Journal of Multivariate Analysis 141 (2015) 35–48

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Simultaneous prediction for independent Poisson processes with different durations

Fumiyasu Komaki *

Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

ARTICLE INFO

Article history:
Received 31 January 2014
Available online 23 June 2015

AMS subject classifications:
60G55
62F15
62M20

Keywords:
Harmonic time
Jeffreys prior
Kullback–Leibler divergence
Predictive density
Predictive metric
Shrinkage prior

ABSTRACT

Simultaneous predictive densities for independent Poisson observables are investigated. The observed data and the target variables to be predicted are independently distributed according to different Poisson distributions parametrized by the same parameter. The performance of predictive densities is evaluated by the Kullback–Leibler divergence. A class of prior distributions depending on the objective of prediction is introduced. A Bayesian predictive density based on a prior in this class dominates the Bayesian predictive density based on the Jeffreys prior.

1. Introduction

Suppose that x_i $(i = 1, \dots, d)$ are independently distributed according to the Poisson distribution with mean $r_i\lambda_i$ and that y_i $(i = 1, \dots, d)$ are independently distributed according to the Poisson distribution with mean $s_i\lambda_i$. Then,

$$p(x \mid \lambda) = \prod_{i=1}^{d} \frac{(r_i\lambda_i)^{x_i}}{x_i!} e^{-r_i\lambda_i}, \tag{1}$$

and

$$p(y \mid \lambda) = \prod_{i=1}^{d} \frac{(s_i\lambda_i)^{y_i}}{y_i!} e^{-s_i\lambda_i}, \tag{2}$$

where $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$. Here, $\lambda := (\lambda_1, \dots, \lambda_d)$ is the unknown parameter, and $r = (r_1, \dots, r_d)$ and $s = (s_1, \dots, s_d)$ are known positive constants. The objective is to construct a predictive density $\hat{p}(y; x)$ for y by using x.

For example, suppose that a shop has a purchase history data set for d customers. The time lengths r_i $(i = 1, \dots, d)$ of the purchase history data are different depending on the customers $i = 1, \dots, d$. The purchasing processes for the customers

* Correspondence to: Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

E-mail address: komaki@mist.i.u-tokyo.ac.jp.

http://dx.doi.org/10.1016/j.jmva.2015.06.008
0047-259X/

are d independent Poisson processes with intensities λ_i $(i = 1, \ldots, d)$. If the shop wants to predict the purchases of the customers in future time intervals with lengths s_i $(i = 1, \ldots, d)$, then the problem can be formulated as above.

The performance of $\hat{p}(y; x)$ is evaluated by the Kullback–Leibler divergence

$$D(p(y \mid \lambda), \hat{p}(y; x)) := \sum_y p(y \mid \lambda) \log \frac{p(y \mid \lambda)}{\hat{p}(y; x)}$$

from the true density $p(y \mid \lambda)$ to the predictive density $\hat{p}(y; x)$. The risk function is given by

$$\mathrm{E}\Big[D(p(y \mid \lambda), \hat{p}(y; x)) \,\Big|\, \lambda\Big] = \sum_x \sum_y p(x \mid \lambda) p(y \mid \lambda) \log \frac{p(y \mid \lambda)}{\hat{p}(y; x)}.$$

It is widely recognized that Bayesian predictive densities

$$p_\pi(y \mid x) := \frac{\int p(y \mid \lambda) p(x \mid \lambda) \pi(\lambda) \mathrm{d}\lambda}{\int p(x \mid \lambda) \pi(\lambda) \mathrm{d}\lambda},$$

where $\mathrm{d}\lambda = \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d$, constructed by using a prior π perform better than plug-in densities $p(y \mid \hat{\lambda})$ constructed by replacing the unknown parameter λ by an estimate $\hat{\lambda}(x)$. The choice of π becomes important to construct a Bayesian predictive density.

The Jeffreys prior

$$\pi_{\mathrm{J}}(\lambda) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d \propto \lambda_1^{-\frac{1}{2}} \cdots \lambda_d^{-\frac{1}{2}} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d \tag{3}$$

for $p(x \mid \lambda)$ coincides with the Jeffreys prior for $p(y \mid \lambda)$ and the volume element prior $\pi_{\mathrm{P}}(\lambda)$ with respect to the predictive metric discussed in Section 4. A natural class of priors including the Jeffreys prior is

$$\pi_\beta(\lambda) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d := \lambda_1^{\beta_1 - 1} \cdots \lambda_d^{\beta_d - 1} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d,$$

where $\beta_i > 0$ $(i = 1, \ldots, d)$.

We introduce a class of priors defined by

$$\pi_{\alpha,\beta,\gamma}(\lambda) \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d := \frac{\lambda_1^{\beta_1 - 1} \cdots \lambda_d^{\beta_d - 1}}{(\lambda_1/\gamma_1 + \cdots + \lambda_d/\gamma_d)^\alpha} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_d,$$

where $0 \le \alpha \le \beta_{\cdot} := \sum_i \beta_i$, $\beta_i > 0$, and $\gamma_i > 0$ $(i = 1, \ldots, d)$. In the following, a dot as a subscript indicates summation over the corresponding index. Note that $\pi_{\alpha,\beta,\gamma} \propto \pi_{\alpha,\beta,c\gamma}$, where $c > 0$ and $c\gamma = (c\gamma_1, \ldots, c\gamma_d)$. The prior $\pi_{\alpha,\beta,\gamma}$ does not depend on $\gamma := (\gamma_1, \ldots, \gamma_d)$ if $\alpha = 0$. If $\alpha > 0$, $\pi_{\alpha,\beta,\gamma}$ puts more weight on parameter values close to 0 than π_β does. In this sense, $\pi_{\alpha,\beta,\gamma}$ with $\alpha > 0$ is a shrinkage prior.

There have been several studies for the simple setting $r_1 = r_2 = \cdots = r_d$ and $s_1 = s_2 = \cdots = s_d$. Decision theoretic properties of linear estimators under the Kullback–Leibler loss is studied by Ghosh and Yang [5]. The theory for Bayesian predictive densities for the Poisson model under the Kullback–Leibler loss is a generalization of that for Bayesian estimators under the Kullback–Leibler loss. A class of priors $\pi_{\alpha,\beta} := \pi_{\alpha,\beta,\gamma}$ with $\gamma_1 = \cdots = \gamma_d = 1$ is introduced in [8]. It is shown that the risk of the Bayesian predictive density based on $\pi_{\tilde{\alpha},\beta}$ with $\tilde{\alpha} := \beta_{\cdot} - 1$ is smaller than the risk of that based on π_β if $\beta_{\cdot} > 1$. For example, if $d \ge 3$, there exists a Bayesian predictive density that dominates the Bayesian predictive density $p_{\mathrm{J}}(y \mid x)$ based on the Jeffreys prior because $\beta_{\cdot} = d/2 > 1$. Here, $p_\pi(y \mid x)$ is said to dominate $p_{\mathrm{J}}(y \mid x)$ if the risk of $p_\pi(y \mid x)$ is not greater than that of $p_{\mathrm{J}}(y \mid x)$ for all λ and the strict inequality holds for at least one point λ in the parameter space.

Bayesian predictive densities based on shrinkage priors are discussed by Komaki [7] and George et al. [2] for normal models. See also [3] for recent developments of the theory of predictive densities. In practical applications, it often occurs that observed data x and the target variable y to be predicted have different distributions parametrized by the same parameter. Regression models with the same parameter and different explanatory variable values are a typical example. Kobayashi and Komaki [6] and George and Xu [4] showed that shrinkage priors are useful for constructing Bayesian predictive densities for normal linear regression models. Komaki [10] has studied asymptotic theory for general models other than normal models when $x(i)$ $(i = 1, \ldots, N)$ and y have different distributions $p(x \mid \theta)$ and $p(y \mid \theta)$, respectively, with the same parameter θ. However, there has been few studies on nonasymptotic theories of Bayesian predictive densities for non-normal models when the distributions of x and y are different.

In the present paper, we develop finite sample theory for prediction when the data x and the target variable y have different Poisson distributions (1) and (2), respectively, with the same parameter λ. The proposed prior depends not only on r corresponding to the data distribution but also on s corresponding to the objective of prediction. Thus, we need to abandon the context invariance of the prior, see e.g. [1]. The Bayesian predictive densities studied in the present paper are not represented by using widely known functions such as gamma or beta functions, contrary to the simple setting $r_1 = \cdots = r_d$ and $s_1 = \cdots = s_d$ [8]. However, the predictive densities are represented by introducing a generalization of the Beta function, and the results are proved analytically.