Lognormal and Gamma Mixed Negative Binomial Regression: Supplementary Material

Mingyuan Zhou, Lingbo Li, David Dunson and Lawrence Carin

Duke University, Durham, NC 27708, USA

I. LEMMA1

Lemma 1: Denote $f(z) = -\ln(1-pz)$ and $\mathbf{F}$ as a lower triangular matrix with $F(1,1) = 1$, $F(m,j) = 0$ if $j > m$ and

$$F(m,j) = \frac{m-1}{m}F(m-1,j) + \frac{1}{m}F(m-1,j-1) \tag{1}$$

if $1 \le j \le m$, then we have

$$\frac{1}{m!}\frac{d^m}{dz^m}f^j(z)\Big|_{z=0} = \sum_{j'=1}^{j} F(m,j')\frac{j!}{(j-j')!}[f'(z)]^m f^{j-j'}(z)\Big|_{z=0} = F(m,j)j!p^m \tag{2}$$

Proof: The first and second derivatives of $f(z) = -\ln(1-pz)$ are $f'(z) = \frac{p}{1-pz}$ and $f''(z) = \frac{p^2}{(1-pz)^2} = (f'(z))^2$, respectively, and $f^j(0) = 0$ for $j \ge 1$, therefore $\frac{d^m}{dz^m}f^j(z)\big|_{z=0} = 0$ for $j > m$ and thus (2) is true for $j > m$.

When $n = j = 1$, we have $\frac{d^n}{dz^n}f^j(z)\big|_{z=0} = f_k'(0) = p$, $\sum_{j'=1}^{n} F(n,j')\frac{j!}{(j-j'!)}[f'(z)]^n f^{j-j'}(z)\Big|_{z=0} = p$, and $F(n,j)j!p^n = p$, therefore, (2) is true for $n = m = 1$.

Assume (2) is true for $n = m-1$, then

$$\begin{aligned}
\frac{1}{m!}\frac{d^m}{dz^m}f^j(z)\Big|_{z=0} &= \frac{1}{m}\left.\frac{d\frac{1}{(m-1)!}\frac{d^{m-1}}{dz^{m-1}}f^j(z)}{dz}\right|_{z=0} \\
&= \frac{1}{m}\sum_{j'=1}^{j} F(m-1,j')\frac{j!}{(j-j')!}\frac{d}{dz}\left\{[f'(z)]^{m-1}f^{j-j'}(z)\right\}\Big|_{z=0} \\
&= \sum_{j'=1}^{j}\left[\frac{m-1}{m}F(m-1,j') + \frac{1}{m}F(m-1,j'-1)\right]\frac{j!}{(j-j')!}[f'(z)]^m f^{j-j'}(z)\Big|_{z=0} \\
&= \sum_{j'=1}^{j} F(m,j')\frac{j!}{(j-j')!}[f'(z)]^m f^{j-j'}(z)\Big|_{z=0} \\
&= F(m,j)j!p^m.
\end{aligned}$$

Therefore, (2) is also true for $n = m$. ■

II. Univariate Count Data Analysis

Assume N iid samples $\boldsymbol{y} = \{y_i, i = 1, \cdots, N\}$ are drawn from the distribution NB(r, p), with a gamma distribution prior Gamma$(r; a, 1/b) = \frac{b^a}{\Gamma(a)} r^{a-1} e^{-br}$ on r, and a beta distribution prior on p. This constitutes the hierarchical model

$$y_i \overset{iid}{\sim} \text{NB}(r, p),\ i = 1, \cdots, N \tag{3}$$

$$r \sim \text{Gamma}(a, 1/b) \tag{4}$$

$$p \sim \text{Beta}(\alpha, \beta). \tag{5}$$

A. Point Estimation Methods for the Negative Binomial Dispersion Parameter r

Parameterizing the NB distribution with the mean $\mu = rp/(1-p)$ and inverse dispersion parameter ϕ (the reciprocal of r), for N iid observations $y_i \sim \text{NB}(\mu, \phi),\ i = 1, \cdots, N$, the log-likelihood function can be expressed as

$$\ell(\mu, \phi) = \sum_{i=1}^{N} \left[\sum_{j=0}^{y_i - 1} \ln(1 + j\phi) + y_i \log \mu - (y_i + \phi^{-1}) \ln(1 + \phi\mu) - \ln(y_i!) \right]. \tag{6}$$

Setting $\frac{\partial \ell(\mu, \phi)}{\partial \mu} = 0$, the ML estimate of μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{7}$$

and setting $\frac{\partial \ell(\hat{\mu}, \phi)}{\partial \phi} = 0$, the ML estimate of ϕ is found by solving

$$\frac{\partial \ell(\hat{\mu}, \phi)}{\partial \phi} = \sum_{i=1}^{N} \left[-\sum_{j=0}^{y_i - 1} \frac{1}{\phi(1 + j\phi)} + \frac{y_i - \hat{\mu}}{\phi(1 + \phi\hat{\mu})} + \phi^{-2} \ln(1 + \phi\hat{\mu}) \right] = 0 \tag{8}$$

which does not have a closed form solution and is typically solved numerically with a nonlinear root finder (Piegorsch, 1990), such as the Newton-Raphson method.

Other popular methods include MM and MQL (Clark and Perry, 1989). The MM and MQL estimates of μ are as in (7). The MM estimate of ϕ is expressed as

$$\hat{\phi}_{\text{MM}} = \frac{\sum_{i=1}^{N} (y_i - \hat{\mu})^2 / (N - 1) - \hat{\mu}}{\hat{\mu}^2} \tag{9}$$

and the MQL estimate $\hat{\phi}_{\text{MQL}}$ is obtained by solving

$$\sum_{i=1}^{N} \left[\frac{1}{\phi^2} \log \left(\frac{1 + \phi\hat{\mu}}{1 + \phi y_i} \right) - \frac{y_i}{1 + \phi y_i} + \frac{1 + 6y_i}{2(\phi + 6 + 6\phi y_i)} - \frac{1}{2(\phi + 6)} \right] = 0. \tag{10}$$

Note that when the sample size is small, the mean is small or when the dispersion parameter is large, the MM, ML and MQL estimates may fail to converge or provide invalid estimates. In these circumstances, following Clark and Perry (1989); Piegorsch (1990), one typically imposes the restriction

$\hat{\phi} > -1/\max\{1, y_1, \cdots, y_N\}$. Confidence intervals for $\hat{\phi}$ of these point estimates may be obtained by assuming that $\hat{\phi}$ is asymptotically normally distributed with mean ϕ and variance var$(\hat{\phi})$ (Saha, 2011).

B. Gibbs Sampling

Gibbs sampling can proceed by alternately sampling from the following equations:

$$(p|-) \sim \text{Beta}\left(\alpha + \sum_{i=1}^{N} y_i, \beta + Nr\right) \tag{11}$$

$$\Pr(L_i = j) = R_r(y_i, j), \quad j = 0, \cdots, y_i \tag{12}$$

$$(r|-) \sim \text{Gamma}\left(a + \sum_{i=1}^{N} L_i, \frac{1}{b - N\ln(1-p)}\right). \tag{13}$$

We show in Figure 1 the top-left 50×50 submatrices of $\mathbf{R}_r$ for $r = .1$, 1, 10 and 100.

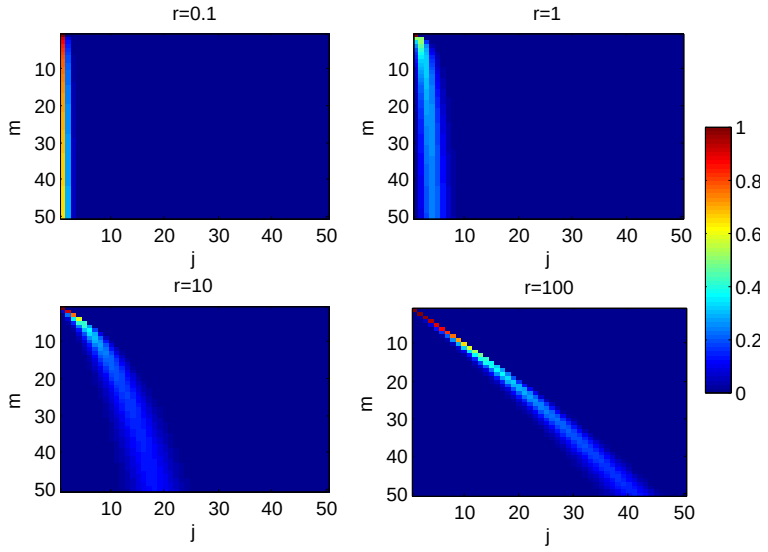

Fig. 1. The top-left 50×50 submatrices of $\mathbf{R}_r$ for $r = 0.1$, 1, 10 and 100. Note that $\mathbf{R}_r = \mathbf{F}$ when $r = 1$.

C. Variational Bayes Inference

Using variational Bayes (VB) inference (Beal, 2003; Bishop and Tipping, 2000), we approximate the posterior distribution $P(r, p, \mathbf{L}|\boldsymbol{y})$ with

$$Q(r, p, \mathbf{L}) = Q_r(r)Q_p(p)\prod_{i=1}^{N} Q_{L_i}(L_i)$$

and we seek to minimize the KL divergence $D_{KL}(Q||P)$. To exploit conjugacy, we define

$$Q_p(p) = \text{Beta}(\tilde{\alpha}, \tilde{\beta}) \tag{14}$$

$$Q_{L_i}(L_i) = \sum_{j=0}^{y_i} R_{\tilde{r}}(y_i, j)\delta_j \tag{15}$$

$$Q_r(r) = \text{Gamma}(\tilde{a}, 1/\tilde{b}) \tag{16}$$

where

$$\tilde{\alpha} = \alpha + \sum_{i=1}^{N} y_i, \quad \tilde{\beta} = \beta + N\langle r\rangle, \quad \tilde{r} = \exp\left(\langle \ln r\rangle\right) \tag{17}$$

$$\tilde{a} = a + \sum_{i=1}^{N} \langle L_i\rangle, \quad \tilde{b} = b - N\langle \ln(1-p)\rangle. \tag{18}$$

These moments can be calculated as

$$\langle r\rangle = \tilde{a}/\tilde{b}, \quad \langle \ln r\rangle = \psi(\tilde{a}) - \ln\tilde{b} \tag{19}$$

$$\langle L_i\rangle = \sum_{j=1}^{y_i} R_{\tilde{r}}(y_i, j)j, \quad \langle \ln(1-p)\rangle = \psi(\tilde{\beta}) - \psi(\tilde{\alpha}+\tilde{\beta}). \tag{20}$$

Equations (17)-(20) constitute the VB inference.

D. Variational Bayes Lower Bound

The VB lower bound can be calculated as

$$\begin{aligned}\mathcal{L} &= \langle \ln P(\mathbf{L}|\boldsymbol{y}, r)\rangle + \langle \ln P(\boldsymbol{y}|r, p)\rangle + \langle \ln P(r)\rangle + \langle \ln P(p)\rangle \\ &\quad - \langle \ln Q(\mathbf{L})\rangle - \langle \ln Q(r)\rangle - \langle \ln Q(p)\rangle \qquad (21)\\ &= \langle \ln P(\boldsymbol{y}|r, p)\rangle + \langle \ln P(r)\rangle + \langle \ln P(p)\rangle - \langle \ln Q(r)\rangle - \langle \ln Q(p)\rangle \qquad (22)\end{aligned}$$

where

$$\langle \ln P(\boldsymbol{y}|r,p)\rangle = N\langle r\rangle\langle \ln(1-p)\rangle + \langle \ln p\rangle \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \ln(y_i!) + \sum_{i=1}^{N} \left[\langle \ln\Gamma(y_i + r) - \ln\Gamma(r)\rangle\right] \tag{23}$$

$$\langle \ln P(r)\rangle = a\ln b + (a-1)\langle \ln r\rangle - b\langle r\rangle - \ln\Gamma(a) \tag{24}$$

$$\langle \ln Q(r)\rangle = \tilde{a}\ln\tilde{b} + (\tilde{a}-1)\langle \ln r\rangle - \tilde{b}\langle r\rangle - \ln\Gamma(\tilde{a}) \tag{25}$$

$$\langle \ln P(p)\rangle = (\alpha-1)\langle \ln p\rangle + (\beta-1)\langle \ln(1-p)\rangle - \ln \mathrm{B}(\alpha,\beta) \tag{26}$$

$$\langle \ln Q(p)\rangle = (\tilde{\alpha}-1)\langle \ln p\rangle + (\tilde{\beta}-1)\langle \ln(1-p)\rangle - \ln \mathrm{B}(\tilde{\alpha},\tilde{\beta}) \tag{27}$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function. Note that there are no analytical forms for expectations $\langle \ln\Gamma(y_i + r) - \ln\Gamma(r)\rangle$, $i = 1, \cdots, N$, for which the Monte Carlo integration (Andrieu et al., 2003) algorithm is used. Therefore, some variations of the calculated lower bound during iterations are expected.

III. Sampling from the Polya-Gamma Distribution

As in Polson and Scott (2011), a random variable $X \sim \text{PG}(a, c)$ has a Polya-Gamma distribution if

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \tag{28}$$

where each g_k is an independent gamma random variable: $g_k \sim \text{Gamma}(a, 1)$. Thus a PG distributed random variable can be generated from an infinite sum of weighted iid gamma random variables. A conventional sampling method is to truncate the infinite sum at a large number K as

$$\hat{X} = \frac{1}{2\pi^2} \sum_{k=1}^{K} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \tag{29}$$

which is guaranteed to be left biased.

To avoid the bias, we propose to sample X as

$$\tilde{X} \quad = \quad \frac{\tilde{a}_{K+1}}{2\pi^2} \sum_{k=1}^{K} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \tag{30}$$

where

$$\tilde{a}_{K+1} \quad = \quad \frac{\mathbb{E}[X]}{\mathbb{E}\left[\frac{1}{2\pi^2} \sum_{k=1}^{K} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)}\right]} \tag{31}$$

$$= \quad \frac{\frac{a}{2c} \tanh\left(\frac{c}{2}\right)}{\frac{1}{2\pi^2} \sum_{k=1}^{K} \frac{a}{(k-1/2)^2 + c^2/(4\pi^2)}} \tag{32}$$

Thus $\mathbb{E}[\tilde{X}] = \mathbb{E}[X]$. We set the truncation level as $K = 2000$ in all experiments. We find that a truncation level as small as $K = 20$ also works for all the examples we considered.

IV. Variational Bayes Lower Bound of the LGNB Regression Model

The VB lower bound of the LGNB model can be calculated as

$$\begin{aligned}\mathcal{L} = &\ln P(\boldsymbol{y}|r, \boldsymbol{\psi})\rangle + \langle \ln P(r|h)\rangle + \langle \ln P(h)\rangle + \langle \ln P(\boldsymbol{\psi}|\boldsymbol{\beta}, \varphi)\rangle + \langle \ln P(\boldsymbol{\beta}|\boldsymbol{\alpha})\rangle + \langle \ln P(\boldsymbol{\alpha})\rangle + \langle \ln P(\varphi)\rangle \\ &- \langle \ln Q(r)\rangle - \langle \ln Q(h)\rangle - \langle \ln Q(\boldsymbol{\psi})\rangle - \langle \ln Q(\boldsymbol{\beta}) - \langle \ln Q(\boldsymbol{\alpha})\rangle - \langle \ln Q(\varphi)\rangle \end{aligned} \tag{33}$$

where

$$\begin{aligned}
\langle \ln P(\boldsymbol{y}|r,\boldsymbol{\psi})\rangle &= -\langle r\rangle \sum_{i=1}^{N}\langle \ln(1+e^{\psi_i})\rangle - \sum_{i=1}^{N}\langle \ln(1+e^{-\psi_i})\rangle y_i - \sum_{i=1}^{N}\ln(y_i!) \\
&\quad + \sum_{i=1}^{N}[\langle \ln\Gamma(y_i+r)-\ln\Gamma(r)\rangle] && (34)\\
\langle \ln P(r|h)\rangle &= a_0\langle \ln h\rangle + (a_0-1)\langle \ln r\rangle - \langle h\rangle\langle r\rangle - \ln\Gamma(a_0) && (35)\\
\langle \ln Q(r)\rangle &= \tilde{a}\ln\tilde{h} + (\tilde{a}-1)\langle \ln r\rangle - \tilde{h}\langle r\rangle - \ln\Gamma(\tilde{a}) && (36)\\
\langle \ln P(h)\rangle &= b_0\ln g_0 + (b_0-1)\langle \ln h\rangle - g_0\langle h\rangle - \ln\Gamma(b_0) && (37)\\
\langle \ln Q(h)\rangle &= \tilde{b}\ln\tilde{g} + (\tilde{b}-1)\langle \ln h\rangle - \tilde{g}\langle h\rangle - \ln\Gamma(\tilde{b}) && (38)\\
\langle \ln P(\boldsymbol{\psi}|\boldsymbol{\beta},\varphi)\rangle &= \frac{N}{2}\langle \ln\varphi\rangle - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\langle\varphi\rangle\left\{\langle \boldsymbol{\psi}^T\boldsymbol{\psi}\rangle - 2\langle\boldsymbol{\psi}\rangle^T\mathbf{X}\langle\boldsymbol{\beta}\rangle + \mathrm{tr}[\mathbf{X}\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle\mathbf{X}^T]\right\} && (39)\\
\langle \ln Q(\boldsymbol{\psi})\rangle &= -N(1+\ln(2\pi))/2 - \ln|\tilde{\boldsymbol{\Sigma}}|/2 && (40)\\
\langle \ln P(\boldsymbol{\beta}|\boldsymbol{\alpha})\rangle &= -\frac{P+1}{2}\ln(2\pi) + \frac{1}{2}\sum_{p=0}^{P}\langle \ln\alpha_p\rangle - \frac{1}{2}\sum_{p=0}^{P}\langle\alpha_p\rangle\langle\beta_p^2\rangle && (41)\\
\langle \ln Q(\boldsymbol{\beta})\rangle &= -(P+1)(1+\ln(2\pi))/2 - \ln|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}|/2 && (42)\\
\langle \ln P(\boldsymbol{\alpha})\rangle &= (P+1)c_0\ln d_0 + (c_0-1)\sum_{p=0}^{P}\langle \ln\alpha_p\rangle - d_0\sum_{p=0}^{P}\langle\alpha_p\rangle - (P+1)\ln\Gamma(c_0) && (43)\\
\langle \ln Q(\boldsymbol{\alpha})\rangle &= \sum_{p=0}^{P}\left\{\tilde{c}_p\ln\tilde{d}_p + (\tilde{c}_p-1)\langle\ln\alpha_p\rangle - \tilde{d}_p\langle\alpha_p\rangle - \ln\Gamma(\tilde{c}_p)\right\} && (44)\\
\langle \ln P(\varphi)\rangle &= e_0\ln f_0 + (e_0-1)\langle\ln\varphi\rangle - f_0\langle\varphi\rangle - \ln\Gamma(e_0) && (45)\\
\langle \ln Q(\varphi)\rangle &= \tilde{e}\ln\tilde{f} + (\tilde{e}-1)\langle\ln\varphi\rangle - \tilde{f}\langle\varphi\rangle - \ln\Gamma(\tilde{e}). && (46)
\end{aligned}$$

Note that there are no analytical forms for expectations $\langle \ln\Gamma(y_i+r)-\ln\Gamma(r)\rangle$, $\langle\ln(1+e^{\psi_i})\rangle$ and $\langle\ln(1-e^{\psi_i})\rangle$, $i=1,\cdots,N$, for which the Monte Carlo integration algorithm is used. Therefore, some variations of the calculated lower bound during iterations are expected.

V. Calculating the Prediction and Quasi-Dispersion

In Gibbs sampling, the posteriors of the lognormal precision parameter $\varphi=\sigma^{-2}$, NB dispersion parameter r, and regression coefficients $\boldsymbol{\beta}$ are represented by S collected samples $\{\varphi^{(s)}, r^{(s)}, \boldsymbol{\beta}^{(s)}\}_{s=1,S}$; whereas in VB, they are represented by $Q_\varphi(\varphi)=\text{Gamma}(\tilde{e}, 1/\tilde{f})$, $Q_r(r)=\text{Gamma}(\tilde{a}, 1/\tilde{h})$ and $Q_{\boldsymbol{\beta}}=\mathcal{N}(\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$. With the collected Gibbs samples, given a covariate vector $\boldsymbol{x}_j$, we can calculate the posterior means of the prediction and quasi-dispersion as

$$\hat{\mu}_j = \frac{1}{S}\sum_{s=1}^{S} r^{(s)} e^{\boldsymbol{x}_j^T\boldsymbol{\beta}^{(s)}+\sigma^{2(s)}/2} \quad (47)$$

$$\hat{\kappa} = \frac{1}{S}\sum_{s=1}^{S}\{e^{(\sigma^2)^{(s)}}(1+1/r^{(s)})-1\}; \quad (48)$$

and with the VB Q functions, they can be calculated as

$$\hat{\mu}_j = \langle r \rangle \langle e^{\boldsymbol{x}_j^T \boldsymbol{\beta} + \sigma^2/2} \rangle \tag{49}$$

$$\hat{\kappa} = \langle e^{\sigma^2} \rangle (1 + \langle r^{-1} \rangle) - 1 \tag{50}$$

where $\langle r^{-1} \rangle$ is equal to $\tilde{h}/(\tilde{a}-1)$ if $\tilde{a} > 1$ (the mean of the inverse gamma distribution) and is set as $\tilde{h}/\tilde{a}$ otherwise, and $\langle e^{\boldsymbol{x}_j^T \boldsymbol{\beta} + \sigma^2/2} \rangle$ and $\langle e^{\sigma^2} \rangle$ are calculated with the Monte Carlo integration; and with the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$, we have point estimates as

$$\hat{\mu}_j = \exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}) \tag{51}$$

and $\hat{\kappa} = 0$ in the Poisson and $\hat{\kappa} = \hat{\phi}$ in the NB regression models.

REFERENCES

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 2003.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, UCL, 2003.

C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *UAI*, 2000.

S. J. Clark and J. N. Perry. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, 1989.

W. W. Piegorsch. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, 1990.

N. G. Polson and J. G. Scott. Default Bayesian analysis for multi-way tables: a data-augmentation approach. *arXiv:1109.4180v1*, 2011.

K. K. Saha. Interval estimation of the over-dispersion parameter in the analysis of one-way layout of count data. *Statistics in Medicine*, 2011.