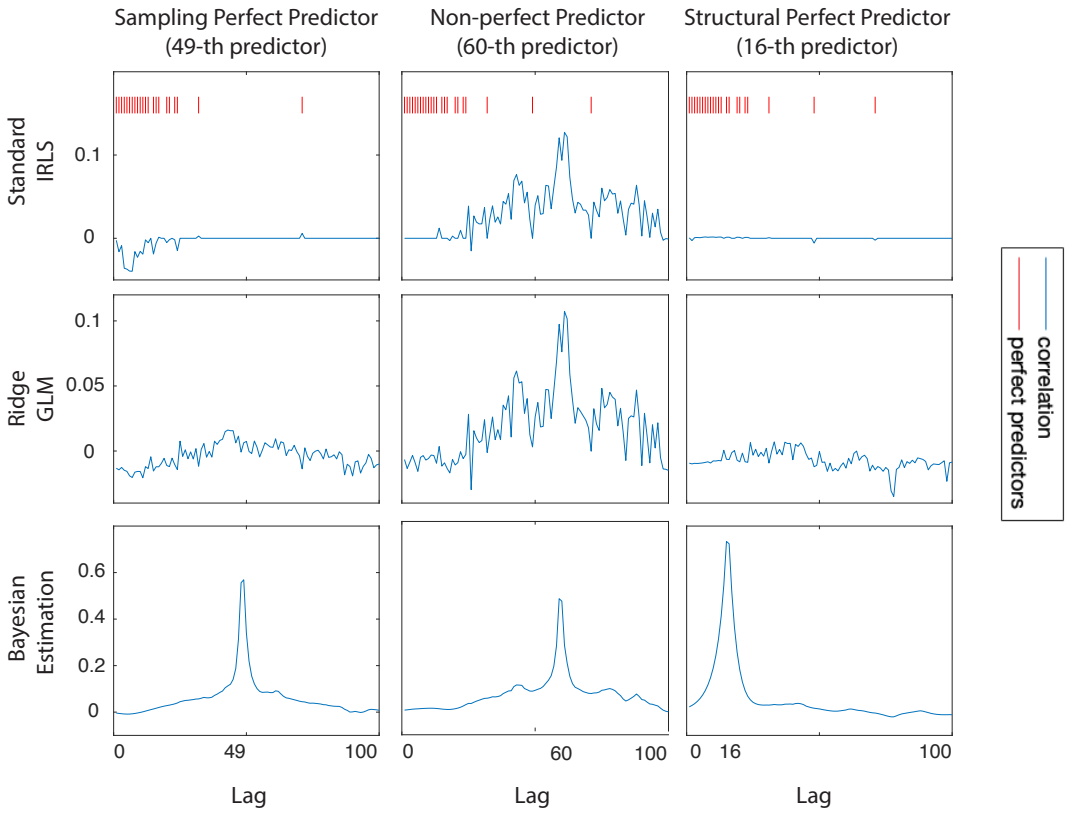
**FIGURE 3** Estimated correlation between a fixed predictor and all other predictors associated with the past history of spike counts at lags 1-100ms. This is shown for a sampling perfect predictor (left column), a non-perfect predictor (middle column), and a putative structural perfect predictor (right column), when Standard IRLS (first row), Ridge GLM (second row), and Bayesian Estimation (third row) were used. The red ticks in the top panel represent the lags associated with perfect predictors.

smooth correlation structure, consistent with the prior distribution. Note that we could use our knowledge of the refractory period of the neuron to adjust the prior to prevent this correlation. Though not shown in this figure, the Cubic Smoothing Spline estimates correlations very similar to the Bayesian Estimation method, and the Bounded Search estimates are very close to those obtained from Ridge GLM. The Maximum Likelihood Limit approach, again, does not provide correlation estimates between any perfect predictor and any other predictor, as it removes all perfect predictors from the model prior to fitting.

## 4 | DISCUSSION

In this paper, we investigated the problem of complete separation in GLMs, which leads to the failure of the IRLS algorithm to converge when a single predictor or a linear combination of multiple predictors completely separate a response from a non-response. This occurs frequently in point process GLMs due first, to structural properties of

neurons and their receptive fields that prevent a response, e. g. refractoriness, and second, incomplete sampling of the signals that influence spiking leading to no observed neural responses. This phenomenon can have a substantial impact on modeling and statistical inference from neural data. We broadly presented various classes of approaches to deal with this problem (section 2), and illustrated how they compare in practice (section 3). Here, we further discuss some of the advantages and disadvantages that are associated with each family of approaches.

**Standard IRLS**

**Advantages.** Already implemented in many statistical software packages. Describes the existing data set optimally. Inference for non-perfect parameters is typically reliable.

**Disadvantages.** Very slow; continues to run until iteration limit reached. Fitted models generalize poorly to other datasets. Poor estimates of variance-covariance structure for sampling perfect parameters make inferences about these parameters challenging.

**Maximum Likelihood Limit**

**Advantages.** Fast. Describes the existing data optimally. Inference for non-perfect parameters is typically reliable.

**Disadvantages.** Fitted models generalize poorly to other datasets. Not typically implemented in statistical software packages. No explicit computation of the Fisher information for estimating the variance-covariance structure of the perfect parameters.

**Bayesian Estimation**

**Advantages.** Provides increased statistical power and prediction accuracy when appropriate priors are available. Allows for precise estimation of confidence bounds and covariance between parameters. Prior can be chosen to distinguish between covariates that may lead to structural or sampling perfect predictors.

**Disadvantages.** Requires methods for selecting appropriate priors. Results may be sensitive to the choice of prior.

**Regularization Methods**

**Advantages.** Already implemented in many software packages. Controls perfect parameters with minimal impact on non-perfect parameters.

**Disadvantages.** Sensitive to tuning parameter, which typically involves an exhaustive search to find an optimal value. This makes this method relatively slow. Can result in diminished statistical power (compared to Bayesian Estimation and Change of Basis) since it treats structural and sampling perfect predictors identically.

**Change of Basis**

**Advantages.** Leads to more parsimonious models that take advantage of known structure in receptive field models. With a judicious choice of model structure, this can lead to a substantial increase in statistical power and model interpretability and reductions in computational time. Prior knowledge can be used to select a model structure that distinguishes structural and sampling perfect predictors.

**Disadvantages.** Selecting an appropriate model structure can be challenging. Some reparameterizations may not solve the perfect prediction problem. The results may be quite sensitive to the choice of model structure. Inferences may require inversion of transformations associated with reparameterization.

**Modifying Search Criterion**

**Advantages.** Has minimal impact on non-perfect parameters.

**Disadvantages.** Results may be sensitive to the choice of criterion. It can result in diminished statistical power (compared to Bayesian Estimation and Change of Basis) since it treats structural and sampling perfect predictors identically.
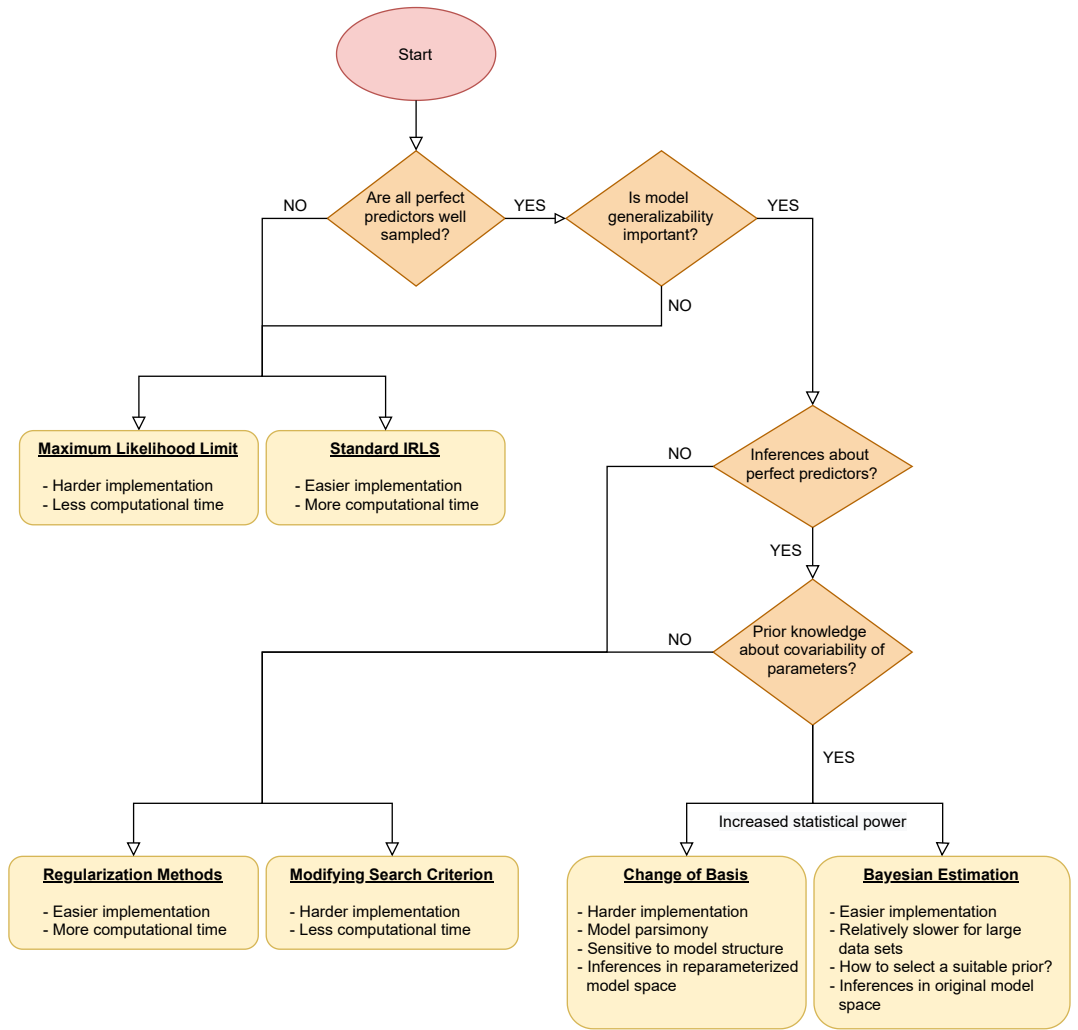
**FIGURE 4** Roadmap for selecting an approach for dealing with perfect prediction, based on the features and goals of the particular analysis question. Each path ends with two approaches to consider with properties that are well-aligned to the resulting analysis issues. The decision between these two can be made according to the pros and cons given for each method (yellow boxes).

Knowing the potential advantages and disadvantages of each of these approaches can help modelers select one, or a subset of methods to use to explore their data. We can also use these to develop a sequence of questions that modelers might ask to help guide them to a decision about which methods might work best for a particular analysis problem. We present an example of such a modeling 'roadmap' in Fig. 4, in flowchart form. Note that this represents just one potential approach to selecting a method; there may be many other roadmaps, depending on what particular set of questions might arise in the modeling process. In the road map depicted in Fig. 4, the central questions focus on how well the signals that influence spiking are sampled, whether the fitted model needs to generalize to other

datasets, whether the perfect predictors relate to the scientific questions to be addressed, and whether there is prior knowledge about receptive field structure to constrain the model. Each path ends with two approaches to consider with properties that are well-aligned to the resulting analysis issues. The decision between these two can be made according to the pros and cons given for each method (yellow boxes). For example, if we had a model for a place cell using indicator functions over space, the Bayesian Estimation approach might be most useful if we have a sensible prior and the scientific questions are focused on inferences about individual parameters that represent coding over a small subregion of the place field. If sensible priors are hard to come by, or if the scientific questions are more focused on more global properties of the place field (e.g. its center and width) than a smoothed estimate based on a change to a spline basis might be more appropriate.

Perfect predictors can arise in neural data due to many different factors, such as structural properties of the receptive field, lack of data for some regions of the receptive field, or even features of the receptive field that cannot be sampled under some experimental conditions. In this paper, we focused on point process GLMs, but complete separation can also happen in other domains of neuroscience. In particular, GLMs are used for Binomial and multinomial data which occur quite often when one looks at behavioral measures in neuroscience experiments. The ideas presented in this paper are not only limited to point process GLMs, and can directly be applied to other studies that involve other kinds of GLMs. Neuroscience experiments going forward are likely to yield larger data sets with many more neurons, and potentially with receptive field structures that involve many predictors simultaneously. As a result, models needed to explain such large and complex data sets are more likely to encounter the problem of complete separation. We anticipate that the types of methods and decision procedures discussed in this paper are going to become more and more important as the experiments and data sets evolve to capture more complex structures of the stimuli.

## acknowledgements

## references

[1] Paninski L, Pillow J, Lewi J. Statistical models for neural encoding, decoding, and optimal stimulus design. Progress in Brain Research 2007;165:493–507.

[2] Paninski L, Brown EN, Iyengar S, Kass RE. Statistical models of spike trains. Stochastic Methods in Neuroscience 2008;.

[3] Eden UT, Kass RE. Statistical models of spike train data. Neuroscience in the 21st Century 2016;2:3137–3151.

[4] Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. Journal of Neurophysiology 2005;93(2):1074–1089.

[5] Chen Z, Brown EN. Generalized linear models for point process analyses of neural spiking activity. Jaeger D, Jung R (eds) Encyclopedia of Computational Neuroscience 2013;p. 1–4.

[6] Keat J, Reinagel P, Reid RC, Meister M. Predicting every spike: a model for the responses of visual neurons. Neuron 2001;30(3):803–817.

[7] Paninski L, Simoncelli EP, Pillow JW. Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. Neural Comput 2004;16(12):2533–2561.

[8] Paninski L. Maximum likelihood estimation of cascade point-process neural encoding models. Network: Computation in Neural Systems 2004;15(4):248–262.

[9] McCullagh P, Nelder JA. Generalized linear models (second edition). ISBN: 0-412-31760-5 1989;Chapman and Hall.

[10] Fisher RA. The case of zero survivors (Appendix to Bliss, C.I.(1935)). Ann Appl Biol 1935;22:164–165.

[11] Green PJ. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). J R Statist Soc 1984;B 46:149–192.

[12] Bryson MC, Johnson ME. The incidence of monotone likelihood in the Cox model. Technometrics 1981;23(4):381–383.

[13] Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984;71(1):1–10.

[14] Santner TJ, Duffy DE. A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 1986;73(3):755–758.

[15] Lesaffre E, Albert A. Partial separation in logistic discrimination. Journal of the Royal Statistical Society, Series B 1989;51(1):109–116.

[16] Clarkson DB, Jennrich RI. Computing extended maximum likelihood estimates for linear parameter models. Journal of the Royal Statistical Society, Series B 1991;53(2):417–426.

[17] E EL, Marx BD. Collinearity in generalized linear regression. Communications in Statistics—Theory and Methods 1993;22(7):1933–1952.

[18] Kolassa JE. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. Scandinavian Journal of Statistics 1997;24(4):523–530.

[19] Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statist Med 2006;25:4216–4226.

[20] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statist Med 2002;21:2409–2419.

[21] D DF. Bias reduction, the Jeffreys prior and GLIM. Fahrmeir L, Francis B, Gilchrist R, Tutz G (eds) Advances in GLIM and Statistical Modelling Lecture Notes in Statistics 1992;78:91–100.

[22] Firth D. Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. Dodge Y, Whittaker J (eds) Computational Statistics 1992;1:553–557.

[23] Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80(1):27–38.

[24] Nagashima K, Sato Y. Information criteria for Firth's penalized partial likelihood approach in Cox regression models. Stat Med 2017;36(21):3422–3436.

[25] Heinze G, Schemper M. A Solution to the Problem of Monotone Likelihood in Cox Regression. Biometrics 2001;57:114–119.

[26] Hauck Jr WW, Donner A. Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association 1977;72(360):851–853.

[27] Vaeth M. On the use of Wald's test in exponential families. International Statistical Review 1985;53(2):199–214.

[28] Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. Springer Series in Statistics 2001;ISBN: 9780387848587.

[29] Boor CD. A Practical Guide to Splines. Applied Mathematical Sciences 1978;27:ISBN: 978–0–387–95366–3.

[30] Lin Y, Zhang HH. Component selection and smoothing in smoothing spline analysis of variance models. Annals of Statistics 2006;34(5):2272–2297.

[31] Pillow JW, J JS, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 2008;454:995–999.

[32] Gerstner W, Naud R. How Good are Neuron Models? Science 2009;326(5951):379–380.

[33] Gerstner W, Naud R, Carandini M, Sincich LC, Horton JC, Adams DL, et al. Quantitative Single-Neuron Modeling Competition for year 2009, developed by researchers at EPFL; data from Wistar rat cortical neurons and rhesus monkey lateral geniculate nucleus and retinal ganglion cells. CRCNSorg;.

[34] Dimatteo I, Genovese CR, Kass RE. Bayesian curve-fitting with free-knot splines. Biometrika 2001;88(4):1055–1071.

[35] Papangelou F. Integrability of expected increments of point processes and a related random change of scale. Trans Amer Math Soc 1972;165(1972):483–506.

[36] Brown EN, Barbieri R, Ventura V, Kass RE, Frank LM. The time-rescaling theorem and its application to neural spike train data analysis. Neural computation 2002;14(2):325–346.

# 5 | APPENDIX

## 5.1 | Cardinal Spline Reparameterization

In this section, we explore a specific example of basis expansion methods, cardinal spline reparameterization, and explain in detail how it can solve the problem of perfect predictors. This basis expansion method is perfectly suitable for cases where predictors are a set of sequential indicator functions, e.g. in a history-dependent model where the $k$-th predictor is the spiking counts at lag $k$. Following the notations introduced in Eq. 7, the $k$-th basis function is given by

$$g_k(x_{j.}) = \sum_{i=1}^{p} x_{ji} s_{ik} \tag{10}$$

where $s_{ik}$ is the $ik$-th element of matrix $S$ that is constructed as follows. First, an increasing sequence of length $q - 4$ from interval $(1, p)$ is augmented by 1, $p$, one value less than 1, and one value greater than $p$ to yield the sequence $\zeta_1 < \zeta_2 < \cdots < \zeta_q$ which is called the set of control points or knots. Then, in order to construct the $j$-th row of $S$, index $i$ is selected such that $\zeta_i \leq j < \zeta_{i+1}$ and $\alpha = \frac{j - \zeta_i}{\zeta_{i+1} - \zeta_i}$ is computed. All elements of $S_{j.}$ are set to zero except the four consecutive elements $\{s_{j,i-1}, s_{j,i}, s_{j,i+1}, s_{j,i+2}\}$ that are given by

$$\begin{bmatrix} s_{j,i-1} & s_{j,i} & s_{j,i+1} & s_{j,i+2} \end{bmatrix} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \end{bmatrix} \begin{bmatrix} -t & 2-t & t-2 & t \\ 2t & t-3 & 3-2t & -t \\ -t & 0 & t & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{11}$$

where $0 \leq t \leq 1$ is called the tension parameter and must be picked prior to constructing $S$. In order to find the relation between original parameters ($\beta_i$ in Eq. 1b) and parameters after changing basis ($\tilde{\beta}_i$ in Eq. 7), we start by simplifying the right hand side of Eq. 7.

$$\lambda_j = exp\Big(\tilde{\beta}_0 + \sum_{k=1}^{q} g_k(x_{j.})\tilde{\beta}_k\Big) = exp\Big(\tilde{\beta}_0 + \sum_{k=1}^{q} \tilde{\beta}_k\Big(\sum_{i=1}^{p} x_{ji}s_{ik}\Big)\Big)$$

$$= exp\Big(\tilde{\beta}_0 + \sum_{i=1}^{p} x_{ji}\Big(\sum_{k=1}^{q} s_{ik}\tilde{\beta}_k\Big)\Big). \tag{12}$$

Comparing Eq. 1b and Eq. 12 and noting that the second sum in Eq. 12 is the $i$-th element of vector $S\beta$ reveals the relation between $\beta$ and $\tilde{\beta}$:

$$\beta_0 = \tilde{\beta}_0, \quad \beta_{1:p} = S\tilde{\beta}_{1:q} \tag{13}$$

Based on this relation, any new predictor is a linear combination of original predictors, and as long as some non-perfect predictor appears with nonzero weights in this combination, the new predictor will not be perfect. This can be assured to happen by choosing the knots such that there exist at least one non-perfect predictor between every two successive knots.