Composite Likelihood concept and sampling for count data

Fotini Panagou and Dimitris Karlis

Athens University of Economics & Business

Introduction

Multivariate Poisson for Count data and Composite Likelihood concept

Let $X = (X_1, X_2, ..., X_m)^T$ be a vector of random variables which follows a Multivariate Poisson distribution with parameters $\theta = (\theta_1, \theta_2, \dots, \theta_q)$. Assume the transformation: $AX = Y$ where $Y = (Y_1, Y_2, ..., Y_q)$, $q = m + \binom{m}{2} \geq m$ and $Y_r \sim Poisson(\theta_r)$, $r = 1, ..., q$.

$A = [A_1, A_2]$ is a $m \times q$ matrix where A_1 is the identity matrix with dimension m and A_2 is an $m \times (m-1)/2$ matrix where each of its columns contain exactly 2 ones and there are no duplicate columns. The generating function of $X's$ is given by the following expression:

$$g(s) = exp\left\{\sum_{t=1}^{m} \theta_t(s_t - 1) + \sum_{t=1}^{m-1} \sum_{j=t+1}^{m} \theta_{ij}(s_t s_j - 1)\right\},$$

where $s = (s_1, s_2, ...s_m)$ and marginal distributions can be found by setting $si = 1$ to the appropriate index i. The joint pmf is too complicated and involves multiple summation.

Composite likelihood is a tool that allows less computational effort.The likelihood needs to be maximized is:

$$\mathcal{L}(X, \theta) = \prod_{k=1}^{K} MP_m(X_k|\theta) \tag{1}$$

Instead we focus on maximizing the composite likelihood below:

$$\mathcal{L}_{CL}(X, \theta) = \prod_{k=1}^{K} \prod_{t=1}^{m-1} \prod_{j=t+1}^{m} BP_2(X_{ik}, X_{jk}|\theta_{CL}) \tag{2}$$

where θ_{CL} are appropriate parameters for each model and $BP_2(X_{ik}, X_{jk}|\theta_{CL})$ is the pmf of the bivariate Poisson distribution.

Sampling approach

In order to reduce the computational effort of maximizing the proposed CL likelihood, we introduce a technique that uses sampling to the pairwise bivariate Poisson pmf's as described in the simulation study below. The idea is to reduce the number of pairs need to be evaluated. We present preliminary results based on simulation.

Simulation study

Assume $X = (X_1, X_2, X_3) \sim MP_3(\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$ and the marginal log likelihoods $l_{CL_j} = log(BP_2(X_{1t}, X_{2t}|uj))$ (3), u_j the corresponding θ's related to the random vector, $j = 1, 2, 3$.

Model 1: Maximize the full likelihood of equation 1 of the 3-variate poisson.

Model 2: Maximize the full composite likelihood of equation (2).

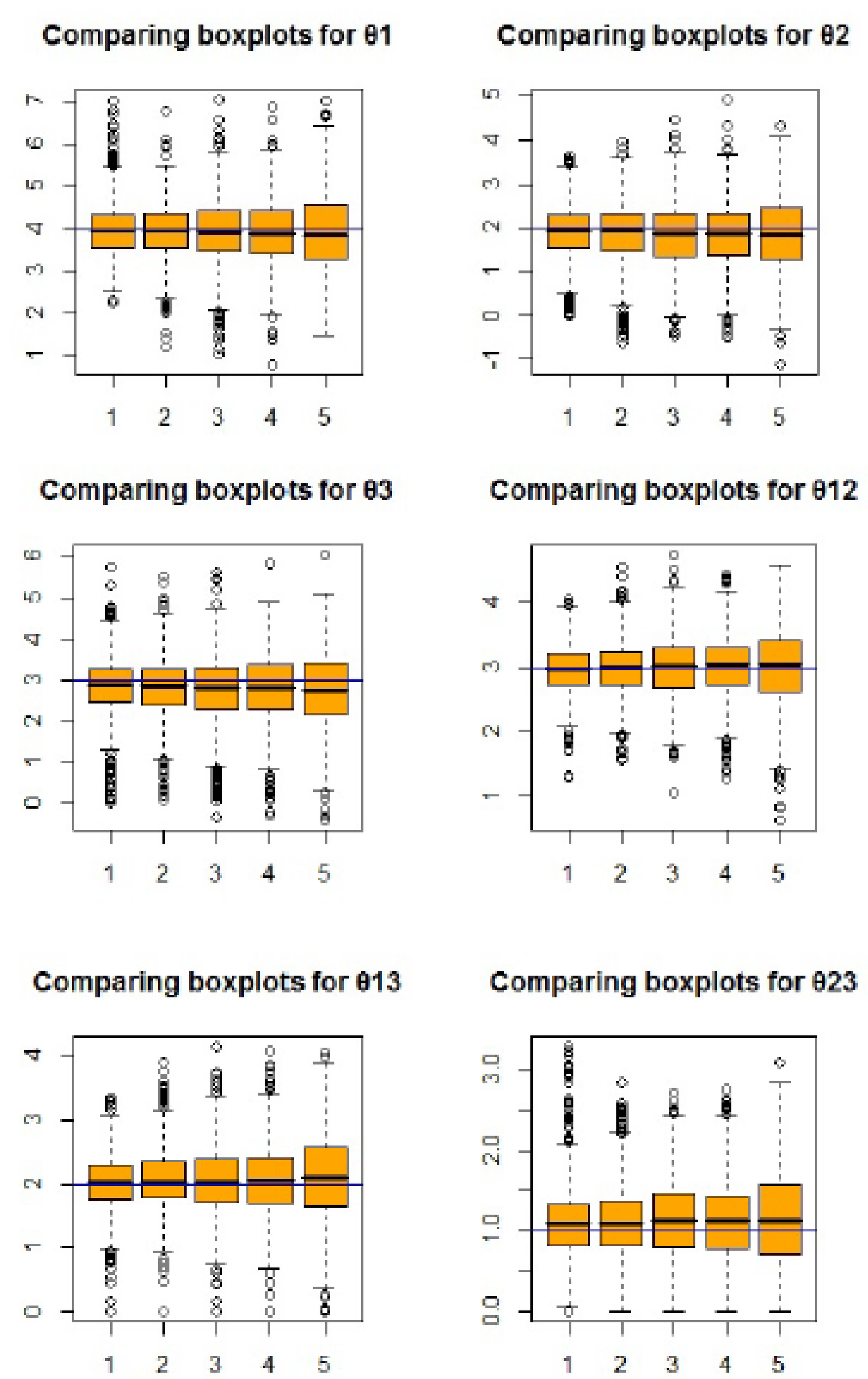
Model 3: Maximize the sampled composite likelihood by choosing 2 out of the 3 pairs of marginal log-likelihoods l_{CL_j} described in equation (3).

Model 4: Maximize the sampled composite likelihood by choosing pairs of marginal log-likelihoods via a bernoulli distribution with probability $p = 2/3$.

Model 5: Maximize the sampled composite likelihood by choosing 1 out of the 3 pairs of marginal log-likelihoods l_{CL_j} described in equation (3).

By performing 1,000 iterations of the simulation study results of the estimated parameters are shown in the attached graph.

Simulation Results



Remarks

The sampling approach provides results close to the actual values and the Composite likelihood parameters and there is no significant loss of efficiency of estimators.

Furthermore, the computational effort has been reduced up to 25% for the Bernoulli sampling, up to 100% for model 3 and up to 400% for model 5.

Next steps

Further considerations of the concept are:

- The sampling concept needs to be studied further for mixtures of Multivariate Poisson and with the use of EM & ECM algorithm.
- Extend the results to other multivariate distributions such as multivariate normal or copula based multivariate models: all we need is the bivariate marginal models
- Extend the results to other mixtures of distributions as for example to case of models with Copulas for mixed mode data.
- Examine/derive the properties of the estimators based on this sampling

References

[1] Alfo M. and Trovato G. (2004), Semi-parametric mixture models for multivariate count data with application, The Econometrics Journal, 7(2), 426-454

[2] Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. Journal of the American Statistical Association, 105(492), 1531-1540

[3] Karlis, D. (2005), EM algorithm for mixed Poisson and other discrete distributions, Astin Bulletin, 35(1), 3-24.