Computational Statistics and Data Analysis 93 (2016) 97–106

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On the estimation of mixtures of Poisson regression models with large number of components

Panagiotis Papastamoulis [a,*], Marie-Laure Martin-Magniette [b,c], Cathy Maugis-Rabusseau [d]

[a] INRA UMR 1165, CNRS ERL 8196, UEVE, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry, France

[b] INRA, Unité de Recherche en Génomique Végétale, UMR 1165, ERL CNRS 8196, Saclay Plant Sciences, CP 5708, F-91057 Evry, France

[c] UEVE, Unité de Recherche en Génomique Végétale, UMR 1165, ERL CNRS 8196, Saclay Plant Sciences CP 5708, F-91057 Evry, France

[d] Institut de Mathématiques de Toulouse, INSA de Toulouse, Département de Génie Mathématique, 135, avenue de Rangueil, 31077 Toulouse Cedex 4, France

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form 8 July 2014
Accepted 10 July 2014
Available online 21 July 2014

Keywords:
Mixtures of distributions
EM algorithm initialization
Multimodal likelihood
Clustering

ABSTRACT

Modelling heterogeneity in large datasets of counts under the presence of covariates demands advanced clustering methods. Towards this direction a mixture of Poisson regressions is proposed. Conditionally on the covariates and a cluster, the multivariate distribution is a product of independent Poisson distributions. A variety of different parameterizations is taken into account for the slope of the conditional log-means. Also considered is the case of partitioning the response variables into sets of replicates sharing the same conditional log-mean up to an additive constant. Model parameters are estimated via an Expectation–Maximization algorithm with Newton–Raphson steps. In particular, an efficient initialization is introduced in order to improve the inference: a splitting scheme is combined with a Small-EM strategy. Simulations and application on two real high-throughput sequencing datasets highlight improvements of parameter estimations. The proposed methodology is implemented in the R package `poisson.glm.mix`, available on CRAN.

1. Introduction

Finite mixture models are a powerful tool for approximating unknown distribution functions in a semiparametric way, as well as to account for unobserved heterogeneity. In past decades both frequentist and Bayesian approaches to estimate a mixture model have become feasible due to the available computing power. From a frequentist point of view, the Expectation–Maximization (EM) algorithm (Dempster et al., 1977) is the most standard approach (see McLachlan and Peel, 2000). Mixture estimation within a Bayesian framework has become popular with advent of Markov Chain Monte Carlo methods, an overview is given in Frühwirth-Schnatter (2006).

An obvious extension of standard mixtures is to include covariates. When observations are counts, it leads to estimate a generalized linear model (Nelder and Wedderburn, 1972) for each component. Mixtures of Poisson regressions have been applied in quality control, as well as in biology and medicine. Aitkin (1996) fitted these models in order to describe the

* Corresponding author.
E-mail addresses: papapast@yahoo.gr, panagiotis.papastamoulis@manchester.ac.uk (P. Papastamoulis), marie_laure.martin@agroparistech.fr (M.-L. Martin-Magniette), cathy.maugis@insa-toulouse.fr (C. Maugis-Rabusseau).

http://dx.doi.org/10.1016/j.csda.2014.07.005
0167-9473/

number of faults in a bolt of fabric in terms of its length, while in a medical context, Wang et al. (1996) presented two examples (seizure frequency and Ames salmonella assay data) with covariate dependent rates. To describe count data with an excess of zeros (see for example Lambert, 1992; Cui and Yang, 2009), zero-inflated Poisson mixture models (with or without covariates) are often devised.

To the best of our knowledge, applications of mixtures of Poisson regressions are usually done on datasets with a reasonable size and a small number of components are usually required to fit well the data. However, a new field of application of such models emerges with technologies of high-throughput sequencing of RNA (RNA-seq) in molecular biology. These technologies allow one to sequence transcripts of genes, followed by a bioinformatic analysis resulting in a sequence of counts per gene. The complex nature of these biological datasets (high-dimensional, highly skewed with dynamic range from zero to 10^5) combined with the genuine multimodality of mixture log-likelihoods impose certain new inferential difficulties. Leisch (2004) and Grün and Leisch (2008b) have developed the `R` package `flexmix` to fit a variety of mixture models but it does not seem to be designed to take the aforementioned complexities into account (see Section 4 in the Supplementary material).

Initialization and ability of the EM algorithm to recover the overall mode of the log-likelihood remains a challenging task. The contribution of this paper is to efficiently estimate mixtures of Poisson regressions when the dataset is sizeable and the number of components could be greater than 10. To overcome such difficulties, an adaptive initialization scheme is proposed: starting from a mixture with a small number of components, the EM algorithm with a Small-EM strategy (Biernacki et al., 2003) is performed to estimate the parameters. Then a mixture with an additional component is estimated by an EM algorithm which exploits information provided by the previous mixture by using a splitting scheme combined with a Small-EM strategy. It is demonstrated by simulations that this estimation procedure avoids local maxima and leads to efficient estimation of parameters of mixtures of Poisson regressions with a large number of components.

The rest of the paper is organized as follows. The proposed mixture model is introduced in Section 2. Section 3 is devoted to parameter estimation. The EM algorithm, along with two initialization schemes is discussed. The proposed method is illustrated in Section 4 using a set of simulation studies and one RNA-seq dataset. The paper concludes with a brief discussion in Section 5. Several details of the initialization schemes, analyses of additional synthetic and real datasets as well as some technical aspects about the maximization step of the EM algorithm are provided in the Supplementary material (see Appendix A).

2. Model parameterization

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ denote n independent observations, where $\boldsymbol{y}_i = \{y_{ij\ell}; j = 1, \ldots, J, \ell = 1, \ldots, L_j\}, \boldsymbol{y}_i \in \mathbb{N}^d, i = 1, \ldots, n$, with $d = \sum_{j=1}^{J} L_j$ and $L_j \geq 1, j = 1, \ldots, J$. In the RNA-seq analysis framework, an observation corresponds to a gene and $y_{ij\ell}$ is a count which measures the expression of gene i in replicate ℓ of condition j. L_j denotes the number of replicates of the jth condition, and the partition of the d variables among the J conditions is determined by the experimental design. In addition to $\boldsymbol{y}$, consider that a vector of V covariates is observed, denoted by $\boldsymbol{x}_i := \{x_{iv}; v = 1, \ldots, V\}, x_i \in \mathbb{R}^V$, for all $i = 1, \ldots, n$. We assume that $\boldsymbol{x}^T\boldsymbol{x}$ is a full rank matrix.

Given a model indicator m taking values in a discrete set and a positive integer K, the distribution of the response $\boldsymbol{y}_i$, conditionally on $\boldsymbol{x}_i$, is a mixture of K distributions, that is,

$$\boldsymbol{y}_i|\boldsymbol{x}_i, m, K \sim \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \prod_{\ell=1}^{L_j} \mathcal{P}(\mu_{ij\ell k;m}),$$

where $\mathcal{P}$ denotes the Poisson distribution. Moreover, $\{y_{ij\ell}; j = 1, \ldots, J, \ell = 1, \ldots, L_j\}$ are assumed to be independent conditionally on the mixture component. The vector $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_K), \pi_k > 0, k = 1, \ldots, K, \sum_{k=1}^{K} \pi_k = 1$ contains the weight of each cluster. Index $m \in \{1, 2, 3\}$ defines a series of parameterizations for the Poisson means $\mu_{ij\ell k;m}$:

$$m = 1 \Rightarrow \ln(\mu_{ij\ell k;1}) = \alpha_{jk} + \gamma_{j\ell} + \sum_{v=1}^{V} \beta_{jkv} x_{iv} \tag{1}$$

$$m = 2 \Rightarrow \ln(\mu_{ij\ell k;2}) = \alpha_{jk} + \gamma_{j\ell} + \sum_{v=1}^{V} \beta_{jv} x_{iv} \tag{2}$$

$$m = 3 \Rightarrow \ln(\mu_{ij\ell k;3}) = \alpha_{jk} + \gamma_{j\ell} + \sum_{v=1}^{V} \beta_{kv} x_{iv}, \tag{3}$$

for $i = 1, \ldots, n, j = 1, \ldots, J, \ell = 1, \ldots, L_j$ and $k = 1, \ldots, K$. Parameterizations (2) and (3) are special cases of (1). If $J = 1$ then Parameterizations (1) and (3) coincide. Finally, for identifiability purposes, we assume for all $j = 1, \ldots, J$

$$\sum_{\ell=1}^{L_j} \gamma_{j\ell} = 0. \tag{4}$$