BIRS Workshop: Adaptive minimax predictive density for sparse Poisson models

Keisuke Yano

Joint work with Ryoya Kaneko, Fumiyasu Komaki

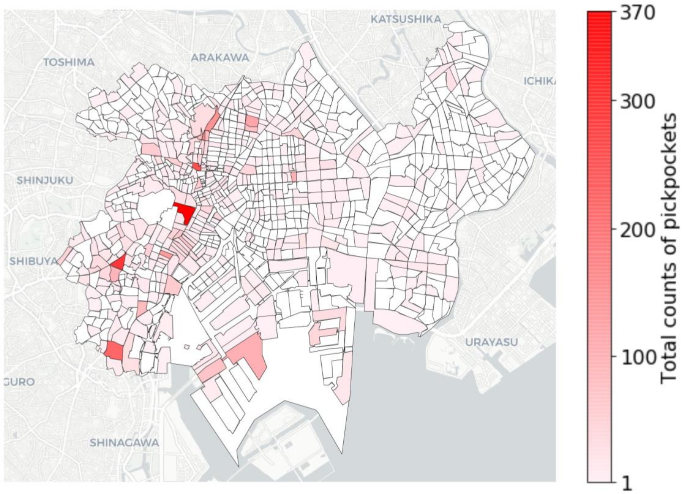
2019/04/11

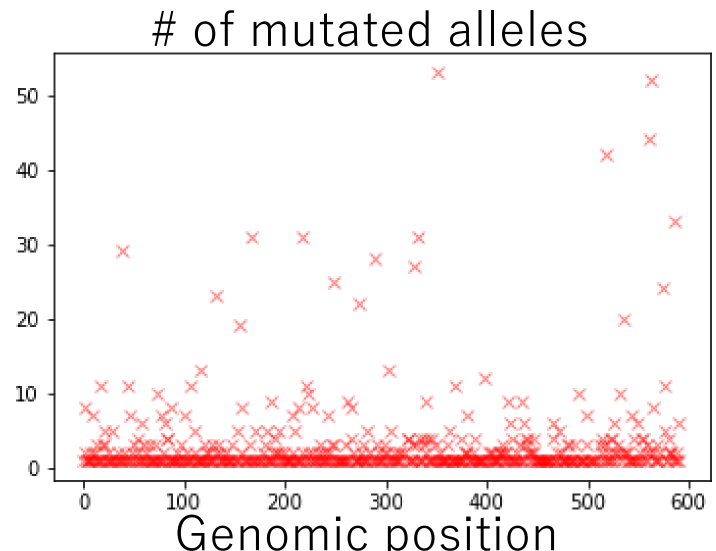
Table of contents

- Background
  - Motivative examples
  - Theoretical framework
- Main results
  - Exact asymptotically minimax risk
  - Exact asymptotically minimax predictive densities
  - Toward adaptation
- Simulation studies and applications to real data

Quick overview

- In count data, there exhibits an overabundance of zeros or near-zeros.



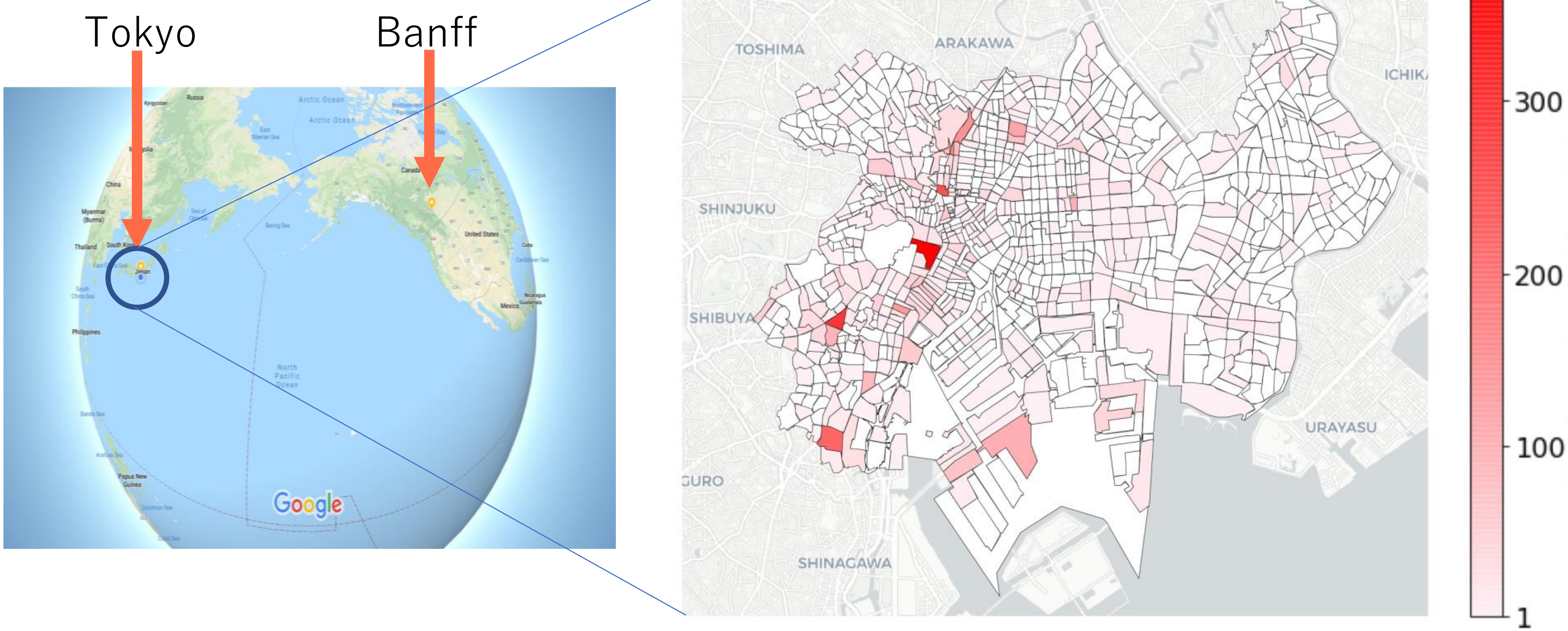


- Sample-size heterogeneity appears together with sparsity or quasi-sparsity.
- We discuss predictive densities for sparse count data with sample-size heterogeneity.

Table of contents

- Background
  - Motivative examples
  - Theoretical framework
- Main results
  - Exact asymptotically minimax risk
  - Exact asymptotically minimax predictive densities
  - Toward adaptation
- Simulation studies and applications to real data

Overabundance of zeros or near-zeros: Pickpocketing in Tokyo

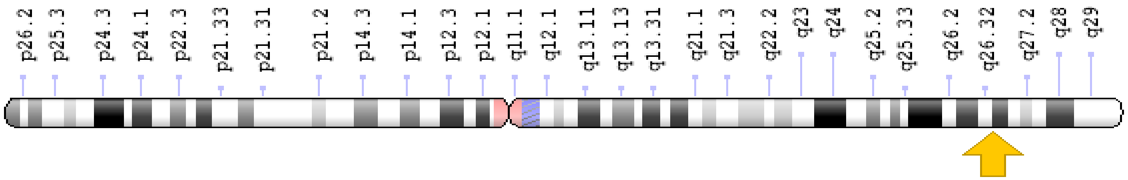


Pickpocketings at towns in 8 wards during 2012-2017 from http://www.keishicho.metro.tokyo.jp/

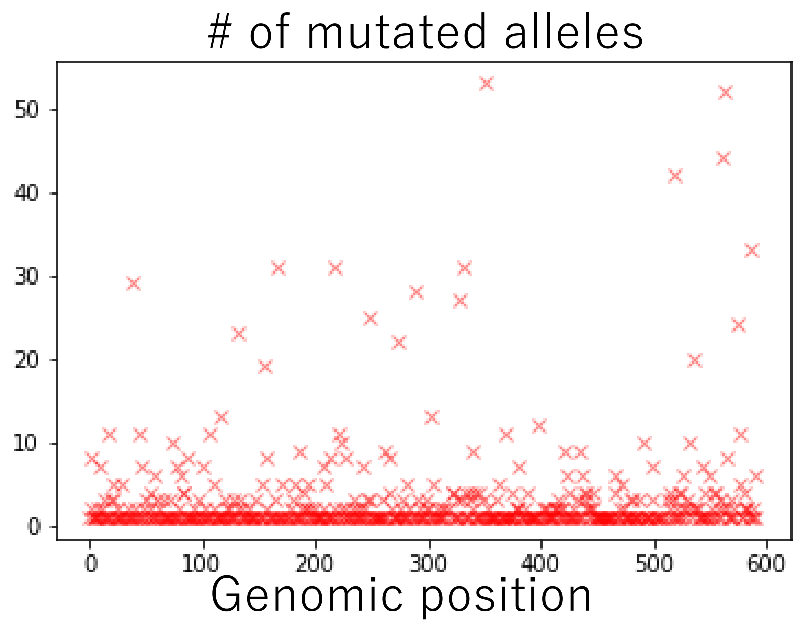
- More pickpocketings in towns with deeper red color
- Less pickpocketings in towns with lighter red color
- No pickpocketings in towns without color

Lots of zeros and near–zeros in the crime count data!!

Overabundance of zeros or near-zeros: Rare allele mutants in PIK3CA



- Genomic location of an oncogene PIK3CA from https://ghr.nlm.nih.gov/gene/PIK3CA
- Focus on rare allele mutants (allele frequencies <0.05)



- very low counts at a majority of genomic positions
- substantially higher counts at functionally relevant positions

Lots of near–zeros in the rare allele mutants!!

Heterogeneity arises in sparse count data

Most of sparse count data have sample-size heterogeneity

Ex.) Longitudinal data of pickpocketings in Tokyo

- $S_{t,i}$: 1 if town i reports crime counts 0 otherwise
- Expectation of crimes for T years at town i = $\sum_{t=1}^{T} S_{t,i}$ × crime rate θ_i

Town＼Year	2013	2014	2015	2016	2017	$\sum_{t=1}^{5} S_{t,i}$
Ginza 1	0	0	0	2	0	5
…						
Ginza 8	7	1	1	12	3	5
Irifune 1	0	0	0	0	0	5
Irifune 2	0	0	0	0	No report	4
…						
Hamarikyu	0	0	No report	0	No report	3

- Sample size $\sum_{t=1}^{T} S_{t,i}$ varies according to towns i

Heterogeneity arises in sparse count data

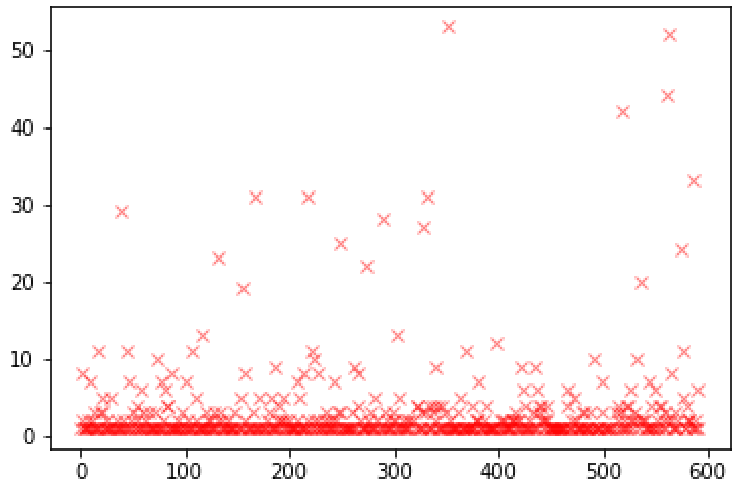
Most of sparse count data have sample-size heterogeneity

Ex.) rare allele mutants in PIK3CA

Mean of # of mutated alleles at each genomic position $\mathrm{E}[X_i]$

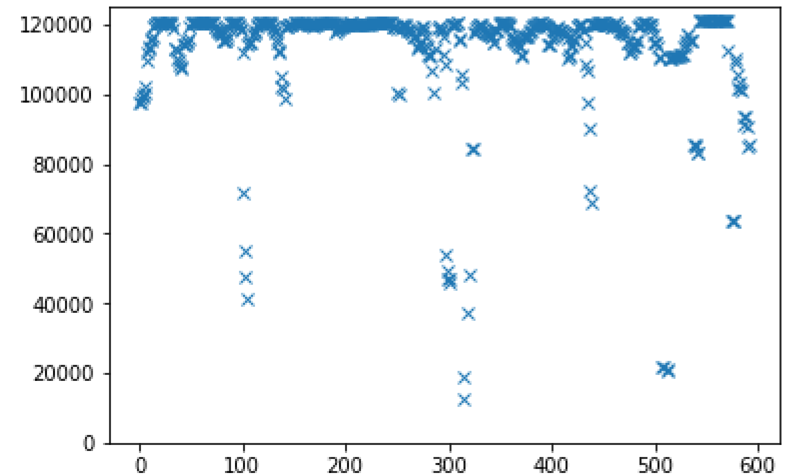
$=$ sequencing depth $r_i \times$ common mutation rate θ_i

# of mutated alleles: X_i



Genomic position i

Sequencing depth r_i



Genomic position i

- Sequencing depth r_i varies according to genomic positions i

Prediction for sparse count data with sample-size heterogeneity

In either example, prediction is of interest

- In the crime data, predicting the behavior of future crime counts based on past crime data is useful for preventing future crimes

- In the rare allele mutants data, predicting the rare allele mutations after normalization of sequencing depth removes the effect of heterogeneous sequencing depths
  - Observe $X_i \sim \mathrm{Po}(r_i\theta_i), indep.$ for $i = 1, \ldots, n$
  - Predict $Y_i \sim \mathrm{Po}(\theta_i), indep.$ for $i = 1, \ldots, n$

Our goal is to find a good predictive density for this set-up

Problem set-up

- Current observation: $X_i \sim \mathrm{Po}(r_i\theta_i), indep.$ for $i = 1, \dots, n$
- Future observation: $Y_i \sim \mathrm{Po}(\theta_i),\ indep.$ for $i = 1, \dots, n$
- Notation: $q(\, y \mid \theta \,) \coloneqq \prod_i (\theta_i^{y_i}/y_i!)\exp(-\theta_i)$
- Known parameter: sample size (ratio) $\{r_i : i = 1, \dots, n\}$
- Unknown parameter: $\theta = (\theta_1, \dots, \theta_n)$
    - θ is assumed to be sparse
      $$\theta = (\theta_1, \dots, \theta_n) \in \Theta[s_n] \coloneqq \{\theta : ||\theta||_0 \leq s_n\}$$
    - θ is assumed to be quasi-sparse
      $$\theta = (\theta_1, \dots, \theta_n) \in \Theta[s_n, \varepsilon_n] \coloneqq \{\theta : (\#i\ s.t.\ \theta_i > \varepsilon_n) \leq s_n\}$$

What is a good strategy for constructing a predictive density?

Decision-theoretic framework for prediction

- Predictive density: $\hat{q}(y; x)$
  - We predict future observations y using a predictive density $\hat{q}(y; x)$ based on current observations x.
  - Ex.) Bayesian predictive density based on a prior Π

$$q_{\Pi}(y \mid x) = \int q(y \mid \theta)\Pi(d\theta \mid x)$$

- Kullback-Leibler loss and risk: $L(x, \hat{q})$ and $R(\theta, \hat{q})$

$$L(x, \hat{q}) \coloneqq \sum_{y} q(y \mid \theta) \log \frac{q(y \mid \theta)}{\hat{q}(y; x)} \quad \& \quad R(\theta, \hat{q}) \coloneqq E_{X \mid \theta}[L(X, \hat{q})]$$

Our goal: find exact asymptotically minimax predictive densities

$$\sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q}) \sim \inf_{\hat{q}} \sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q}) \text{ as } n \to \infty \text{ and } \frac{s_n}{n} \to 0$$

Related literature on sparse count data analysis/ prediction for Poisson

- Sparse (or quasi-sparse) count data analysis
  - Manufacturing; c.f., Lambert (1992)
  - Micropropagation; c.f., Yang, Hardin, and Addy (2010)
  - Terrorist attacks; c.f., Datta and Dunson (2016)
- Estimation and Prediction using Poisson models under Kullback-Leibler loss
  - Simultaneous estimation; Ghosh and Yang (1988)
  - Shrinkage priors; Komaki (2004,2015)

This work discusses prediction (as well as estimation) using sparse Poisson models under Kullback-Leibler loss!

Related literature on exact asymptotically minimaxity

Table for Estimation	Gaussian	Poisson
Ellipsoidal constraint	Pinsker (1980)	Johnstone and MacGibbon (1992)
Sparsity constraint	Donoho, Johnstone, Hoch and Stern (1992)	This work

Table for Prediction	Gaussian	Poisson
Ellipsoidal constraint	Xu and Liang (2010)	*
Sparsity constraint	Mukherjee and Johnstone (2015,2017)	This work

Table of contents

- Background
  - Motivative examples
  - Theoretical framework
- Main results (we focus on sample-size homogeneous cases)
  - Exact asymptotically minimax risk
  - Exact asymptotically minimax predictive densities
  - Toward adaptation
- Simulation studies and applications to real data

Problem set-up (in this talk)

- Current observation: $X_i \sim \mathrm{Po}(r\theta_i), indep.$ for $i = 1, \dots, n$
- Future observation: $Y_i \sim \mathrm{Po}(\theta_i),\ indep.$ for $i = 1, \dots, n$
- Notation: $q(\, y \mid \theta \,) \coloneqq \prod_i (\theta_i^{y_i}/y_i!) \exp(-\theta_i)$
- Known parameter: sample size (ratio) $\{r_i : i = 1, \dots, n\}$
- Unknown parameter: $\theta = (\theta_1, \dots, \theta_n)$
  - θ is assumed to be sparse
    $$\theta = (\theta_1, \dots, \theta_n) \in \Theta[s_n] \coloneqq \{\theta : ||\theta||_0 \le s_n\}$$
  - θ is assumed to be quasi-sparse
    $$\theta = (\theta_1, \dots, \theta_n) \in \Theta[s_n, \varepsilon_n] \coloneqq \{\theta : (\#i\ s.t.\ \theta_i > \varepsilon_n) \le s_n\}$$

What is a good strategy for constructing a predictive density?

Exact asymptotically minimax risk

Let $\mathcal{C} := \left(\frac{r}{r+1}\right)^r \left(\frac{1}{r+1}\right)$

Theorem 2.1 of [Y., Kaneko, Komaki arXiv]

Fix $r \in (0, \infty)$ and fix a sequence $s_n \in (0, n)$ such that $\eta_n := s_n/n = o(1)$.

(a) For $\Theta[s_n]$

$$\inf_{\hat{q}} \sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q}) \sim \mathcal{C} s_n \log(\eta_n^{-1})$$

(b) For $\Theta[s_n, \varepsilon_n]$ with $\varepsilon_n = o(\eta_n)$

$$\inf_{\hat{q}} \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, \hat{q}) \sim \mathcal{C} s_n \log(\eta_n^{-1})$$

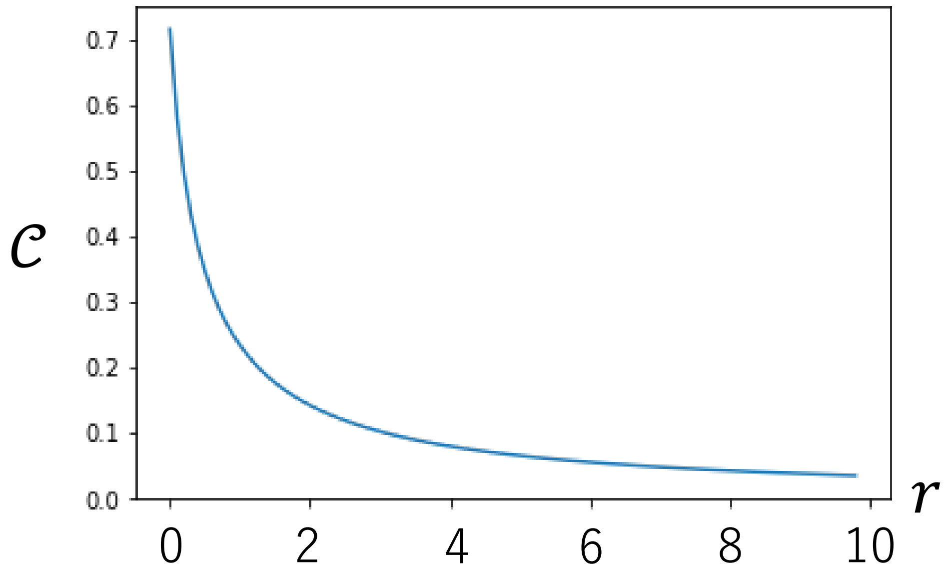
Implication of the theorem

- The rate is identical to that of sparse Gaussian models

$$\inf_{\hat{q}} \sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q}) \sim \mathcal{C} s_n \log(\eta_n^{-1})$$

- The exact constant depends on r

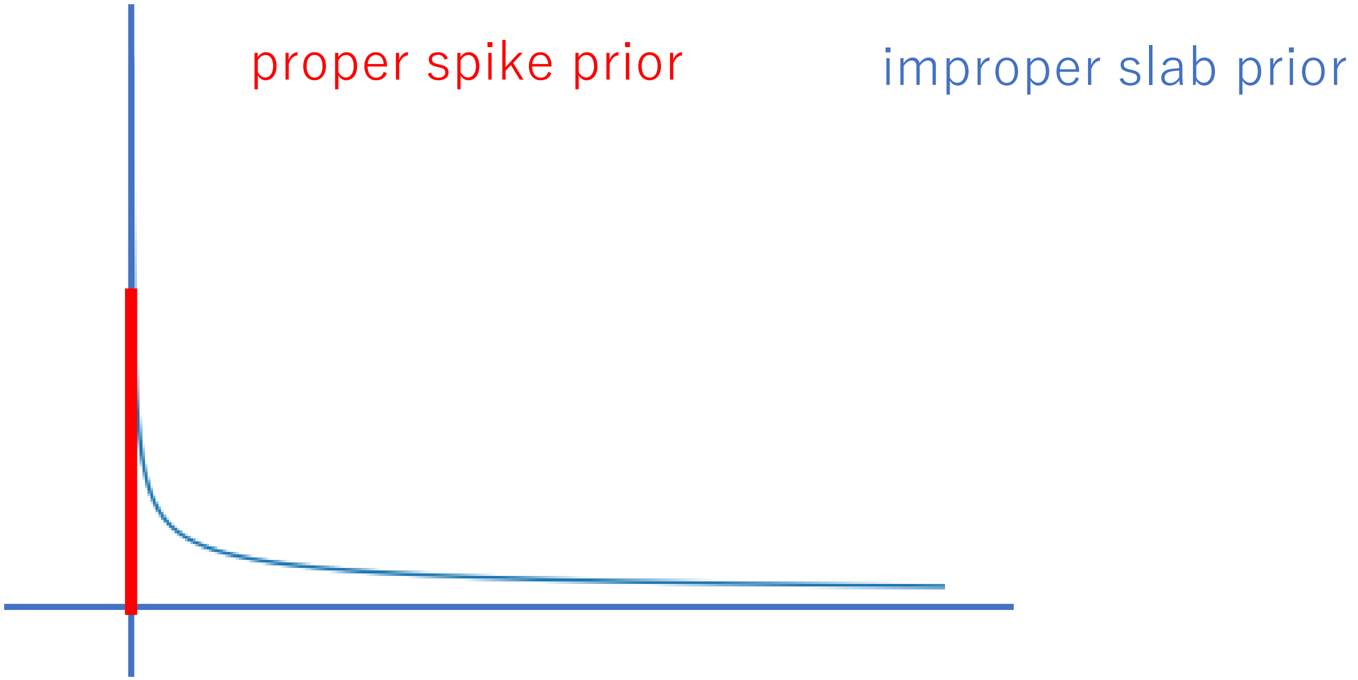
$$\inf_{\hat{q}} \sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q}) \sim \mathcal{C} s_n \log(\eta_n^{-1}) \text{ with } \mathcal{C} := \left(\frac{r}{r+1}\right)^r \left(\frac{1}{r+1}\right)$$



Spike-and-slab prior with improper slab

For $\kappa > 0$ and $h > 0$

$$\Pi[h,\kappa](d\theta_i) := \bigotimes_{i=1}^{n} \left[\delta_0(d\theta_i) + h\theta_i^{\kappa-1}1_{\theta_i \geq 0}\, d\theta_i\right]$$



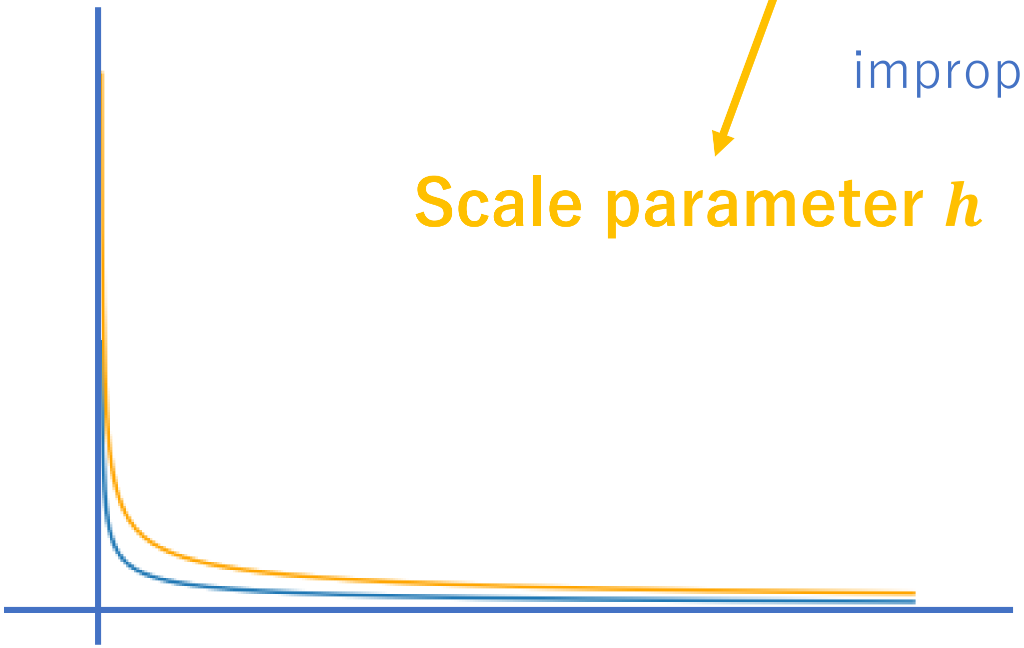
Spike-and-slab prior with improper slab

For $\kappa > 0$ and $h > 0$

$$\Pi[h,\kappa](d\theta_i) := \bigotimes_{i=1}^{n} \left[\delta_0(d\theta_i) + h\theta_i^{\kappa-1}1_{\theta_i \geq 0}\, d\theta_i\right]$$

improper slab prior

Scale parameter h



The scale of an improper prior within their mixture impacts on the posterior

Resulting Bayes predictive density

The resulting Bayes predictive density is controlled by h and κ of $\Pi[h,\kappa]$

$$q_{\Pi[h,\kappa]}(\, y \mid x \,)$$
$$= \prod_{1}^{n} \left\{ \omega_i \delta_0(y_i) + (1-\omega_i) \binom{x_i + y_i + \kappa - 1}{y_i} \left(\frac{r}{r+1} \right)^r \left(1 - \frac{r}{r+1} \right) \right\}$$

$$\text{where } \omega_i = \begin{cases} \frac{1}{1+h\Gamma(\kappa)/r^{\kappa}}, & x_i = 0 \\ 0, & x_i \geq 1 \end{cases}$$

- When $x_i \geq 1$, $q_{\Pi[h,\kappa]}(y_i \mid x_i)$ is just negative binomial
- When $x_i = 0$, $q_{\Pi[h,\kappa]}(y_i \mid x_i)$ is zero-inflated negative binomial

Our prior switches the predictive density according to the value of x !

Risk bounds for Bayes predictive densities based on $\Pi[h,\kappa]$

Let $\mathcal{C} := \left(\frac{r}{r+1}\right)^r \left(\frac{1}{r+1}\right)$ and $\mathcal{K} := \frac{r^{-\kappa} - (r+1)^{-\kappa}}{\kappa}$

Theorem 2.2 of [Y., Kaneko, Komaki arXiv]

Fix $r \in (0,\infty)$ and $\kappa > 0$.

Fix also $s_n \in (0,n)$ s.t. $\eta_n := s_n/n = o(1)$.

The predictive density $q_{\Pi[L\eta_n,\kappa]}$ with $L > 0$ and $\kappa > 0$satisfies

$$\sup_{\theta\in\Theta[s_n]} R\left(\theta, q_{\Pi[L\eta_n,\kappa]}\right) \leq \mathcal{C}s_n \log(\eta_n^{-1}) - \mathcal{C}s_n \log L + \mathcal{K}s_n L + \Upsilon_1$$

$$\sup_{\theta\in\Theta[s_n,\varepsilon_n]} R\left(\theta, q_{\Pi[L\eta_n,\kappa]}\right) \leq \mathcal{C}s_n \log(\eta_n^{-1}) - \mathcal{C}s_n \log L + \mathcal{K}s_n L + \Upsilon_2$$

where Υ_1, Υ_2 represent terms independent of L or $O(s_n\eta_n)$.

Implication of the theorem

- Theoretical consideration:

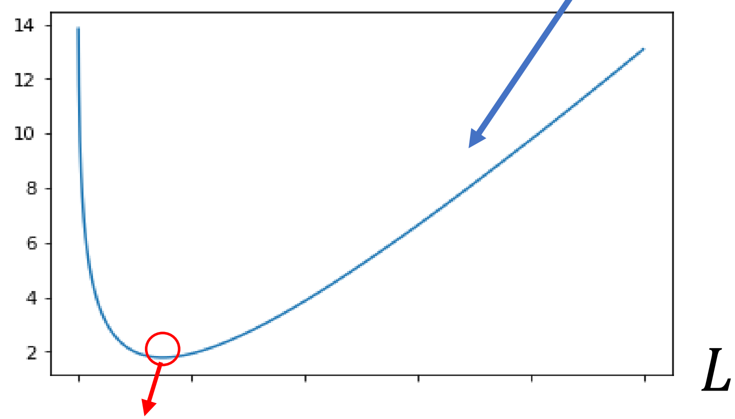
  $q_{\Pi[L\eta_n,\kappa]}$ is exact asymptotically minimax for any $L, \kappa > 0$.

- Practical consideration:

Tuning parameters L and κ even when $\eta_n \coloneqq s_n/n$ is known.

Our theorem also provides a theoretical guidepost for L.

$$\sup_{\theta\in\Theta[s_n]} R\left(\theta, q_{\Pi[L\eta_n,\kappa]}\right) \le \mathcal{C} s_n \log(\eta_n^{-1}) \underline{- \mathcal{C} s_n \log L + \mathcal{K} s_n L} + \Upsilon_1$$



- $L^* \coloneqq \mathcal{C}/\mathcal{K}$ minimizes the upper bound w.r.t . L !
- Prediction gives indication of how to select tuning parameter!

Implication of the theorem

- Theoretical consideration:

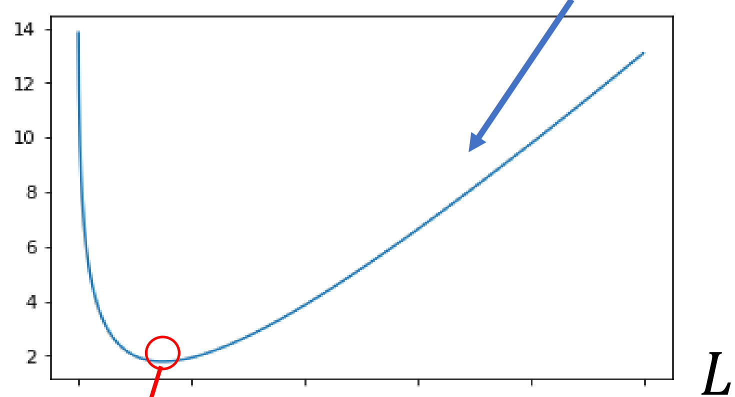
  $q_{\Pi[L\eta_n,\kappa]}$ is exact asymptotically minimax for any $L, \kappa > 0$.

- Practical consideration:

Tuning parameters L and κ even when $\eta_n := s_n / n$ is known.

Our theorem also provides a theoretical guidepost for L.

$$\sup_{\theta\in\Theta[s_n]} R\left(\theta, q_{\Pi[L\eta_n,\kappa]}\right) \le \mathcal{C} s_n \log(\eta_n^{-1}) \underline{- \mathcal{C} s_n \log L + \mathcal{K} s_n L} + \Upsilon_1$$



- $L^* := \mathcal{C}/\mathcal{K}$ minimizes the upper bound w.r.t . L !
- Prediction gives indication of how to select tuning parameter!

Adaptation to s_n

Plugging-in an estimator for s_n works to some extent.

Let $\hat{s}_n := \max\{1, \#\{i: X_i \geq 1\}\}$ and $\hat{\eta}_n := \hat{s}_n/n$.

Theorem 2.3 of [Y., Kaneko, Komaki arXiv]

Fix $r \in (0, \infty)$ and $\kappa > 0$.

For any $s_n \in (0, n)$ s.t. $s_n = o(n^{1/2})$,

the predictive density $q_{\Pi[L^* \hat{\eta}_n, \kappa]}$ with $\kappa > 0$ satisfies

$$\sup_{\theta \in \Theta[s_n]} R\left(\theta, q_{\Pi[L^* \hat{\eta}_n, \kappa]}\right) \sim \inf_{\hat{q}} \sup_{\theta \in \Theta[s_n]} R(\theta, \hat{q})$$

$$\sup_{\theta \in \Theta[s_n, \varepsilon_n]} R\left(\theta, q_{\Pi[L^* \hat{\eta}_n, \kappa]}\right) \sim \inf_{\hat{q}} \sup_{\theta \in \Theta[s_n, \varepsilon_n]} R(\theta, \hat{q}).$$

Table of contents

- Background
  - Motivative examples
  - Theoretical framework
- Main results
  - Exact asymptotically minimax risk
  - Exact asymptotically minimax predictive densities
  - Toward adaptation
- Simulation studies and applications to real data

Simulation studies

Comparisons using ℓ_1 point prediction; $\mathrm{E}[\log q(\mathrm{Y};\mathrm{X})]$; predictive coverage.

$$\theta_i \sim v_i e_{S,i} \mid \quad v_i \sim Gamma(10,1), \qquad S \sim \text{Unif on all } s\text{−sparse subsets}$$

1. Set-up: $(n, s, r) = (200, 5, 1)$

	$\boldsymbol{\Pi[L^* \, \hat{\eta}_n, 0.5]}$	$\boldsymbol{\Pi[L^* \, \hat{\eta}_n, 1]}$	**Gauss hypergeometric in Datta and Dunson(2016)**	**Shrinkage in Komaki (2004)**
Point prediction	18.8	21.9	104	96.5
$\mathrm{E}[\log q(\mathrm{Y};\mathrm{X})]$	-15.4	-16.1	-66.3	-86.2
90% Prediction Coverage	92.6	95.8	92.0	40.5

2. Set-up: $(n, s, r) = (200, 5, 20)$

	$\boldsymbol{\Pi[L^* \, \hat{\eta}_n, 0.5]}$	$\boldsymbol{\Pi[L^* \, \hat{\eta}_n, 1]}$	**Gauss hypergeometric in Datta and Dunson(2016)**	**Shrinkage in Komaki (2004)**
Point prediction	14.0	14.5	15.7	22.5
$\mathrm{E}[\log q(\mathrm{Y};\mathrm{X})]$	-13.3	-13.5	-15.6	-21.6
90% Prediction Coverage	90.0	89.4	97.6	97.5

Application to pickpocketing in Tokyo

- Pickpocketings at all towns of 8 wards in Tokyo
- Current observations X: data from 2012 to 2017
- Future observation Y: data from 2018/1 to 2018/6

	$\boldsymbol{\Pi[L^{*}\ \hat{\eta}_n, 0.5]}$	**Gauss hypergeometric in Datta and Dunson(2016)**	**Shrinkage in Komaki (2004)**
Point prediction	273	293	273
$[\log q(\mathrm{Y};\mathrm{X})]$	-399	-399	-429
90% Prediction marginal Coverage	93.0	27.0	84.2

Conclusion

- Prediction for Poisson models under sparsity (and quasi-sparsity) constraints
    - Many motivative examples
- Main results
    - Exact asymptotically minimax risks are identified
    - Exact asymptotically minimax predictive densities are constructed using spike-and-slab priors with improper slab priors.
    - Optimal scale of improper slab priors is specified by the predictive risk bound.
    - Plugging-in strategy works for adaptation
    - Sample-size heterogeneous versions are also obtained.
- This talk is based on our arXiv manuscript
    - K. Yano, R. Kaneko and F. Komaki: Exact Minimax Predictive Density for Sparse Count Data
    - arXiv:1812.06037v2

References 1

- J. Datta and D. Dunson (2016) Bayesian inference on quasi-sparse count data, *Biometrika*, **103**, 971-983.
- D. Donoho, I. Johnstone, J. Hoch and A. Stern (1992) Maximum entropy and the nearly black object. *Journal of Royal Statistical Society Series B*, **54**,41-81.
- M. Ghosh and M.-C. Yang (1988) Simultaneous Estimation of Poisson Means Under Entropy Loss, *The Annals of Statistics*, **16**, 278-291.
- I. Johnstone and B. MacGibbon (1992) MINIMAX ESTIMATION OF A CONSTRAINED POISSON VECTOR, *The Annals of Statistics*, **20**, 807-831.
- F. Komaki (2004) SIMULTANEOUS PREDICTION OF INDEPENDENT POISSON OBSERVABLES, *The Annals of Statistics*, **32**, 1744-1769.
- F. Komaki (2015) Simultaneous prediction for independent Poisson processes with different durations, *Journal of Multivariate Analysis*, **141**, 35-48.
- Lambert (1992) Zero-inflated Poisson regression, with an application to random defects in manufacturing, *Technometrics*, **34**, 1-14.

References 2

- G. Mukherjee and I. Johnstone (2015) EXACT MINIMAX ESTIMATION OF THE PREDICTIVE DENSITY IN SPARSE GAUSSIAN MODELS, *The Annals of Statistics,* **43**, 937-961.
- G. Mukherjee and I. Johnstone (2017) ON MINIMAX OPTIMALITY OF SPARSE BAYES PREDICTIVE DENSITY ESTIMATES, arXiv:1707.04380
- M. Pinsker (1980) Optimal filtration of square-integrable signals in Gaussian Noise, *Problems in Information Transmission*.
- X. Xu and F. Liang (2010) Asymptotic minimax risk of predictive density estimation for non-parametric regression, *Bernoulli*, **16**, 543-560.
- Z. Yang, J. Hardin and C. Addy (2009) Testing overdispersion in the zero-inflated Poisson model, *Journal of Statistical Planning and Inference*, **139**, 3340-3353.
- K. Yano, R. Kaneko and F. Komaki (2018) Exact Minimax Predictive Density for Sparse Count Data, arXiv:1812.06037v2