Shrinkage priors for nonparametric Bayesian prediction of nonhomogeneous Poisson processes

Fumiyasu Komaki

The University of Tokyo, RIKEN CBS

Banff workshop
April 9, 2019

Outline

Statistical modeling and data analysis based on point process models using nonparametric Bayesian methods have various applications.

We consider nonparametric Bayesian inference of intensity functions of inhomogeneous Poisson processes, which are a basic point process model.

It is shown that Bayesian inference and prediction based on a class of improper priors is useful.

Several theorems corresponding to those for finite dimensional models also hold for the inhomogeneous Poisson process model.

1. Preliminary results: finite dimensional problems
2. Nonparametric Bayesian inference for nonhomogeneous Poisson processes

Prediction

- An observation x from a distribution with density $p(x \mid \theta)$ that belongs to a model $\{p(x \mid \theta) \mid \theta \in \Theta\}$.
- The objective is to predict an unobserved random variable y from the same (or closely related) distribution is predicted by using a predictive density $\hat{p}(y; x)$.

Loss function

The Kullback-Leibler divergence from the true density $p(y \mid \theta) = \prod_{i=1}^{m} p(x_{i+1} \mid \theta)$ to a predictive density $\hat{p}(y; x)$:

$$D\{p(y \mid \theta), \hat{p}(y; x)\} = \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y; x)} \mathrm{d}y.$$

- If we adopt a plug-in distribution $p(y \mid \hat{\theta}(x))$ as a predictive distribution, the loss for the plug-in distribution can be regarded as a loss for the estimator $\hat{\theta}$.
- Predictive distribution theory is a natural generalization of estimation theory under the Kullback-Leibler loss.

Risk function

$$\mathrm{E}[D(p,q) \mid \theta] := \int p(x \mid \theta) \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y;x)} \mathrm{d}y\mathrm{d}x.$$

Bayes risk

$$\mathrm{E}_\pi[D(p,q)] := \int \pi(\theta) \int p(x \mid \theta) \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y;x)} \mathrm{d}y\mathrm{d}x\mathrm{d}\theta.$$

$\pi(\theta)$ is a prior density.

Examples

- The multivariate Normal model with a known covariance matrix.
- Simultaneous predictive distributions for independent Poisson observables.

In these examples, Bayesian predictive distributions based on shrinkage priors dominate the Bayesian predictive distribution based on the Jeffreys prior when the dimension is not less than three.

The multivariate Normal model with a known covariance matrix

The d-dimensional Normal model $N(\mu, \Sigma)$.

μ: an unknown mean vector
Σ: a known variance-covariance matrix.

We assume that $\Sigma = I$ without loss of generality.

It suffices to consider the problem of predicting

$$y \sim N(\mu, \tau^2 I) \ \text{ using } \ x \sim N(\mu, \sigma^2 I).$$

The Lebesgue prior (the Jeffreys prior) $\pi_{\mathrm{I}}(\mu) = 1$ is commonly used as a non-informative prior for μ.

The Bayesian predictive density with $\pi_{\mathrm{I}}(\mu)$

$$p_{\mathrm{J}}(y \mid x) = \frac{1}{\{2\pi(\sigma^2 + \tau^2)\}^{d/2}} \exp\left\{ -\frac{1}{2(\sigma^2 + \tau^2)} \|y - x\|^2 \right\}.$$

The Bayesian predictive density $p_{\mathrm{J}}(y \mid x)$ dominates the plug-in density

$$p(y \mid \hat{\mu}) = \frac{1}{\{2\pi\sigma^2\}^{d/2}} \exp\left\{ -\frac{1}{2\sigma^2} \|y - x\|^2 \right\},$$

where $\hat{\mu} = x$.

Stein (1974) introduced the generalized Bayes estimator constructed using the prior

$$\pi_{\mathrm{S}}(\mu) = \|\mu\|^{-(d-2)}, \qquad d \geq 3.$$

The generalized Bayes estimator dominates the best invariant estimator $\hat{\mu} = x$.

Here, we consider the Bayesian predictive density $p_{\mathrm{S}}(y \mid x)$ based on Stein's prior $\pi_{\mathrm{S}}(\mu)$.

Theorem (K2001)

For all μ, the inequality

$$\mathrm{E}[D\{p(y \mid \mu), p_{\mathrm{J}}(y \mid x)\}|\mu] - \mathrm{E}[D\{p(y \mid \mu), p_{\mathrm{S}}(y \mid x)\}|\mu] > 0$$

holds. □

$p_\pi(y \mid x)$ 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.

George, E. I., Liang, F. and Xu, X. (2006):
sufficient conditions for general priors other than the Stein prior.

The multidimensional Poisson model

We assume that $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$ are distributed according to

$$\begin{aligned} p(x \mid \lambda) &= \prod_{i=1}^{d} p(x_i \mid \lambda) \\ &= \exp\{-(a\lambda_1 + a\lambda_2 + \cdots + a\lambda_d)\} \frac{(a\lambda_1)^{x_1}}{x_1!} \frac{(a\lambda_2)^{x_2}}{x_2!} \cdots \frac{(a\lambda_d)^{x_d}}{x_d!} \end{aligned}$$

and

$$\begin{aligned} p(y \mid \lambda) &= \prod_{i=1}^{d} p(y_i | \lambda) \\ &= \exp\{-(b\lambda_1 + b\lambda_2 + \cdots + b\lambda_d)\} \frac{(b\lambda_1)^{y_1}}{y_1!} \frac{(b\lambda_2)^{y_2}}{y_2!} \cdots \frac{(b\lambda_d)^{y_d}}{y_d!}, \end{aligned}$$

respectively.

We consider the problem of predicting

$$y =(y_1, y_2, \ldots, y_d)$$

using

$$x = (x_1, x_2, \ldots, x_d),$$

under the Kullback–Leibler loss

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

Here,

$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)$: unknown parameters,

a, b: known positive real numbers.

We consider a class of improper prior densities

$$\pi_{\alpha,\gamma}(\lambda)\mathrm{d}\lambda_1\mathrm{d}\lambda_2\cdots\mathrm{d}\lambda_d \propto \frac{\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\cdots\lambda_d^{\alpha_d-1}}{(\lambda_1+\lambda_2+\cdots+\lambda_d)^{\gamma}}\mathrm{d}\lambda_1\mathrm{d}\lambda_2\cdots\mathrm{d}\lambda_d$$

with $0 < -\gamma + \sum_i \alpha_i \leq 1$ and $\alpha_i > 0$ $(i = 1, 2, \ldots, d)$.

The Jeffreys prior

$$\pi_{\mathrm{J}}(\lambda)\mathrm{d}\lambda_1\mathrm{d}\lambda_2\cdots\mathrm{d}\lambda_d \propto \frac{1}{(\lambda_1\lambda_2\cdots\lambda_d)^{\frac{1}{2}}}\mathrm{d}\lambda_1\mathrm{d}\lambda_2\cdots\mathrm{d}\lambda_d$$

is equal to $\pi_{\alpha=(\frac{1}{2},\ldots,\frac{1}{2}),\gamma=0}(\lambda)\mathrm{d}\lambda_1\mathrm{d}\lambda_2\cdots\mathrm{d}\lambda_d$.

A limit of Gamma prior $\mathrm{Ga}(\alpha, \gamma)$

The Jeffreys prior is a limit of Gamma prior $\mathrm{Ga}(\alpha, \gamma)$ because

$$\lambda^{\alpha-1} \propto \lim_{\beta \to \infty} \beta^{\alpha} \frac{1}{\Gamma(\alpha)} \frac{\lambda^{\alpha-1}}{\beta^{\alpha}} \exp(-\lambda/\beta).$$

Intuitively speaking, the Jeffreys prior ${\lambda_i}^{1/2}$ is $\mathrm{Ga}(\alpha_i = 1/2, \beta = \infty)$.

Theorem (K2004)

The Bayesian predictive distribution based on the prior

$$\pi_{\alpha,\gamma}(\lambda)d\lambda_1 d\lambda_2 \cdots d\lambda_d \propto \frac{\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\cdots\lambda_d^{\alpha_d-1}}{(\lambda_1+\lambda_2+\cdots+\lambda_d)^{\gamma}} d\lambda_1 d\lambda_2 \cdots d\lambda_d$$

with $-\gamma + \sum_i \alpha_i > 0$ *and* $\alpha_i > 0$ $(i = 1, 2, \ldots, d)$ *is given by*

$$p_{\pi_{\alpha,\gamma}}(y \mid x) =$$

$$\left(\frac{a}{a+b}\right)^{\sum x_i - \gamma + \sum \alpha_i} \left(\frac{b}{a+b}\right)^{\sum y_i} \frac{\Gamma(\sum x_i + \sum y_i - \gamma + \sum \alpha_i)\Gamma(\sum x_i + \sum \alpha_i)}{\Gamma(\sum x_i - \gamma + \sum \alpha_i)\Gamma(\sum x_i + \sum y_i + \sum \alpha_i)}$$

$$\times \frac{\Gamma(x_1 + y_1 + \alpha_1)\Gamma(x_2 + y_2 + \alpha_2)\cdots\Gamma(x_d + y_d + \alpha_d)}{\Gamma(x_1 + \alpha_1)\Gamma(x_2 + \alpha_2)\cdots\Gamma(x_d + \alpha_d) y_1! y_2! \cdots y_d!}.$$

□

Theorem (K2004)

When

$$-\gamma + \sum_i \alpha_i > 1 \ \ \textit{and} \ \ \alpha_i > 0 \quad (i = 1, 2, \ldots, d),$$

the Bayesian predictive distribution

$$p_{\pi_{\alpha,\gamma}}(y \mid x) \ \ \textit{based on} \ \ \pi_{\alpha,\gamma}(\lambda)$$

is dominated by the Bayesian predictive distribution

$$p_{\pi_{\tilde{\alpha},\tilde{\gamma}}}(y \mid x) \ \ \textit{based on} \ \ \pi_{\tilde{\alpha},\tilde{\gamma}}(\lambda),$$

where $\tilde{\gamma} := \sum_i \alpha_i - 1$ *and* $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_d) := (\alpha_1, \alpha_2, \ldots, \alpha_d)$. □

We set

$$\pi_{\mathrm{S}}(\lambda) := \pi_{\alpha=(\frac{1}{2},\dots,\frac{1}{2}),\,\gamma=\frac{d}{2}-1}(\lambda) \quad \text{(a shrinkage prior).}$$

π_{S} gives more weight to parameter values close to the origin than the Jeffreys prior does.

Corollary (K2004)

When $d \geq 3$, the Bayesian predictive distribution $p_{\pi_{\mathrm{S}}}(y \mid x)$ based on the shrinkage prior $\pi_{\mathrm{S}}(\lambda)$ dominates the Bayesian predictive distribution $p_{\pi_{\mathrm{J}}}(y \mid x)$ based on the Jeffreys prior

$$\pi_{\mathrm{J}}(\lambda) d\lambda_1 d\lambda_2 \cdots d\lambda_d \propto \frac{1}{(\lambda_1 \lambda_2 \cdots \lambda_d)^{\frac{1}{2}}} d\lambda_1 d\lambda_2 \cdots d\lambda_d.$$

□

1. Preliminary results: finite dimensional problems
2. Nonparametric Bayesian inference for nonhomogeneous Poisson processes

Basic properties of nonparametric inference of nonhomogeneous Poisson processes using Gamma process priors are given by Lo (1982) and Lo and Weng (1989).

Corresponding results for probability density estimation is given by Lo (1984).

The results investigated in the following in this talk depends on that inference is for intensity functions not for probability densities.

The nonhomogeneous Poisson model

We conseder nonhomogeneous Poisson processes on $[0, 1]$ for simplicity of explanation.

Nonhomogeneous Poisson processes on more general spaces such as multidimensional Euclidean spaces can be treated exactly the same way.

$u \in [0, 1]$: time of the Poisson process.

$\mathscr{Po}(Y; t\lambda)$, $t > 0$: an nonhomogeneous Poisson process.
λ: a intensity measure
λ is also used for the intensity function $\lambda(u)$ by abuse of notation.

Y: a sample from the Poisson point process
with intensity measure $t\lambda$.

$\mathscr{Ga}(\lambda; \alpha, \beta)$: Gamma process prior.
$\alpha(u)$ is a function of u (density), β is a scalar not depending on u.

A Simple Setting: Gamma–Poisson Processes

The mixture of nonhomogeneous Poisson processes

$$Y \sim \mathscr{Po}(t\lambda)$$

with respect to the prior

$$\lambda \sim \mathscr{Ga}(\alpha, \beta)$$

is negative binomial process

$$Y \sim \mathscr{NeBi}(\alpha, t\beta/(1 + t\beta)).$$

We observe the nonhomogeneous Poisson process $X \sim \mathscr{Po}(s\lambda)$ $(s > 0)$.
We assume a Gamma prior $\lambda \sim \mathscr{Ga}(\alpha, \beta)$.

1. The posterior is
$\lambda \sim \mathscr{Ga}(\alpha + \sum_i \delta_{x_i}, 1/(s + 1/\beta)) = \mathscr{Ga}(\alpha + \sum_i \delta_{x_i}, \beta/(1 + s\beta))$.

2. The predictive process is
$Y \sim \mathscr{NeBi}(\alpha + \sum_i \delta_{x_i}, (s + 1/\beta)^{-1} t/\{1 + (s + 1/\beta)^{-1} t\})$
$= \mathscr{NeBi}(\alpha + \sum_i \delta_{x_i}, t\beta/\{1 + (s + t)\beta\})$.

□

Posterior

$\lambda \sim \mathscr{G}a(\alpha + \sum_i \delta_{x_i}, 1/(s + 1/\beta)) = \mathscr{G}a(\alpha + \sum_i \delta_{x_i}, \beta/(1 + s\beta))$.

The expectation of λ:

$$(\alpha + \sum_i \delta_{x_i}) \frac{\beta}{1 + s\beta}.$$

The variance of λ:

$$(\alpha + \sum_i \delta_{x_i}) \left(\frac{\beta}{1 + s\beta} \right)^2.$$

Bayesian predictive process:

$$\begin{aligned} Y \sim & \mathscr{NeBi}(\alpha + \sum_i \delta_{x_i}, (s+1/\beta)^{-1} t / \{1 + (s+1/\beta)^{-1} t\}) \\ = & \mathscr{NeBi}(\alpha + \sum_i \delta_{x_i}, t\beta/\{1+(s+t)\beta\}). \end{aligned}$$

The expectation of Y:

$$(\alpha + \sum_i \delta_{x_i}) \frac{t\beta}{1+(s+t)\beta} \left(1 - \frac{t\beta}{1+(s+t)\beta}\right)^{-1} = (\alpha + \sum_i \delta_{x_i}) \frac{t\beta}{1+s\beta}.$$

The variance of Y:

$$\begin{aligned} & (\alpha + \sum_i \delta_{x_i}) \frac{t\beta}{1+(s+t)\beta} \left(1 - \frac{t\beta}{1+(s+t)\beta}\right)^{-2} \\ & \qquad = (\alpha + \sum_i \delta_{x_i}) \frac{t\beta\{1+(s+t)\beta\}}{(1+s\beta)^2}. \end{aligned}$$

It is difficult to determine the scale parameter β in advance.

One method is to consider the limit $\beta \to \infty$. By applying the Bayes rule to the improper prior $\lambda \sim \mathscr{G}a(\alpha, \infty)$, we construct the posterior and the predictive processes.

Formerly, the density of the improper prior is $\lambda^{\alpha-1}$.

Theorem

If we observe $X \sim \mathscr{P}o(s\lambda)$ (nonhomogeneous Poisson process) and assume the improper prior $\lambda \sim \mathscr{G}a(\alpha, \infty)$, then

1. *the posterior is $\lambda \sim \mathscr{G}a(\alpha + \sum_i \delta_{x_i}, 1/s)$.*
2. *the predictive process is $Y \sim \mathscr{N}e\mathscr{B}i(\alpha + \sum_i \delta_{x_i}, t/(s+t))$.*

□

We show a theorem corresponds to theorems in K (2004) for finite dimensional models.

Let

$$|\alpha| := \int_U \mathsf{d}\alpha, \quad |\lambda| := \int_U \mathsf{d}\lambda.$$

Theorem

Assume that λ is absolutely continuous with respect to α. If $|\alpha| > 1$, the Bayesian predictive process based on the improper prior $\lambda \sim \mathscr{G}\mathscr{a}(\alpha, \infty)$ is dominated by the predictive process with improper prior

$$\frac{\lambda^{\alpha-1}}{|\lambda|^{|\alpha|-1}}.$$

□

Kernel mixture models

$$\mu \sim \mathcal{G}a(\alpha, \beta)$$

A (known) kernel function:

$$k(t, v) \ \text{ s.t. } \int k(t, v) \mathrm{d}t = 1$$

Example.

$$k(t, v) \propto \exp\left\{-\frac{1}{2\sigma^2}(t - v)^2\right\}$$

□

Kernel mixture

$$\lambda(x) = \int k(x, v)\mu(\mathrm{d}v)$$

Nonhomogeneous Poisson process:

$$\{x_1, x_2, \ldots, x_N\} \Big| \lambda \sim \mathcal{P}o(\lambda)$$

Likelihood for kernel mixture models

$$L(\lambda \mid \{x_1, x_2, \ldots, x_N\}) = \left\{\prod_{i=1}^{N} \lambda(x_i)\right\} \exp\left\{-\lambda(u)\mathrm{d}u\right\},$$

$$L(\mu \mid \{x_1, x_2, \ldots, x_N\}) = \left\{\prod_{i=1}^{N} \int k(x_i, v)\mu(\mathrm{d}v)\right\} \exp\left\{-\int\int k(u, v)\mu(\mathrm{d}v)\mathrm{d}u\right\}.$$

The posterior and Bayesian predictive processes for kernel mixture models have more complex forms than those for the simple Gamma–Poisson processes.

Theorem

Assume that λ is absolutely continuous with respect to the Lebesgue measure.

1. *The Bayesian predictive process based on the If $|\alpha| > 1$, improper prior $\lambda \sim \mathcal{G}a(\alpha, \infty)$ is dominated by the predictive process with improper prior*

$$\frac{\lambda^{\alpha-1}}{|\lambda|^{|\alpha|-1}}.$$

2. *The Bayesian predictive process based on the improper prior*

$$\frac{\lambda^{\alpha-1}}{|\lambda|^{\gamma}}$$

is admissible under the Kullback–Leibler loss if $|\alpha| - 1 \leq \gamma < |\alpha|$.

□

Conclusion

- A class of improper shrinkage priors for the nonhomogeneous Poisson processes is considered.
- A class of improper priors could be useful as objective priors for nonhomogeneous Poisson models.

Thank you for your attention!

References

George, E. I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Annals of Statistics*, vol. 34, 78–91.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate Normal observables. *Biometrika*, vol. 88, 859–864.

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables, *Annals of Statistics*, vol. 32, pp. 1744–1769.

Lo, A.Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 59, pp. 55–66.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density Estimates, *Annals of Statistics*, vol. 12, pp. 351–357.

Lo, A.Y. and Weng, C. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates, *Annals of the Institute of Statistical Mathematics*, vol. 41, pp. 227–245.

Stein, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, ed.) 345–381. Universita Karlova, Prague.