Model based clustering in group life insurance via Bayesian nonparametric mixtures

Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico

Laura D'Angelo

Abstract Experience rating allows insurance companies to adjust the premium for a certain policyholder on the basis of the loss experienced by similar insured parties. In group life insurance, because of the lack of complete information, clustering policyholders is particularly challenging. To address this issue we adopt a model-based clustering approach using flexible Bayesian nonparametric mixtures, which take into account the discrete nature of data, consisting of claim counts. We consider a Pitman-Yor mixture of Rounded Gaussian kernels, which provides more flexible and robust results than standard Poisson mixtures. We show how this approach leads to more accurate inference compared to standard Dirichlet process mixtures of Poisson, through an application to data arising from a portfolio of groups of workers.

Abstract *Nell'ambito delle assicurazioni sulla vita collettive è di particolare importanza definire il rischio associato ad un determinato gruppo: ciò viene svolto sulla base del numero di richieste di risarcimento osservate per gruppi con caratteristiche simili. Tuttavia, in questo contesto, individuare i gruppi simili risulta particolarmente complesso a causa della presenza di una grande eterogeneità non osservata. Per rispondere a questo problema, di seguito è proposto un approccio Bayesiano nonparametrico basato su un modello mistura. In particolare, per tenere conto della natura discreta dei dati, si propone una mistura di kernel Gaussiane arrotondate basata sul processo di Pitman-Yor. Attraverso un'applicazione a un portafoglio assicurativo, si mostra come questo modello fornisca risultati più accurati rispetto a una mistura basata sul kernel Poisson.*

Key words: Pitman-Yor Process, Rounded Gaussian, Dirichlet Process, Poisson mixtures

---

Laura D'Angelo
Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy, e-mail: laura.dangelo.1@phd.unipd.it

1 Introduction

An accurate assessment of the risk associated to policyholders is a fundamental task in group life insurance. This is usually achieved through experience rating, which allows to adjust the premium for a certain policyholder on the basis of the loss experienced by similar insured parties. While in individual life insurance one can record many relevant factors affecting risk, in group life insurance clustering observations becomes a tough challenge because of the presence of many unrecorded risk factors, which lead to a great unobserved heterogeneity.

Our study is motivated by the analysis of data of a real group life insurance portfolio, which consists of the claim counts and total risk exposure of 72 groups of workers insured in the period 1982-1985. The data are also analyzed in [7], [6] and [1]. In their work, [1] assert how standard parametric models do not allow to adequately describe the heterogeneity between groups, and adopt a more flexible approach using Bayesian nonparametric mixtures: specifically, they propose to use a Dirichlet Process mixture of Poisson kernels. Although apparently flexible, a deeper analysis reveals how this model has some drawbacks, arising from both choices of the kernel and of the mixing measure. In fact, even if the Poisson kernel is a natural choice in the context of count data, it lacks flexibility, having a single parameter controlling both the mean and the variance, and forcing them to be equal. This lack of flexibility is extended also to nonparametric mixtures: for example, all distributions which are under-dispersed can not be consistently estimated [2, 3]. We propose to use a Rounded Gaussian kernel, a more flexible kernel for count data introduced in [2], based on rounding of continuous kernels.

Regarding the mixing measure, even if the Dirichlet Process [4, 5] is a popular nonparametric prior, it does not allow to have full control of the clustering structure and the results are indeed heavily affected by the prior choice of the concentration parameter. Using a more general prior can relieve this problem: specifically, we adopt a Pitman-Yor process [8], a tractable generalization of the Dirichlet Process, which introduces one further parameter. In their work, [3] show how also in the case of count data, using this process as a prior in nonparametric mixtures can lead to more robust inference than the Dirichlet process, as for varying prior centering one obtains more stable results.

2 Model

Let $Y_i \in \mathscr{Y}$ be the discrete random variable representing the claim count for group i, $i = 1, \ldots, n$. We propose to model the distribution $p(y)$ as a nonparametric mixture

$$p(y) = \int K(y|\psi) dP(\psi),$$

where $K(y|\psi)$ is a kernel on $(\mathscr{Y},\Psi)$ and Ψ is a parametric space. As a prior for the unknown mixing measure P, we assume a Pitman-Yor process, $P \sim PY(\sigma,\theta,P_0)$, which includes the Dirichlet Process for $\sigma = 0$.

Concerning the choice of the kernel $K(y|\psi)$, we adopt a flexible Rounded Gaussian (RG) kernel. Briefly, the idea is to consider an underlying continuous variable $Y^* \in \mathscr{Y}^*$ with density f: the probability mass function of a discrete variable Y can be obtained through a set of thresholds on the support of Y^* as

$$p(j) = \int_{a_j}^{a_{j+1}} f(y^*)\,dy^*$$

where $\{a_j\}_{j=0,1,\ldots,+\infty}$ is a predefined sequence of thresholds such that $a_0 = min\{y^* : y^* \in \mathscr{Y}^*\}$ and $a_{+\infty} = max\{y^* : y^* \in \mathscr{Y}^*\}$. For the underlying continuous kernel, we assumed a Gaussian distribution, with thresholds $\{a_j\}_{j=0}^{+\infty} = \{-\infty, 0, 1, 2, \ldots, +\infty\}$. Indicating with μ and τ respectively the location and precision parameter of the rounded Gaussian kernel, $RG(\mu,\tau^{-1})$, the resulting model can be expressed through its hierarchical representation as:

$$\begin{aligned} Y_i|\mu_i,\tau_i &\sim RG(\mu_i,\tau_i^{-1}) \\ (\mu_i,\tau_i)|P &\sim P \\ P &\sim PY(\sigma,\theta,P_0) \end{aligned} \tag{1}$$

For simplicity of computation we chose a conjugate base measure: $P_0(\mu,\tau) = N(\mu;\mu_0,\kappa\tau^{-1})Gamma(\tau;\alpha,\beta)$, where the parameters μ_0 and κ were fixed equal to the sample mean and variance, respectively.

3 Application to insurance data

The dataset consists of the number of claims and risk exposure for 72 groups of workers, where the risk exposure is computed as the total exposure for all individuals in the group. Figure 1 shows the histogram of the number of claims normalized by the corresponding exposure: for many groups the number of deaths results to be zero, regardless of the amount of exposure; most of the remaining groups concentrate around small values, but there are few isolated groups for which the number of claims is definitely high: it seems indeed reasonable to assume a non-homogeneous portfolio.

On these data we computed the posterior under model (1), where the exposure was introduced as a multiplicative factor on the mean of the latent variable. In this way, the expected number of deaths is proportional to the duration of the exposure. On the contrary, we assumed no restrictions for the variance and it is independent of the exposure: the possibility to distinguish its effect on the mean and on the variance is another advantage of the Rounded Gaussian kernel, as for the Poisson kernel the only way is to constrain both the mean and the variance to be equally affected.

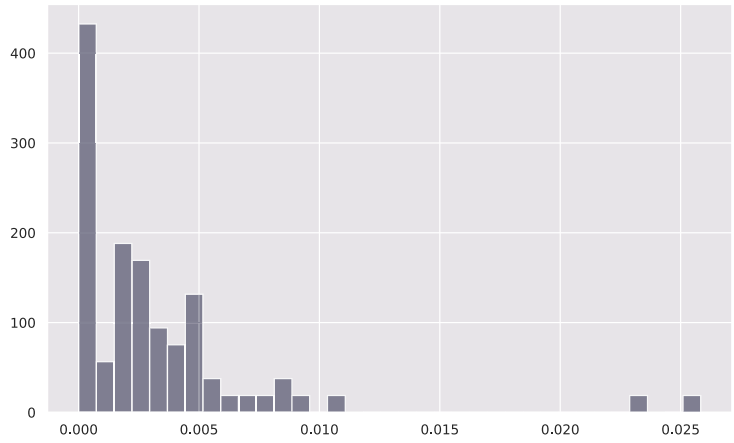

Fig. 1 Histogram of the number of deaths per unit of exposure.

Concerning the prior specification, the parameter σ of the Pitman-Yor process was fixed equal to 0.6, as positive values of this parameter lead to more robust results in the clustering structure of the mixture. This behavior is well displayed in [3], where through a simulation study the authors show how adopting a Dirichlet process (which corresponds to $\sigma = 0$) causes the posterior estimates of the number of clusters to be heavily affected by the prior expectation, while, for increasing σ, the posterior mean stabilizes even for different prior specifications. Concerning the parameter θ, we fixed it conditionally to the sample size and σ to have a prior expected number of clusters equal to 15. This prior specification also allows us to obtain a better comparison between the results of our model and of a Dirichlet Process mixture of Poisson kernels, as we are able to center both processes on the same prior expectation.

The posterior mean probability mass function (pmf) obtained with both mixtures, with the exposure fixed equal to three different levels, are displayed in Figures 2 and 3. At a first glance, the pmfs from the two models look very different, however, analyzing the resulting distributions for the same level of exposure, we can observe some similarities. In fact, the location of the mixture components is similar for both models: for an exposure level $E = 1000$, both pmf have maximum in 2 and, in general, most probability mass is assigned to values between 0 and 10. When the exposure is set to 5000 and 10000, the differences increase, as only for the components closest to zero we can identify clearly the same location. In fact, for the Poisson mixture, the components away from zero are less clear, as they are spread because of the forced increased variance, which is itself multiplied by a factor equal the exposure. This behavior is clearly enhanced for an exposure equal to 10000, as the components further from zero are completely flattened. This drawback does

not exist for the RG mixture and, indeed, all mixture components are well defined for every level of exposure. Moreover, as it is reasonable to guess, higher levels of exposure shift the location of the distribution to higher values of claim counts.

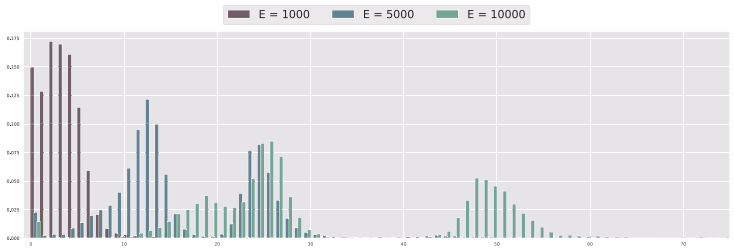

Fig. 2 Estimated pmf using a PYP mixture of RG kernels. Colors correspond to levels of exposure.

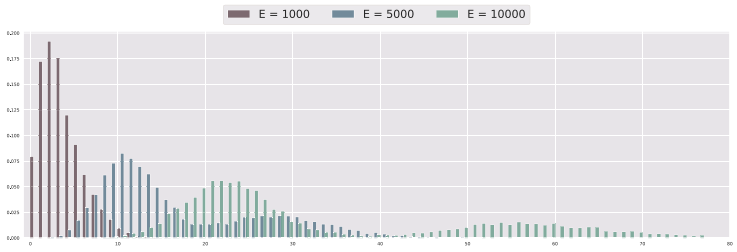

Fig. 3 Estimated pmf using a DP mixture of Poisson kernels. Colors correspond to levels of exposure.

4 Conclusion

In this application we analyzed a group life insurance portfolio, where for each group the number of deaths and the overall exposure were recorded. We showed how the heterogeneity between groups required a nonparametric approach and, to address this issue, we proposed a flexible Pitman-Yor process mixture of Rounded Gaussian kernels. To compare this model with a more "classical" approach, we also estimated a Dirichlet Process mixture of Poisson kernels, although some previous results suggested the inadequacy of this model under many conditions. The application of these mixtures to our data further supported the use of the Rounded Gaussian kernel compared to the Poisson kernel, as the constraints on the variance of the latter led to unlikely flattened and over-dispersed mixture components.

References

1. Brown G.O., Buckley W.S.: Experience rating with Poisson mixtures, Annals of Actuarial Science **2** 304321 (2015)
2. Canale A., Dunson D.: Bayesian Kernel Mixtures for Counts. J. Am. Stat. Assoc. **106** 1528–1539 (2011)
3. Canale A., Pruenster I.: Robustifying Bayesian nonparametric mixtures for count data. Biometrics **73** 174–184 (2017)
4. Ferguson T.S.: A Bayesian Analysis of Some Nonparametric Problems. Ann. Statist. 1 **2** 209–230 (1973)
5. Ferguson T.S.: Prior Distributions on Spaces of Probability Measures. Ann. Statist. 2 **4** 615–629 (1974)
6. Haastrup, S.: Comparison of Some Bayesian Analyses of Heterogeneity in Group Life Insurance. Scand. Actuar. J. **1** 2–16 (2000)
7. Norberg, R.: Experience Rating in Group Life Insurance. Scand. Actuar. J. **4**, 194–224 (1989)
8. Pitman J., Yor M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. 25 **2** 855–900 (1997)