1 Conway-Maxwell Poisson (COM-Poisson, CMP) and Properties

The p.m.f. for Conway-Maxwell Poisson (CMP) is:

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}$$

$Z(\lambda, \nu) = Z$ is the normalization constant, i.e. $Z(\lambda, \nu) = Z = \sum_{y=0}^\infty \frac{\lambda^y}{(y!)^\nu}$, which doesn't have closed form in general. The domain for parameters is $\lambda, \nu > 0$ and $0 < \lambda < 1, \nu = 0$. The parameter ν controls the dispersion: 1) when $\nu = 1$, the CMP is Poisson distribution, 2) when $\nu < 1$, the distribution is over-dispersed and 3) when $\nu > 1$, the distribution is under-dispersed. When $\nu \to \infty$, the CMP approaches a Bernoulli distribution, while $\nu = 0$, it reduces to a geometric distribution.

In neuroscience, modeling dispersion is nearly as important as modeling mean, and both under-and over-dispersion are common. Therefore, I will consider modeling parameters by 2 linear models, i.e. $log(\lambda) = \boldsymbol{x}'\boldsymbol{\beta}$ and $log(\nu) = \boldsymbol{g}'\boldsymbol{\gamma}$, in this manuscript. To write things more compactly, define $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')$, then:

$$\begin{pmatrix} log(\lambda) \\ log(\nu) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}' & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{g}' \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}$$

To derive the SSP for CMP regression, the keys are gradient and Hessian for log-likelihood. In the following part of this section, I will define some notations and give some necessary properties for CMP.

Assume there are n independent $Y_i \sim CMP(\lambda_i, \nu_i)$. Denote $\boldsymbol{\eta}_i = (log(\lambda_i), \nu_i)'$, $Z_i = Z(\lambda_i, \nu_i)$ The log-likelihood for i^{th} observation is:

$$l_i(\boldsymbol{\eta}_i) = y_i log(\lambda_i) - log(y_i!)\nu_i - log(Z_i)$$

Since $E(Y_i) = \frac{\partial log(Z_i)}{\partial log(\lambda_i)}$, $Var(Y_i) = \frac{\partial^2 log(Z_i)}{\partial log(\lambda_i)^2}$, $E(log(Y_i!)) = -\frac{\partial log(Z_i)}{\partial \nu_i}$, $Var(log(Y_i!)) = \frac{\partial^2 log(Z_i)}{\partial \nu_i^2}$ and $Cov(Y_i, log(Y_i!)) = -\frac{\partial^2 log(Z_i)}{\partial log(\lambda_i)\partial \nu_i}$, the gradient for l_i (w.r.t $\boldsymbol{\eta}_i$) is:

$$\frac{\partial l_i}{\partial \boldsymbol{\eta}_i} = \begin{pmatrix} y_i - E(Y_i) \\ E(log(Y_i!)) - log(y_i!) \end{pmatrix}$$

And the Hessian for l_i (w.r.t $\boldsymbol{\eta}_i$) is:

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\eta_i}\partial \boldsymbol{\eta_i}'} = \begin{pmatrix} -Var(Y_i) & Cov(Y_i, log(Y_i!)) \\ Cov(Y_i, log(Y_i!)) & -Var(log(Y_i!)) \end{pmatrix}$$

The $E(Y_i)$, $Var(Y_i)$, $E(log(Y_i!))$, $Var(log(Y_i!))$ and $Cov(Y_i, log(Y_i!))$ can be evaluated numerically. For example, $E(Y_i) = \frac{1}{Z_i}\sum_{y=0}^\infty \frac{y(\lambda_i)^y}{(y!)^{\nu_i}}$. When $\nu_i \leq 1$

or $\lambda_i > 10^{\nu_i}$, these moments can be approximated well, for example $E(Y_i) \approx \lambda_i^{1/\nu_i} - \frac{\nu_i - 1}{2\nu_i}$.

However, these conditions are not held for neuroscience. In neuroscience, the observations are usually in the scale of milliseconds. When there's no information input, the observations are pretty sparse (therefore previously some researchers suggested to use zero-inflated Poisson). But when there's information inputs, the observation will usually produce a peak. In other words, λ_i is usually very small. Moreover, both over- and under-dispersion are common. But that make numerical summation easy, since λ is small while ν is not very small, the numerical approximation can be done in a few steps.

Note that $\frac{\partial log(\lambda_i)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{x_i} \\ \boldsymbol{0} \end{pmatrix}$ and $\frac{\partial \nu_i}{\partial \boldsymbol{\theta}} = \nu_i \frac{\partial log(\nu_i)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{0} \\ \nu_i \boldsymbol{g_i} \end{pmatrix}$. Therefore, the gradient for l_i (w.r.t $\boldsymbol{\theta}$) is:

$$\frac{\partial l_i}{\partial \boldsymbol{\theta}} == \begin{pmatrix} [y_i - E(Y_i)]\boldsymbol{x_i} \\ \nu_i[E(log(Y_i!)) - log(y_i!)]\boldsymbol{g_i} \end{pmatrix}$$

And the Hessian for l_i (w.r.t $\boldsymbol{\theta}$) is:

$$\frac{\partial^2 l_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} -Var(Y_i)\boldsymbol{x_i}\boldsymbol{x_i}' & \nu_i Cov(Y_i, log(Y_i!))\boldsymbol{x_i}\boldsymbol{g_i}' \\ \nu_i Cov(Y_i, log(Y_i!))\boldsymbol{g_i}\boldsymbol{x_i}' & -\nu_i[\nu_i Var(log(Y_i!)) - E(log(Y_i!)) + log(y_i!))]\boldsymbol{g_i}\boldsymbol{g_i}' \end{pmatrix}$$

2 SSP for CMP

Denote subsample size as r, with subsampling probabilities π_i for all dta points. Denote the covariates in the subsample as $\boldsymbol{x}_i^*$ and $\boldsymbol{g}_i^*$, while responses and subsampling probabilities in the subsample as y_i^* and π_i^*. Denote the weighted log-likelihood as:

$$l^*(\boldsymbol{\theta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[y_i^* log(\lambda_i(\boldsymbol{\theta})) - log(y_i^*!)\nu_i(\boldsymbol{\theta}) - log(Z_i(\boldsymbol{\theta}))]$$

To simplify notations, denote $E(Y_i|\boldsymbol{x_i^*}, \boldsymbol{g_i^*}, \boldsymbol{\theta}) = E^*(Y_i|\boldsymbol{\theta})$ and $E(Y_i|\boldsymbol{x_i}, \boldsymbol{g_i}, \boldsymbol{\theta}) = E(Y_i|\boldsymbol{\theta})$. Similar for $E(log(Y_i!))$, $Var(Y_i)$, $Var(log(Y_i!))$ and $Cov(Y_i, log(Y_i!))$ Note that

$$\frac{1}{n}\frac{\partial l^*(\hat{\boldsymbol{\theta}}_{MLE})}{\partial \hat{\boldsymbol{\theta}}_{MLE}} = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\begin{pmatrix} [y_i^* - E^*(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x_i^*} \\ \nu_i^*(\hat{\boldsymbol{\theta}}_{MLE})[E^*(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i^*!)]\boldsymbol{g_i^*} \end{pmatrix} = \frac{1}{r}\sum_{i=1}^{r}\boldsymbol{\zeta}_i$$

Let $a = [y_i - E(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]$ and $b = \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i!)]$.

Given the full data matrix $\mathcal{F}_n = (\boldsymbol{X}, \boldsymbol{y})$, $\boldsymbol{\zeta}_1, \dots \boldsymbol{\zeta}_n$ are i.i.d, with mean $\boldsymbol{0}$ and variance:

$$Var(\boldsymbol{\zeta_i}|\mathcal{F}_n) = r\boldsymbol{V}_c = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\begin{pmatrix} a^2\boldsymbol{x_i}\boldsymbol{x_i}' & ab\boldsymbol{x_i}\boldsymbol{g_i}' \\ ba\boldsymbol{g_i}\boldsymbol{x_i}' & b^2\boldsymbol{g_i}\boldsymbol{g_i}' \end{pmatrix}$$

2

where $\boldsymbol{V}_c = \frac{1}{rn^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \begin{pmatrix} a^2 \boldsymbol{x_i} \boldsymbol{x_i}' & ab\boldsymbol{x_i}\boldsymbol{g_i}' \\ ba\boldsymbol{g_i}\boldsymbol{x_i}' & b^2 \boldsymbol{g_i}\boldsymbol{g_i}' \end{pmatrix}$. Follow the same steps in OS-MAC,

$$\frac{1}{n} \boldsymbol{V}_c^{-1/2} \frac{\partial l^*(\hat{\boldsymbol{\theta}}_{MLE})}{\partial \hat{\boldsymbol{\theta}}_{MLE}} \to N(0, \boldsymbol{I})$$

Denote

$$\boldsymbol{M}_X = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where

$A = Var(Y_i|\hat{\boldsymbol{\theta}}_{MLE})\boldsymbol{x_i}\boldsymbol{x_i}'$

$B = -\nu_i(\hat{\boldsymbol{\theta}}_{MLE})Cov(Y_i, log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE})\boldsymbol{x_i}\boldsymbol{g_i}'$

$C = -\nu_i(\hat{\boldsymbol{\theta}}_{MLE})Cov(Y_i, log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE})\boldsymbol{g_i}\boldsymbol{x_i}'$

$D = \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[\nu_i(\hat{\boldsymbol{\theta}}_{MLE})Var(log(Y_i!|\hat{\boldsymbol{\theta}}_{MLE}) - E(log(Y_i!|\hat{\boldsymbol{\theta}}_{MLE})) + log(y_i!))]\boldsymbol{g_i}\boldsymbol{g_i}'$

(Since $\boldsymbol{M}_X$ converges to Fisher information matrix, we may replace D with $D* = \nu_i(\hat{\boldsymbol{\theta}}_{MLE})^2 Var(log(Y_i!|\hat{\boldsymbol{\theta}}_{MLE})\boldsymbol{g_i}\boldsymbol{g_i}')$. Since $D*$ doesn't include $log(y_i!)$, it will be more robust to the outliers.)

Still follow the same steps as in OSMAC, we can show that as $n \to \infty$ and $r \to \infty$, conditional on $\mathcal{F}_n$ in probability,

$$\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \to N(0, \boldsymbol{I})$$

where $\boldsymbol{V} = \boldsymbol{M}_X^{-1} \boldsymbol{V}_c \boldsymbol{M}_X^{-1}$

2.1 SSP under A-optimality Criteiron

$$tr(V) = tr(\boldsymbol{M}_X^{-1} \boldsymbol{V}_c \boldsymbol{M}_X^{-1})$$

$$= \frac{1}{rn^2} \sum_{i=1}^{n} \frac{1}{\pi_i} tr(\boldsymbol{M}_X^{-1} \begin{pmatrix} a^2 \boldsymbol{x_i} \boldsymbol{x_i}' & ab\boldsymbol{x_i}\boldsymbol{g_i}' \\ ba\boldsymbol{g_i}\boldsymbol{x_i}' & b^2 \boldsymbol{g_i}\boldsymbol{g_i}' \end{pmatrix} \boldsymbol{M}_X^{-1})$$

$$= \frac{1}{rn^2} [\sum_{i=1}^{n} \pi_i][\sum_{i=1}^{n} \frac{1}{\pi_i} ||\boldsymbol{M}_X^{-1} \begin{pmatrix} a\boldsymbol{x_i} \\ b\boldsymbol{g_i} \end{pmatrix} ||^2]$$

$$\geq \frac{1}{rn^2} [\sum_{i=1}^{n} ||\boldsymbol{M}_X^{-1} \begin{pmatrix} a\boldsymbol{x_i} \\ b\boldsymbol{g_i} \end{pmatrix} |||^2$$

By equality condition for Cauchy-Schwartz inequality, $\pi_i \propto ||\boldsymbol{M}_X^{-1} \begin{pmatrix} a\boldsymbol{x_i} \\ b\boldsymbol{g_i} \end{pmatrix} || = $

$$||\boldsymbol{M}_X^{-1} \begin{pmatrix} [y_i - E(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_i \\ \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i!)]\boldsymbol{g}_i \end{pmatrix} ||$$ So, the SSP is:

$$\pi_i^{mMSE} = \frac{||\boldsymbol{M}_X^{-1} \begin{pmatrix} [y_i - E(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_i \\ \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i!)]\boldsymbol{g}_i \end{pmatrix} ||}{\sum_{j=1}^n ||\boldsymbol{M}_X^{-1} \begin{pmatrix} [y_j - E(Y_j|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_j \\ \nu_j(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_j!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_j!)]\boldsymbol{g}_j \end{pmatrix} ||}$$

2.2 SSP under L-optimality Criteiron

$$tr(\boldsymbol{V}_c) = \frac{1}{rn^2}[\sum_{i=1}^n \pi_i][\sum_{i=1}^n \frac{1}{\pi_i}|| \begin{pmatrix} a\boldsymbol{x}_i \\ b\boldsymbol{g}_i \end{pmatrix} ||^2]$$

$$\geq \frac{1}{rn^2}[\sum_{i=1}^n || \begin{pmatrix} a\boldsymbol{x}_i \\ b\boldsymbol{g}_i \end{pmatrix} ||]^2$$

By equality condition for Cauchy-Schwartz inequality, $\pi_i \propto || \begin{pmatrix} [y_i - E(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_i \\ \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i!)]\boldsymbol{g}_i \end{pmatrix} ||$
So, the SSP is:

$$\pi_i^{mVc} = \frac{|| \begin{pmatrix} [y_i - E(Y_i|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_i \\ \nu_i(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_i!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_i!)]\boldsymbol{g}_i \end{pmatrix} ||}{\sum_{j=1}^n || \begin{pmatrix} [y_j - E(Y_j|\hat{\boldsymbol{\theta}}_{MLE})]\boldsymbol{x}_j \\ \nu_j(\hat{\boldsymbol{\theta}}_{MLE})[E(log(Y_j!)|\hat{\boldsymbol{\theta}}_{MLE}) - log(y_j!)]\boldsymbol{g}_j \end{pmatrix} ||}$$