Bayesian Approach to Point-Process Generalized Linear Models

M. Rule

January 14, 2012

1 Background

In neuroscience, we are interested in the problem of how neurons encode, process, and communicate information. Neurons communicate over long distances using brief all-or-nothing events called spikes. We are often interested in how the spiking rate of a neuron depends on other variables, such as stimuli, motor output, or other ongoing signals in the brain.

To model this, we consider spikes as events that occur at a point in time with an underlying variable rate, or conditional intensity, λ. There are many approaches to estimating λ. These notes cover point-process generalized linear models, and Bayesian approaches. These are closely related, and in some cases the same thing.

1.0.1 Point-process generalized linear models

In point-process Generalized Linear Models (GLM) we estimate $\ln(\lambda)$ (or $\text{logit}(\lambda)$), which can take on any real value. The GLM estimates $\ln(\lambda)$ directly as a linear combination $\ln(\lambda) = \sum_i a_i x_i$ of some variables x_i. The coefficients a are optimized using likelihood.

For computational tractability we work with a discretized version of the point process. A point process can be converted into a series of non-negative integers by counting the number of events y that occur in a fixed time interval Δ. If Δ is small, such that λ is approximately constant within the interval, then the distribution of this new count process can be written as:

$$\Pr(y{=}k) = (\lambda\Delta)^k(1-\lambda\Delta)^{1-k},\ k \in \mathbb{Z}$$

If we choose Δ sufficiently small such that $\Pr(y{>}1) \approx 0$, we can restrict analysis to the approximation

$$\Pr(y{=}1) = \lambda\Delta,\ \Pr(y{=}0) = 1-\lambda\Delta$$

For an alternative derivation, consider the Poisson distribution, which defines the probability of observing k events in time Δ where the expected number of events is $\lambda\Delta$:

$$\Pr(y{=}k) = \frac{(\lambda\Delta)^k e^{-\lambda\Delta}}{k!}$$

In the limit of Δ sufficiently small, $\Pr(y > 1)$ becomes negligible, and we consider only

$$\Pr(y{=}1) = (\lambda\Delta)e^{-\lambda\Delta}$$

Since Δ is small, $e^{-\lambda\Delta} \approx 1$, and we have

$$\Pr(y{=}1) \approx \lambda\Delta$$

Since $\Pr(y{=}0)$ can be computed from $\Pr(y{=}1)$ in this case, we focus on estimating $\Pr(y{=}1)$, which we will denote, for convenience, as $\Pr(1)$ or simply P. Note that the log-linear model also works in the discrete case, since multiplication of λ by Δ can be absorbed into a constant term in the model.

$$\Pr(1) \approx \ln(\lambda\Delta) = \sum_i a_i x_i$$

For now, restrict analysis to a single covariate x. In the event that the relationship between $\ln(\lambda\Delta)$ and x is not entirely captured by a log-linear model, we may want to add nonlinear functions of x to the model. In this way, we can capture a wider range of possible relationships between x and $\ln(\lambda\Delta)$. However, it is sometimes challenging to select nonlinear terms, and adding irrelevant terms can increase the time required to fit the model, and increase over-fitting. The general form of this case is given by

$$\Pr(1) \approx \ln(\lambda\Delta) = \sum_i a_i f_i(x)$$

1.0.2 Connection between GLM and Bayesian approach

We can also derive a model for conditional intensity using Bayes rule. We are interested in learning how $\Pr(1)$ might depend on another variable x. That is, we would like to know how the conditional intensity $\lambda_x\Delta \approx \Pr(1|x)$ differs from baseline $\Pr(x)$. We can apply Bayes rule to directly solve for this conditional distribution in terms of more easily observed $\Pr(x|1)$:

$$\lambda_x\Delta \approx \Pr(1|x) = \Pr(x|1)\frac{\Pr(1)}{\Pr(x)}$$

For computational efficiency we work with the natural logarithm of probability. This yields an expression that allows us to directly estimate conditional intensity in terms of the log probability density functions (PDF) of x and $x|1$.

$$\ln(\lambda_x\Delta) \approx \ln(\Pr(1|x)) = \ln(\Pr(x|1)) - \ln(\Pr(x)) + \ln(P) \tag{1}$$

The performance of this approach depends on accurate modeling of $\Pr(x|1)$ and $\Pr(x)$. If these quantities follow distributions whose parameters can be estimated quickly from available data, this approach can be faster than likelihood maximization of the more general log-linear model. If the expression for the log-likelihood can be factored into a linear combination of fixed nonlinear functions of x, then we can directly relate the Bayesian approach to the log-linear model :

$$\ln(\Pr(1|x)) = \ln(\Pr(x|1)) - \ln(\Pr(x)) + \ln(P) = \sum_i a_i f_i(x) \tag{2}$$

While we can always approximate the above using a series expansion of $\ln(\Pr(1|x))$, a closed form solution of this relationship exists if the log PDFs of $\Pr(x|1)$ and $\Pr(x)$ can each be factored as $\sum_i a_i f_i(x)$, even if $\Pr(x|1)$ and $\Pr(x)$ follow different distributions.

We show in the next section that distributions from the exponential family meet these conditions, allowing direct solution of the parameters to a log-linear model from data statistics. Choosing a parametric distribution for $\Pr(x|1)$ and $\Pr(x)$ also determines the set of nonlinear terms required for a log-linear model to capture the relationship between x and $\ln(\lambda)$, as well as their weights.

2 Selecting nonlinear features for a GLM point process model

In the previous section we discussed a direct solution for the conditional intensity using Bayes rule, and how this solution might relate to log-linear models. Specific examples are given in this section. The general steps are as follows :

- Look at the marginal distribution of the covariate $\Pr(x)$, and the distribution conditioned on the presence of a spike $\Pr(x|1)$.
- Pick a parametric probability distribution family that fits the observed distributions well.
- Look up the log probability-density function.
- Write down the log-likelihood ratio in terms of the log PDF as in (**??**).
- Expand this function until it is of the form $\sum_i a_i f_i(x)$ where a_i is a real valued parameter and f_i is a function of the covariate.
- The functions $f_i(x)$ are the nonlinear features of x that you should include in a log-linear model, and the coefficients a_i provide an approximate fit of that model.

2.1 Exponential data imply a log-linear model

If a covariate follows an exponential distribution, the Bayesian method provides parameters for a log-linear point process model. The probability density for an exponential distribution has one scale parameter λ, $\Pr(x;\lambda_x) = \lambda_x e^{-\lambda_x x}$, which gives the log probability as $\ln(\Pr(x;\lambda_x)) = \ln(\lambda_x) - \lambda_x x$. Substituting this into equation (**??**) yields

$$\ln(\lambda\Delta) \approx \left(\ln(\lambda_{x|1}) - \lambda_{x|1} x\right) - (\ln(\lambda_x) - \lambda_x x) + \ln(P)$$

Collecting terms yields the the linear and constant terms in a log linear inhomogenous poisson process model:

$$\ln(\lambda\Delta) \approx \left(\lambda_x - \lambda_{x|1}\right) x + \left(\ln(\lambda_{x|1}) - \ln(\lambda_x) + \ln(P)\right)$$

2.2 Normally distributed data imply a log-quadratic model

If a covariate follows a Gaussian distribution, the Bayesian classifier provides parameters for a log-quadratic point process model $\ln(\lambda) = ax^2 + bx + c$. The probability density for a Gaussian variable is

$$\mathcal{N}(\mu, \sigma)(x) = \exp[-(x-\mu)^2/(2\sigma^2)]/(\sigma\sqrt{2\pi}),$$

which gives the log-density

$$\ln(\mathcal{N}(\mu, \sigma)(x)) = -(x-\mu)^2/(2\sigma^2) - \ln(\sigma) - \ln(2\pi)/2.$$

Substituting this into equation (**??**) and ignoring the $\ln(2\pi)/2$ terms, which immediately cancel, gives :

$$\begin{aligned}\ln(\lambda\Delta) = &\frac{1}{2\sigma_x^2}(x-\mu_x)^2 + \ln(\sigma_x) \\ &- \left(\frac{1}{2\sigma_{x|1}^2}(x-\mu_{x|1})^2 + \ln(\sigma_{x|1})\right) \\ &+ \ln(P)\end{aligned}$$

Expanding the quadratic expressions and collecting terms yields the constant, linear, and quadratic terms for a log-quadratic inhomogeneous Poisson model:

$$\begin{aligned}\ln(\lambda\Delta) = &\left(1/(2\sigma_x^2) - 1/(2\sigma_{x|1}^2)\right) x^2 \\ &+ \left(\mu_{x|1}/\sigma_{x|1}^2 - \mu_x/\sigma_x^2\right) x \\ &+ \mu_x^2/(2\sigma_x^2) - \mu_{x|1}^2/(2\sigma_{x|1}^2) + \ln(\sigma_x) - \ln(\sigma_{x|1}) + \ln(P)\end{aligned}$$

If the covariate x has been z-scored such that $\sigma_x = 1$ and $\mu_x = 0$, the expression simplifies to:

$$\begin{aligned}\ln(\lambda\Delta) = &\left(1/2 - 1/(2\sigma_{x|1}^2)\right) x^2 \\ &+ \left(\mu_{x|1}/\sigma_{x|1}^2\right) x \\ &+ \ln(P) - \mu_{x|1}^2/(2\sigma_{x|1}^2) - \ln(\sigma_{x|1})\end{aligned}$$

2.3 Gamma distributed data imply $\ln(x)$ nonlinear features

The Gamma PDF in terms of a shape α and inverse scale β parameter is

$$\Pr(x; \alpha, \beta) = \beta^\alpha x^{\alpha-1} \exp(-x\beta)/\Gamma(\alpha),$$

which gives the log PDF

$$\ln(\Pr(x; \alpha, \beta)) = \alpha \ln(\beta) + (\alpha - 1)\ln(x) - \beta x - \ln(\Gamma(\alpha)).$$

Substituting this into equation (**??**) yields

$$\begin{aligned}\ln(\lambda\Delta) = &\left(\alpha_{x|1}\ln(\beta_{x|1}) + (\alpha_{x|1} - 1)\ln(x) - \beta_{x|1}x - \ln(\Gamma(\alpha_{x|1}))\right) \\ &- (\alpha_x \ln(\beta_x) + (\alpha_x - 1)\ln(x) - \beta_x x - \ln(\Gamma(\alpha_x))) \\ &+ \ln(P)\end{aligned}$$

Collecting terms yields a fit for a model that, in addition to linear and constant terms, includes a $log(x)$ term

$$\begin{aligned}\ln(\lambda\Delta) = &\left(\beta_x - \beta_{x|1}\right) x \\ &+ \left(\alpha_{x|1} - \alpha_x\right) \ln(x) \\ &+ \ln(\Gamma(\alpha_x)) - \ln(\Gamma(\alpha_{x|1})) + \alpha_{x|1} \ln(\beta_{x|1}) - \alpha_x \ln(\beta_x) \\ &+ \ln(P)\end{aligned}$$

2.4 Von Mises imply sine and cosine nonlinear terms

There is also a GLM formulation for von Mises distributed data. This is left left as an exercize to reader. To derive it, move the preferred phase parameter out of the cosine using trigonometric identities. The nonlinearities implied for the GLM are $\sin(\theta)$ and $\cos(\theta)$.

3 Generalization to the exponential family

The exponential family of distributions inclues all the distributions discussed so far in this section, and follows the canonical form

$$\Pr(x|\theta) = h(x)g(\theta)e^{\theta T(x)}$$

Where $x \in \mathbb{R}^n$, $n \in \mathbb{N}$ is an n dimensional real valued vector space, $\theta \in \mathbb{R}^m$, $m \in \mathbb{N}$ is an m dimensional real valued parameter, $h : \mathbb{R}^n \to \mathbb{R}$ maps from x to a real number, $g : \mathbb{R}^m \to \mathbb{R}$ maps from θ to a real number, and $T : \mathbb{R}^n \to \mathbb{R}^m$ maps from x to θ. This impies the canonical log PDF

$$\ln(\Pr(x|\theta)) = \ln(h(x)) + \ln(g(\theta)) + \theta T(x)$$

This form is already suitable for mapping onto a log linear point process model. The term $\ln(g(\theta))$ is a constant offset implied by the parameters, and present only for normalization. The term $\theta T(x)$ a weighted sum of functions of x, and the term $\ln(h(x))$ is also a function of x with weight 1. If $\theta = [\theta_1, ..., \theta_m]$, $x = [x_1, ..., x_n]$, and $T = [f_1(x), ..., f_m(x)]$, the log PDF can be written as:

$$\ln(\Pr(x)) = \sum_i a_i f_i(x) = \ln(g(\theta)) + \ln(h(x)) + \sum_{j=1}^{m} \theta_j f_j(x)$$

From this, the coditional intensity in the general case where x and $x|1$ may have different distributions from the exponential family is:

$$\begin{aligned}\ln(\lambda\Delta) &= \sum_i a_i f_i(x) \\ &= \left(\ln(g_{x|1}(\theta_{x|1})) + \ln(h_{x|1}(x)) + \sum_{j=1}^{m_{x|1}} \theta_{x|1,j} f_{x|1,j}(x)\right) \\ &- \left(\ln(g_x(\theta_x)) + \ln(h_x(x)) + \sum_{j=1}^{m_x} \theta_{x,j} f_{x,j}(x)\right) + \ln(P)\end{aligned}$$

If x and $x|1$ are both in the exponential family, it is always possible to pick a more general distribtuion from the exponential family that includes both x and $x|1$, and so we may consider

$$\begin{aligned}\ln(\lambda\Delta) &= \sum_i a_i f_i(x) \\ &= \left(\ln(g(\theta_{x|1})) + \ln(h(x)) + \sum_{j=1}^{m} \theta_{x|1,j} f_j(x)\right) \\ &- \left(\ln(g(\theta_x)) + \ln(h(x)) + \sum_{j=1}^{m} \theta_{x,j} f_j(x)\right) + \ln(P)\end{aligned}$$

The $\ln(h(x))$ terms cancel, although logarithmic terms may still be introduced via f_j

$$\begin{aligned}\ln(\lambda\Delta) &= \sum_i a_i f_i(x) \\ &= \left(\ln(g(\theta_{x|1})) + \sum_{j=1}^{m} \theta_{x|1,j} f_j(x)\right) \\ &- \left(\ln(g(\theta_x)) + \sum_{j=1}^{m} \theta_{x,j} f_j(x)\right) + \ln(P)\end{aligned}$$

Collecting terms

$$\begin{aligned}\ln(\lambda\Delta) &= \sum_i a_i f_i(x) \\ &= \left(\ln \frac{g(\theta_{x|1})}{g(\theta_x)} + \ln(P)\right) + \sum_{j=1}^{m} (\theta_{x|1,j} - \theta_{x,j}) f_j(x)\end{aligned}$$

If x and $x|1$ can be modeled by any distribution in the exponential family, we can write a log linear model directly from Bayes' rule. If the canonical form of the distribution has a closed form, we can also solve for the model weights.

4 Relationships determined by fitting a GLM can be simpler than those implied by a Bayesian approach

Although solving directly for conditional intensity using Bayes rule can imply a log-linear model and its coefficients, the converse is not in general true. For example, the quadratic terms from a Gaussian model vanish when $\sigma_x = \sigma_{x|1}$, reducing it to a simple log-linear model.

In general, a purely log-linear GLM will be able to fit the data if the distributions $\Pr(x)$ and $\Pr(x|1)$ differ only by a location parameter. For a particular choice of features $f(x)$, the GLM can be more accurate than a parametric Bayesian approach, because it can flexibly model data from a broader class of distributions.

Additionally, the GLM directly optimizes parameters that summarize the difference between $\Pr(x)$ and $\Pr(x|1)$, rather than finding parameters for each distribution separately. This makes

the GLM more statistically efficient, as less data is required to constrain a smaller number of parameters.

Nevertheless, inspecting the distributions $\Pr(x)$ and $\Pr(x|1)$ provides important clues for which nonlinear features $f(x)$ to include in a GLM regression.

5 Relationship to Likelihood Ratio Classification

Consider a continuous valued covariate x, and a discrete valued variable y that takes on k possible values. Variable y induces k classes on the values of x. Because of this, reconstructing y from x can be phrased a classification problem. For an observation of x', choosing the class for which x' is most likely gives us the maximum a posteriori estimate. In the special case of binary classification where $k = 2$, we can summarize this comparison with the likelihood ratio. Say we have two values or classes 0 and 1, a covariate x, and want to know the respective probabilites $\Pr(1|x)$ and $\Pr(0|x)$. The likelihood ratio is defined as $\Pr(1|x)/\Pr(0|x)$, and summarizes the relative likelihoods of the two possible classes. The likelihood ratio can be converted to a classification by thresholding. We can adjust the threshold to change the proportion of false positives and false negatives, or use the likelihood ratio directly in ROC analyis. We can use Bayes' rule to write down the likelihood ratio in terms of experimentally observable conditional distributions $\Pr(x|1)$ and $\Pr(x|0)$:

$$\Pr(1|x) = \Pr(x|1)\frac{\Pr(1)}{\Pr(x)},\ \Pr(0|x) = \Pr(x|0)\frac{\Pr(0)}{\Pr(x)}$$

The likelihood ratio can be written using the above expressions as:

$$\mathcal{L} = \frac{\Pr(1|x)}{\Pr(0|x)} = \frac{\Pr(x|1)\Pr(1)}{\Pr(x|0)\Pr(0)}$$

A point process may be approximated as a discrete Bernoulli process: rather than a list of event times, we consider a series of bins length Δ, where the probability that a bin contains an event is approximately $\lambda\Delta$. This creates a time series with two possible values : a bin contains at least one event or it does not. Let 0 denote bins that do not contain events, and 1 denote bins that contain at least one event. We can then apply a likelihood ratio classifier as a model of the point process.

How is the likelihood ratio related to the conditional intensity? For Δ sufficiently small to treat a point process as a Bernoulli process, events are extremely rare, and $\Pr(0) \to 1$ and $\Pr(x|0) \to \Pr(x)$ in the limit $\Delta \to 0$. Under these conditions the likelhood ratio is nearly the same as the conditional probability $\Pr(1|x)$. Therefore, we can say that $\mathcal{L} \sim \lambda\Delta_t$. In the limit $\Delta_t \to 0$, $\mathcal{L} = \lambda\Delta_t$.

$$\lim_{\Delta\to 0}\mathcal{L} = \lim_{\Delta\to 0}\left(\frac{\Pr(x|1)\Pr(1)}{\Pr(x|0)\Pr(0)}\right) = \Pr(x|1)\frac{\Pr(1)}{\Pr(x)} = \Pr(1|x) = \lambda\Delta$$

Likelihood ratio classification of a binned point process is approaches estimating the conditional intensity in the small Δ limit.

6 Kullback-Leibler divergence $D_{KL}(x|1 \parallel x)$ is an easily computed predictor model performance

Mutual information is a statistic used to summarize how related two variables are. Consider the problem of measuring the mutual information between a variable x and point process y.

Estimating the mutual information between a point process and a continuous covariate reduces to estimating the Kullback-Leibler divergence of the conditional $\Pr(x|1)$ from background $\Pr(x)$, both of which we have already computed in a Bayesian fit of the point process model. entropy, as:

$$I(x, y) = \mathbb{E}_y D_{KL}(x|y \parallel x)$$

Spikes are rare in a point process, so $\Pr(x|0) \approx \Pr(x)$. Since D_{KL} is zero if both distributions are identical, mutual information is (in the sparse limit) reduces to

$$I(x, y) = \Pr(y{=}1) D_{KL}(x|1 \parallel x)$$

The mutual information between x and y is the KL divergence of $\Pr(x|1)$ and $\Pr(x)$, multiplied by $\Pr(y = 1)$. Since $\Pr(y = 1)$ is a background term that depends on the choice of Δ, it is not especially relevant, except when comparing two point processes with different underlying rates.

Let's explore this in the case that $\Pr(x|1)$ and $\Pr(x)$ are normally distributed. Substituting $\Pr(x|1)$ and $\Pr(x)$ into the formula for the GL divergence between two Gaussian variables yields

$$D_{KL}(x|1 \parallel x) = \frac{(\mu_{x|1} - \mu_x)^2}{2\sigma_x^2} + \frac{1}{2}\left(\frac{\sigma_{x|1}^2}{\sigma_x^2} - 1 - \ln \frac{\sigma_{x|1}^2}{\sigma_x^2}\right)$$

If x has been z scored and has unit variance zero mean,

$$D_{KL}(x|1 \parallel x) = \frac{1}{2}\left(\mu_{x|1}^2 + \sigma_{x|1}^2 - 1 - \ln \sigma_{x|1}^2\right)$$

If we assume $\sigma_{x|1} \approx \sigma_x = 1$ (which is the implicit assumption if we were to fit a GLM with only the linear feature x), this simplifies further;

$$D_{KL}(x|1 \parallel x) = \frac{1}{2}\mu_{x|1}^2$$

From this we see that the fraction of information about y captured by x depends mainly on the squared deviation of the spike-triggered average $\mu_{x|1}$ from the baseline $\mu_x = 0$.

Incidentally, this also suggests that fancy Bayesian and GLM models might not tell you that much more than the Spike Triggered Average (STA), in some cases.

Another curious observation which holds empirically, at least for low information variables examined so far, is the relationship

$$2\ln(2\mathrm{AUC} - 1) + \frac{1}{2} = \ln(D_{KL}(x|1 \parallel x))$$

Where AUC is the area under the receiver operating characteristic (ROC) curve, and is used to summarize the accuracy of a point-process decoder. (I'm not sure whether this approximation is really valid or under what conditions it might hold)

7 Overall,

Models that fit the log-intensity of an inhomogeneous Poisson point process are related to likelihood-ratio based Bayesian classifiers. Solving for the conditional intensity using Bayes' rule, and distributions from the exponential family, is one way to find parameters for a point process model. The distribution family of x also provides clues as to which nonlinear features to incorporate into a point process GLM. The parametric Bayesian approaches also suggest simple closed-form formulae for measuring mutual information between a spike train and an external variable.