The empirical Bayes estimation of an instantaneous spike rate with a Gaussian process prior

Shinsuke Koyama
Department of Physics, Kyoto University
Department of Statistics and Center for Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213
koyama@stat.cmu.edu

Takeaki Shimokawa
Department of Physics, Kyoto University
Sakyo-ku, Kyoto 606-8502, Japan
shimokawa@ton.scphys.kyoto-u.ac.jp

Shigeru Shinomoto
Department of Physics, Kyoto University
Sakyo-ku, Kyoto 606-8502, Japan
shinomoto@scphys.kyoto-u.ac.jp

Abstract

We investigate a Bayesian method for systematically capturing the underlying firing rate of a neuron. Any method of rate estimation requires a prior assumption about the flatness of the underlying rate, which can be represented by a Gaussian process prior. This Bayesian framework enables us to adjust the very assumption by taking into account the distribution of raw data: A hyperparameter of the Gaussian process prior is selected so that the marginal likelihood is maximized. It takes place that this hyperparameter diverges for spike sequences derived from a moderately fluctuating rate. By utilizing the path integral method, we demonstrate two cases that exhibit the divergence continuously and discontinuously.

1 Introduction

Bayesian methods based on Gaussian process priors have recently become quite popular, and provide flexible non-parametric methods to machine learning tasks such as regression and classification problems [1, 2, 3]. The important advantage of Gaussian process models is the explicit probabilistic formulation, which not only provides probabilistic predictions but also gives the ability to infer hyperparameters of models [4].

One of the interesting applications of the Gaussian process is the analysis of neuronal spike sequences. In the field of neuroscience, we often wish to determine the underlying firing rate from observed spike sequences [5, 6, 7, 8, 9]. The firing rate may be a complex fluctuating function that sensitively reflects the detailed density of data points, or simple function that represents the gross features of the data points. The posterior distribution function representing the data is influenced by such prior assumption about the jaggedness of the underlying rate. In terms of the Bayesian framework, the prior assumption for the underlying rate can be represented by a Gaussian process representing the tendency of the rate to be relatively flat. According to the empirical Bayes method, a hyperparameter of the prior distribution is selected so that the marginal likelihood function is maximized [4, 10, 11, 12, 13].

The characteristic specific to the rate estimation is that a sequence of events (spikes) is described by a point process that is clearly a typical continuous time non-Gaussian process [14, 15, 16]. This means that the exact derivation of the marginal likelihood function is no longer possible.

In this manuscript, we show that in some limiting conditions the marginal likelihood function can be obtained analytically using the path integral method which is the standard technique in quantum physics [17, 18, 19, 20]. By utilizing the path integral method, we demonstrate an interesting phenomenon, i.e., the optimal time scale for the rate estimation exhibits the divergence continuously and discontinuously for two examples.

2 Methods

2.1 Time dependent Poisson process

We consider the time-dependent Poisson process in which point events (spikes) are derived independently from a given time-dependent rate $\lambda(t)$. In this process, the probability density for spikes to occur at $\{t_i\} \equiv \{t_1, t_2, \cdots, t_n\}$ within the interval $(0, T]$ for a given time-dependent rate $\lambda(t)$ is given by

$$p\left(\{t_i\} \mid \{\lambda(t)\}\right) = \left[\prod_{i=1}^{n} \lambda(t_i)\right] \exp\left(-\int_0^T \lambda(t)dt\right), \tag{1}$$

where the exponential term is the survivor function that represents the probability that no spikes occur in the inter-spike intervals [14, 15, 16]. Eq. (1) satisfies the normalization condition:

$$\begin{aligned}
&\sum_{n=0}^{\infty} \frac{1}{n!} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T t_n p\left(\{t_i\} \mid \{\lambda(t)\}\right) \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \left(\int_0^T \lambda(t)dt\right)^n \exp\left(-\int_0^T \lambda(t)dt\right) = 1.
\end{aligned} \tag{2}$$

2.2 The empirical Bayes method

For a sequence of point events derived from a time-dependent Poisson process, we apply the Bayes method to the inference of the time-dependent rate $\lambda(t)$. We introduce here a Gaussian process prior of $\lambda(t)$, which can be written in the general form:

$$p(\{\lambda(t)\}) = \frac{1}{Z} \exp\left[-\frac{1}{2}\lambda(t)^T A \lambda(t)\right], \tag{3}$$

where A is a linear operator, and Z denotes a normalization constant. Here the inner product of two functions is defined by $a(t)^T b(t) = \int dt a(t) b(t)$ [24]. In this paper, we consider $A \equiv [\Delta \frac{d}{dt}]^T [\Delta \frac{d}{dt}]$, where the hyperparameter Δ is proportional to the time scale of the estimation of the instantaneous rate. This represents the tendency of the of the estimated rate to be relatively flat. Then, the prior distribution is obtained as

$$p_\Delta(\{\lambda(t)\}) = \frac{1}{Z(\Delta)} \exp\left[-\frac{\Delta^2}{2} \int_0^T \left(\frac{d\lambda(t)}{dt}\right)^2 dt\right]. \tag{4}$$

The posterior distribution of $\lambda(t)$, given the data $\{t_i\}$, can be calculated with the Bayes formula:

$$p_\Delta(\{\lambda(t)\}|\{t_i\}) = \frac{p(\{t_i\}|\{\lambda(t)\}) p_\Delta(\{\lambda(t)\})}{p_\Delta(\{t_i\})}, \tag{5}$$

where $p_\Delta(\{t_i\})$ is the "marginal likelihood function" or the "evidence" for the hyperparameter Δ, with the given data $\{t_i\}$, defined as

$$p_\Delta(\{t_i\}) = \int p(\{t_i\}|\{\lambda(t)\}) p_\Delta(\{\lambda(t)\}) d\{\lambda(t)\}. \tag{6}$$

The integration above is functional integration over possible paths of $\lambda(t)$. According to the empirical Bayes theory, the hyperparameter can be adjusted so that the marginal likelihood function is maximized [4, 10, 11, 12, 13].

The maximum *a posteriori* (MAP) estimate $\hat{\lambda}(t)$ for a given hyperparameter Δ can be obtained by applying the variational method to the posterior distribution:

$$p_\Delta(\{\lambda(t)\}|\{t_i\}) \propto p(\{t_i\}|\{\lambda(t)\})p_\Delta(\{\lambda(t)\}). \tag{7}$$

2.3 The path integral method

The marginal likelihood function $p_\Delta(\{t_i\})$ defined by Eq.(6) can be represented in the form of the path integral

$$p_\Delta(\{t_i\}) = \frac{1}{Z(\Delta)} \int \exp\left[-\int_0^T L(\dot{\lambda}, \lambda, t)\, dt\right] d\{\lambda(t)\}, \tag{8}$$

where $L(\dot{\lambda}, \lambda, t)$ is the "Lagrangian" of the form:

$$L(\dot{\lambda}, \lambda, t) = \frac{\Delta^2}{2}\dot{\lambda}^2 + \lambda - \sum_{i=1}^{n} \delta(t - t_i) \log \lambda. \tag{9}$$

The MAP estimate $\hat{\lambda}(t)$ corresponds to the "classical path" obtained from the Euler-Lagrange equation,

$$\frac{d}{dt}(\frac{\partial L}{\partial \dot{\lambda}}) - \frac{\partial L}{\partial \lambda} = 0. \tag{10}$$

The marginal likelihood $p_\Delta(\{t_i\})$ for a given hyperparameter Δ can be computed by performing the integration, Eq.(7), over all paths $\lambda(t)$ around the above mentioned "classical path." The path integral can be obtained analytically for a Lagrangian in the quadratic form. By approximating the "action integral" to a range quadratic in the deviation from the classical path, the path integral is obtained as [18, 19, 20],

$$R \exp\left[-\int_0^T L(\dot{\hat{\lambda}}, \hat{\lambda}, t)\, dt\right], \tag{11}$$

where R represents the "quantum" contribution of the quadratic deviation to the path integral, obtained as

$$R = \sqrt{\frac{\frac{\partial^2 L}{\partial \dot{\lambda}^2}}{2\pi T}} \left[\frac{\det(\frac{\partial^2 L}{\partial \dot{\lambda}^2}\partial_t^2 + \frac{\partial^2 L}{\partial \lambda \partial \dot{\lambda}}\partial_t - \frac{\partial^2 L}{\partial \lambda^2})}{\det(\frac{\partial^2 L}{\partial \dot{\lambda}^2}\partial_t^2)}\right]^{-\frac{1}{2}} \tag{12}$$

$$= \sqrt{\frac{\frac{\partial^2 L}{\partial \dot{\lambda}^2}}{2\pi T}} \left[\frac{f_1(T,0)}{f_2(T,0)}\right]^{-\frac{1}{2}}. \tag{13}$$

Here, f_1 and f_2 respectively satisfy

$$(\frac{\partial^2 L}{\partial \dot{\lambda}^2}\partial_t^2 + \frac{\partial^2 L}{\partial \lambda \partial \dot{\lambda}}\partial_t - \frac{\partial^2 L}{\partial \lambda^2})f_1(t,0) = 0, f_1(0,0) = 0, \left.\frac{df_1(t,0)}{dt}\right|_{t=0} = 1, \tag{14}$$

$$\frac{\partial^2 L}{\partial \dot{\lambda}^2}\partial_t^2 f_2(t,0) = 0, f_2(0,0) = 0, \left.\frac{df_2}{dt}(t,0)\right|_{t=0} = 1. \tag{15}$$

The path integral method presented here can be regarded as a *functional* version of the Laplace approximation used in the field of machine learning [22, 23, 24]

2.4 Evaluation of the marginal likelihood

We formulate a method for computing the marginal likelihood function using the path integral method for a Poisson process with underlying rate

$$\lambda_0(t) = \mu + \sigma f(t), \tag{16}$$

where μ is the mean rate and $\sigma f(t)$ represents a fluctuation from the mean. The occurrence of spikes fluctuates around this underlying rate. Under the condition that the rate fluctuation is small compared to the mean, i.e., $\sigma/\mu \ll 1$, the fluctuating rate can be approximated as a stochastic process [25],

$$\sum_{i=1}^{n} \delta(t - t_i) = \lambda_0(t) + \sqrt{\lambda_0(t)}\,\xi(t), \tag{17}$$

where the fluctuation $\xi(t)$ is a Gaussian white noise characterized by the ensemble averages $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \delta(t - t')$.

By expressing the deviation of the estimated rate from the mean as $x(t) \equiv \lambda(t) - \mu$ and expanding the Lagrangian, Eq.(9), up to the quadratic order in x, we obtain

$$L(\dot{x}, x, t) \simeq \frac{\Delta^2 \dot{x}^2}{2} - \frac{\sigma f(t) + \sqrt{\mu}\,\xi(t)}{\mu} x + \frac{1}{2\mu} x^2 + \mu(1 - \log \mu). \tag{18}$$

We will hereafter ignore the constant $\mu(1-\log\mu)$, which is irrelevant to the dynamics. The "equation of motion" of the "classical path" is obtained with the Euler-Lagrange equation (10) as

$$\Delta^2 \frac{d^2 \hat{x}}{dt^2} + \frac{\sigma f(t) + \sqrt{\mu}\,\xi(t)}{\mu} - \frac{\hat{x}}{\mu} = 0. \tag{19}$$

The solution of this Euler-Lagrange equation is given as

$$\hat{x}(t) = \sigma \tilde{f}(t) + \sqrt{\mu}\,\tilde{\xi}(t), \tag{20}$$

where

$$\tilde{f}(t) = \int_0^T \frac{1}{2\Delta\sqrt{\mu}} \exp\left(-\frac{|t-s|}{\Delta\sqrt{\mu}}\right) f(s) ds, \tag{21}$$

$$\tilde{\xi}(t) = \int_0^T \frac{1}{2\Delta\sqrt{\mu}} \exp\left(-\frac{|t-s|}{\Delta\sqrt{\mu}}\right) \xi(s) ds. \tag{22}$$

Using the solution, the classical action is represented as

$$\frac{1}{T}\int_0^T L(\dot{\hat{x}}, \hat{x}, t) dt = \frac{\mu\Delta^2 C_{\dot{\tilde{\xi}}\dot{\tilde{\xi}}}}{2} - C_{\xi\tilde{\xi}} + \frac{C_{\tilde{\xi}\tilde{\xi}}}{2} + \frac{\sigma^2}{\mu}\left(\frac{\mu\Delta^2 C_{\dot{\tilde{f}}\dot{\tilde{f}}}}{2} - C_{f\tilde{f}} + \frac{C_{\tilde{f}\tilde{f}}}{2}\right), \tag{23}$$

where C_{AB} is defined as

$$C_{AB} \equiv \frac{1}{T}\int_0^T A(t)B(t) dt. \tag{24}$$

The contributions of the noises representing the sample fluctuation can be evaluated analytically in the limit $T \gg 1$. Applying the Wiener-Khinchine theorem to Eq. (24) and using the fact that the value of the power spectrum of the Gaussian white noise is constant, we obtain

$$\frac{1}{\mu}\left(\frac{\mu\Delta^2 C_{\dot{\tilde{\xi}}\dot{\tilde{\xi}}}}{2} - C_{\xi\tilde{\xi}} + \frac{C_{\tilde{\xi}\tilde{\xi}}}{2}\right) = -\frac{1}{4\Delta\sqrt{\mu}}. \tag{25}$$

The "free energy" that represents the negative marginal likelihood function is given by

$$F(\Delta) \equiv -\frac{1}{T}\log p_\Delta(\{t_i\}_{i=0}^n) = -\frac{1}{T}\left(\log R - \int_0^T L(\dot{\hat{x}}, \hat{x}, t) dt\right). \tag{26}$$

The contribution of the "quantum" part is computed as

$$R = \left(\frac{\Delta}{\pi\sqrt{\mu}} \right)^{\frac{1}{2}} \exp\left(-\frac{T}{2\Delta\sqrt{\mu}} \right) \tag{27}$$

As a result, the free energy is given as

$$F(\Delta) = \frac{1}{4\Delta\sqrt{\mu}} + \frac{\sigma^2}{\mu} \left(\frac{\mu\Delta^2 C_{\dot{\tilde{f}}\dot{\tilde{f}}}}{2} - C_{f\tilde{f}} + \frac{C_{\tilde{f}\tilde{f}}}{2} \right). \tag{28}$$

The hyperparameter is selected so that the free energy $F(\Delta)$ is minimized,

$$\hat{\Delta} = \arg\min_{\Delta} F(\Delta). \tag{29}$$

3 Results

3.1 Sinusoidally regulated Poisson process

First, we apply the formula (28) developed above to the case in which the rate is modulated sinusoidally in time, or

$$f(t) = \sin\frac{t}{\tau}. \tag{30}$$

In this case, the free energy is explicitly

$$F(\Delta) = \frac{1}{4\Delta\sqrt{\mu}} - \frac{\sigma^2}{4\mu} \frac{1}{1 + \frac{\Delta^2\mu}{\tau^2}}. \tag{31}$$

$\Delta = \infty$ implies that a sequence is interpreted as a constant rate. For a small amplitude of rate fluctuation, σ, this free energy has a local minimum at $\Delta = \infty$, or

$$\left. \frac{dF(\Delta)}{d(1/\Delta)} \right|_{1/\Delta=0} > 0. \tag{32}$$

By increasing the amplitude of rate fluctuation σ, there appears another local minimum at a finite Δ in the free energy. There is the local minimum at a finite Δ if

$$z \equiv \frac{\sigma^2\tau}{\mu} > \left(\frac{4}{3} \right)^{\frac{3}{2}}. \tag{33}$$

By increasing the rate fluctuation σ further, the local minimum of F at a finite Δ becomes smaller than zero, implying that the sequence should be interpreted as derived from a fluctuating rate, $\hat{\Delta} < \infty$, provided that

$$z > 2. \tag{34}$$

In the limit of a large fluctuation $z \gg 1$, the optimized time scale $\hat{\Delta}\sqrt{\mu}$ obeys the scaling relation,

$$\hat{\Delta}\sqrt{\mu} \sim \sigma^{-\frac{2}{3}} \tau^{\frac{2}{3}} \mu^{\frac{1}{3}}. \tag{35}$$

This is consistent with the result found by Bialek et al., [17].

3.2 Doubly stochastic Poisson process

Next, we consider the doubly stochastic Poisson process in which the rate fluctuation $f(t)$ obeys the Ornstein-Uhlenbeck process,

$$\frac{df}{dt} = -\frac{f}{\tau} + \sqrt{\frac{2}{\tau}}\xi(t). \tag{36}$$

In this case, the free energy is

$$F(\Delta) = \frac{1}{4\Delta\sqrt{\mu}} - \frac{\sigma^2}{2\mu}\frac{1}{1+\frac{\Delta\sqrt{\mu}}{\tau}} \tag{37}$$

In this case the free energy has a single minimum. The condition for this free energy function to have a minimum at finite Δ is

$$z \equiv \frac{\sigma^2\tau}{\mu} > \frac{1}{2}. \tag{38}$$

In the limit of a large fluctuation $z \gg 1$, the optimized time scale $\hat{\Delta}\sqrt{\mu}$ obeys the scaling relation

$$\hat{\Delta}\sqrt{\mu} \sim \sigma^{-1}\tau^{\frac{1}{2}}\mu^{\frac{1}{2}}. \tag{39}$$

4 Summary

We applied the empirical Bayes method based on the Gaussian process prior to the rate estimation of point events or spikes. A hyperparameter representing the time scale of the rate estimation kernel is selected so that the marginal likelihood function is maximized. We then obtained the marginal likelihood function analytically using the path integral method.

The time scale of the rate estimation kernel diverges in the case that the rate fluctuation of the underlying rate is small. This implies that under some conditions, it is likely that trying to say something about the rate fluctuation is misleading: *Speech is silver, silence is golden*. By means of the path integral method, we found two cases that exhibit continuous and discontinuous transitions.

Note that the incapability of rate-estimation is not necessarily due to the Bayesian method applied to the kernel estimation. In selecting a bin size for the time histogram method so that the mean integrated square error (MISE) from the underlying rate is minimized, the optimal bin size diverges when the spikes are derived from moderately fluctuating rates [26]. The bin size of the time histogram plays a similar role to the time scale of the rate-estimation kernel in determining the smoothness of the rate. The parametric condition for the divergence of the optimal bin size and that of the time scale of the rate-estimation kernel studied here are very similar. The asymptotic characteristics Eqs. (35) and (39) are respectively the same as those for the optimal bin size determined with the MISE criteria. It would be interesting to investigate the relationship of these apparently independent principles.

In the present study we consider the first order derivative of $\lambda(t)$ in the prior distribution (4). A higher order derivative of $\lambda(t)$ would be preferable in the case that a more smooth estimate of the rate is required. Further investigation is necessary for the case in which a higher order derivative, or another covariance function is taken into consideration.

The results here are obtained under the limiting conditions of $\sigma/\mu \ll 1$ and $T \gg 1$. We also applied the second order (Laplace) approximation to the evaluation of the marginal likelihood function. It remains for a future work to examine the extent to which these results are relevant for realistic parameter choices by simulation studies.

Acknowledgments

This study is supported in part by Grants-in-Aid for Scientific Research to SS from the Ministry of Education, Culture, Sports, Science and Technology of Japan (16300068, 18020015) and the 21st Century COE "Center for Diversity and Universality in Physics". SK is supported by the Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

References

[1] Williams, K.I. & Rasmussen, C.E. (1996) Gaussian processes for regression. *Advances in Neural Information Processing Systems*, **8**:598-604.

[2] Williams, K.I. & Barber, D. (1998) Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**:1342-1351.

[3] Rasmussen, C.E. & Williams, K.I. (2006) *Gaussian Processes for Machine Learning.* Cambridge: The MIT Press.

[4] MacKay, D.J.C. (1992) Bayesian interpolation. *Neural Computation* **4**:415-447.

[5] Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. (1996) *Spikes: Exploring the Neural Code.* Cambridge: The MIT Press.

[6] Dayan, P. & Abbott, L.F. (2001) *Theoretical Neuroscience.* Cambridge: The MIT Press.

[7] Ventura, V., Carta, R., Kass, R.E., Olson, C.R. & Gettner, S.N. (2002) Statistical analysis of temporal evolution in single-neuron firing rates. *Biostatistics* **1**:1-20.

[8] Kass, R.E., Ventura, V. & Cai, C. (2003) Statistical smoothing of neuronal data. *Network: Computation in Neural Systems* **14**:5-15.

[9] Smith, A.C. & Brown, E.N. (2003) Estimating a state-space model from point process observations. *Neural Computation* **15**:965-991.

[10] Good, I.J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge: MIT Press.

[11] Akaike, H. (1980) Likelihood and Bayes procedure. Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M. (eds.) *Bayesian Statistics*, pp. 143-166. Valencia: University Press.

[12] Bruce, A.B. & Saad, D. (1994) Statistical mechanics of hypothesis evaluation. *Journal of Physics A: Mathematical and General* **27**:3355-3364.

[13] Carlin, B.P. & Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. New York: Chapman and Hall.

[14] Cox, D.R. & Lewis, P.A.W. (1966) *The Statistical Analysis of Series of Events.* New York: John Wiley & Sons, Inc.

[15] Snyder, D.L. (1975) *Random Point Processes.* New York: John Wiley & Sons, Inc.

[16] Daley, D.J. & Ver-Jones, D. (2003) *An Introduction to the Theory of Point Processes* 2nd ed. New York: Springer-Verlag.

[17] Bialek, W, Callan, C.G. & Strong, S. P. (1996) Field theories for learning probability distributions. *Physical Review Letters* **77**:4693-4697.

[18] Feynman, R.P. & Hibbs, A.R. (1965) *Quantum Mechanics and Path Integrals.* New York: McGraw-Hill, Inc.

[19] Coleman, S. (1975) *Aspects of Symmetry.* Cambridge: Cambridge University Press.

[20] Schulman, L.S. (1981) *Techniques and Applications of Path Integration.* New York: John Wiley & Sons Inc.

[21] MacKay, D.J.C. (1998) Introduction to Gaussian processes. Bishop, C.M (eds.) *Neural Networks and Machine Learning*, Vol.168 of *NATO Series*, pp. 133-165. Berlin: Springer.

[22] Tierney, L. & Kadane, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**:82-86.

[23] Kass, R.E., Tierney, L. & Kadane, J.B. (1991) Laplace's method in Bayesian analysis. Flournoy, N. & Tsutakawa R.K. (eds.) *Statistical Multiple Integration*, pp. 89-99. *Contemporary Mathematics* Providence: American Mathematical Society.

[24] MacKay, D.J.C. (1998) Choice of basis for Laplace approximation. *Machine Learning* **33**:77-86.

[25] Tuckwell, H.C. (1988) *Introduction to Theoretical Neurobiology.* Cambridge: Cambridge University Press.

[26] Koyama, S. & Shinomoto, S. (2004) Histogram bin width selection for time-dependent Poisson processes. *Journal of Physics A: Mathematical and General* **37**:7255-7265.