Addendum to `vglmer`

Max Goplerud

July 2, 2020

Abstract

This addendum extends Goplerud (2020) to new algorithms implemented in `vglmer`. At present, results are shown for negative binomial outcomes. Future updates will include multinomial and linear likelihoods.

1 Negative Binomial

The generative model is standard and ensures that $E[y_i] = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$ while $Var(y_i) = E[y_i]\left(1 + E[y_i]/r\right)$. Thus, r is interpretable in the usual way as an dispersion parameter where $r \to \infty$ recovers the original Poisson model. This differs from other negative binomial implementations (e.g. Pillow and Scott 2012; Zhou et al. 2012) but matches standard practice in applied social scientific research. Note that this definition of r agrees with implementations in `R` (e.g. `glm.nb`) and with the "reciprocal dispersion" from `rstanarm`.

$$y_i \sim \mathrm{NB}\left(r, 1 - p_i\right) \quad p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)} \quad \psi_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\alpha} - \ln r \tag{1a}$$

$$f(y_i) = \frac{\Gamma(y_i + r)}{\Gamma(y_i + 1)\Gamma(r)} \frac{\exp(\psi_i)^{y_i}}{\left(1 + \exp(\psi_i)\right)^{y_i + r}} \tag{1b}$$

The same hierarchical structure is placed on $\boldsymbol{\alpha}$ as described in Goplerud (2020). Applying Polya-Gamma augmentation yields the following augmented likelihood:

$$f(y_i, \omega_i) = \frac{\Gamma(y_i + r)}{\Gamma(y_i + 1)\Gamma(r)} 2^{-(y_i + r)} \exp\left([y_i - r]/2\psi_i - \omega_i \psi_i^2\right) f_{PG}(\omega_i | y_i + r, 0) \tag{2}$$

Using $\boldsymbol{\theta}$ to collect $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\{\boldsymbol{\Sigma}_j\}$, and $\{\omega_i\}$, the complete data log-posterior can be expressed as follows where $p_0(r)$ is the prior placed on r and $p_0(\boldsymbol{\Sigma}_j)$ is the Inverse-Wishart prior on $\boldsymbol{\Sigma}_j$:

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}, r) = & \sum_{i=1}^{N} \ln \Gamma(y_i + r) - \ln \Gamma(r) - \ln \Gamma(y_i + 1) - (y_i + r) \ln(2) + \\
& \sum_{i=1}^{N} (y_i - r)/2 \left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\alpha} - \ln r\right) - \omega_i/2 \left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\alpha} - \ln r\right)^2 + \ln f_{PG}(\omega_i | y_i + r, 0) \\
& \left[\sum_{j=1}^{J} -G_j/2 \ln(|2\pi \boldsymbol{\Sigma}_j|) - \frac{1}{2} \left(\sum_{g=1}^{G_j} \boldsymbol{\alpha}_{j,g}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\alpha}_{j,g}\right) + \ln p_0(\boldsymbol{\Sigma_j})\right] + \ln p_0(r)
\end{aligned}
\tag{3}
$$

It is useful to characterize the ELBO in two ways; first, the main objective ELBO that depends on $q(\boldsymbol{\theta})$ and $q(r)$. Second, a "conditional" ELBO (ELBO^c) that holds r fixed and depends thus only on $q(\boldsymbol{\theta})$.

$$
\begin{aligned}
\mathrm{ELBO}(q(\boldsymbol{\theta}, q(r)) &= E_{q(r)} \left[\mathrm{ELBO}^c(q(\boldsymbol{\theta}); r) - \ln q(r)\right] \\
\mathrm{ELBO}^c(q(\boldsymbol{\theta}); r) &= E_{q(\boldsymbol{\theta})} \left[f(\boldsymbol{y}, \boldsymbol{\theta}, r) - \ln q(\boldsymbol{\theta})\right]
\end{aligned}
\tag{4}
$$

Considering first the conditional ELBO, it is clear this has updates of a nearly identical form to those in Goplerud (2020). Thus, for any fixed r, the conditional ELBO can be maximized. When the expectation is taken over r, however, the main ELBO becomes intractable because of both the log-gamma terms involving r but also the log-Polya-Gamma density ($\ln f_{PG}(\omega_i | y_i + r)$). To proceed, I outline two strategies.

First, one can assume that $q(r)$ has a degenerate point-mass distribution and thus perform variational EM with r in the 'M-Step' and all other parameters in the variational E-Step. One proceeds by optimizing $q(\boldsymbol{\theta})$ given r and then maximizing ELBO$^c(q(\boldsymbol{\theta}; r)$ over r. The CAVI updates for ELBO^c are straightforward and involve only slight adjustments to the updates for the logistic case. All of the variational distributions maintain the same form.

Updating r is more challenging insofar as ELBO^c involves an intractable expectation over $q(\omega_i)$. Note, however, that this can be evaluated and optimized after profiling out the Polya-Gamma parameters, i.e. noting that $\tilde{b}_i$ and $\tilde{c}_i$ are functions of the other variational parameters and r and thus can be substituted out. The profiled ELBO (PELBO^c) is shown below:

$$
\begin{aligned}
&\text{PELBO}^c(q(\boldsymbol{\theta}, r)) = \sum_{i=1}^{N} \ln\Gamma(y_i + r) - \ln\Gamma(r) - \ln\Gamma(y_i + 1) - (y_i + r)\ln(2) + \\
&\frac{1}{2}(\boldsymbol{y} - r)^T[\boldsymbol{X}\tilde{\boldsymbol{\mu}}_\beta + \boldsymbol{Z}\tilde{\boldsymbol{\mu}}_\alpha - \ln r] + E_{q(\boldsymbol{\alpha})q\left(\{\boldsymbol{\Sigma}_j\}_{j=1}^J\right)}[\ln p(\boldsymbol{\alpha})] + \sum_{j=1}^{J} E_{q(\boldsymbol{\Sigma}_j)}\left[\ln p(\boldsymbol{\Sigma}_j)\right] + \ln p_0(r) \\
&\frac{1}{2}\ln\left[2\pi e|\tilde{\boldsymbol{\Lambda}}_{\alpha-\beta}|\right] + \sum_{i=1}^{N} -(y_i + r)\ln\cosh\left[\frac{1}{2}\sqrt{\begin{matrix}\left[\boldsymbol{x}_i^T\tilde{\boldsymbol{\mu}}_\beta + \boldsymbol{z}_i^T\tilde{\boldsymbol{\mu}}_\alpha - \ln r\right]^2 + \\ [\boldsymbol{x}_i^T, \boldsymbol{z}_i^T]\tilde{\boldsymbol{\Lambda}}_{\beta-\alpha}\begin{bmatrix}\boldsymbol{x}_i \\ \boldsymbol{z}_i\end{bmatrix}\end{matrix}}\right] + \\
&\sum_{j=1}^{J} E_{q(\boldsymbol{\Sigma}_j)}[-\ln q(\boldsymbol{\Sigma}_j)]
\end{aligned}
\tag{5}
$$

This objective monotonically increases at each iteration and thus can be monitored for convergence. The limitation of this approach is that it does not quantify uncertainty in r nor propagate it through to the other parameters.

Thus, I outline a second approximate variational strategy following the spirit of Wang and Blei (2013). First, consider the updates for $q(\boldsymbol{\theta})$: As per standard mean-field CAVI, we can examine the expectation of $f(\boldsymbol{y}, \boldsymbol{\theta}, r)$ over r. The only challenging term is $E_{q(r)}[\ln f(\omega_i|y_i + r, 0)]$. I rely on the delta-method and perform a second-order Taylor expansion around $E_{q(r)}[r]$.

$$
\begin{aligned}
E_{q(r)}[\ln f(\omega_i|y_i+r,0)] &\approx \ln f(\omega_i|y_i+E_{q(r)}[r]) + \frac{1}{2}\left[\frac{\partial}{\partial^2 r}\ln f(\omega_i|y_i+r)\ \right]_{r=E_{q(r)}[r]} Var_{q(r)}[r] \\
&\approx \ln f(\omega_i|y_i+E_{q(r)}[r]) + \frac{1}{2}\left[-\frac{9}{4}\psi^{(1)}\left(\frac{3}{2}[y_i+r]\right)\right]_{r=E_{q(r)}[r]} Var_{q(r)}[r]
\end{aligned}
\tag{6}
$$

As $\ln f(\omega_i|y_i+r)$ is itself intractable, I approximate it by matching the moments to a Gamma random variable; specifically, if $\omega_i \sim PG(b,0)$, then Gamma$(3/2\cdot b, 6)$ has the same mean and variance. Taking the second derivative of the approximation with respect to r gives the second line of Equation 6.

Given this approximation, the CAVI updates for $q(\boldsymbol{\theta})$ are very similar to the logistic case. Further, note that if $Var_{q(r)}[r]=0$, the variational EM updates are found.

For $q(\omega_i)$, $q(\omega_i) \sim PG\left(\tilde{b}_i, \tilde{c}_i\right)$ where

$$
\tilde{b}_i = y_i + E_{q(r)}[r]; \quad \tilde{c}_i = \sqrt{\begin{array}{c}\left[\boldsymbol{x}_i^T\tilde{\boldsymbol{\mu}}_\beta + \boldsymbol{z}_i^T\tilde{\boldsymbol{\mu}}_\alpha - E_{q(r)}[\ln r]\right]^2 + \\ [\boldsymbol{x}_i^T, \boldsymbol{z}_i^T]\tilde{\boldsymbol{\Lambda}}_{\beta-\alpha}\begin{bmatrix}\boldsymbol{x}_i \\ \boldsymbol{z}_i\end{bmatrix} + Var_{q(r)}[\ln r]\end{array}}
$$

For $q(\boldsymbol{\beta},\boldsymbol{\alpha})$, I explicitly state Scheme III (weak factorization). Updates for the other schemes (e.g. Schemes I and II) are similarly structured.

$$q(\boldsymbol{\beta}, \boldsymbol{\alpha}) \sim N\left(\begin{bmatrix} \tilde{\boldsymbol{\mu}}_\beta \\ \tilde{\boldsymbol{\mu}}_\alpha \end{bmatrix}, \quad \tilde{\boldsymbol{\Lambda}}_{\beta-\alpha}\right) \quad \tilde{\boldsymbol{\Lambda}}_{\beta-\alpha} = \left(\left[\boldsymbol{X}^T, \boldsymbol{Z}^T\right] \tilde{\boldsymbol{\Lambda}}_\Omega \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Z} \end{bmatrix} + \boldsymbol{T}^{-1}\right)^{-1}$$

$$\boldsymbol{T}^{-1} = \begin{pmatrix} \mathbf{0}_{p \times p} & \mathbf{0}_{p \times \sum_j d_j g_j} \\ \mathbf{0}_{\sum_j d_j g_j \times p} & \text{blockdiag}\left(\{\boldsymbol{I}_{g_j} \otimes E_{q(\boldsymbol{\Sigma}_j)}[\boldsymbol{\Sigma}_j^{-1}]\}_{j=1}^J\right) \end{pmatrix}$$

$$\begin{bmatrix} \tilde{\boldsymbol{\mu}}_\beta \\ \tilde{\boldsymbol{\mu}}_\alpha \end{bmatrix} = \tilde{\boldsymbol{\Lambda}}_{\beta-\alpha}[\boldsymbol{X}^T, \boldsymbol{Z}^T] \left(\boldsymbol{y}/2 - \mathbf{1}E_{q(r)}[r]/2 + \tilde{\boldsymbol{\Lambda}}_\Omega \mathbf{1} E_{q(r)}[\ln r]\right)$$

For $q(\boldsymbol{\Sigma}_j)$, these are identical to the logistic case.

Next, to update r, I first perform a change of variables to put the posterior in terms of $\ln r$. After doing so, I then assume that $\ln r \sim N(\tilde{\mu}_r, \tilde{\sigma}_r^2)$. With this assumption, the objective becomes as follows:

$$\tilde{\mu}_r, \tilde{\sigma}_r^2 = \text{argmax}_{\mu_r, \sigma_r^2} E_{q(\ln r)}\left[\text{ELBO}^c(q(\boldsymbol{\theta}), \exp(\ln r))\right] + \mu_r + \frac{1}{2}\ln\left(2\pi e \sigma_r^2\right) \tag{7}$$

Unfortunately, this is intractable as, indeed, $\text{ELBO}^c(q(\boldsymbol{\theta}), r)$ cannot be evaluated directly. To tackle this, consider a simple Laplace approximation (Wang and Blei 2013): This would approximate ELBO^c around its maximum with respect to r. I adopt a different strategy and maximize the *profiled* objective (PELBO^c) instead. With this optimum in hand, the Hessian of the objective in Equation 5 is then evaluated at $\hat{\ln r}$. As before, this remains intractable. To address this, I instead use the Hessian of the profiled objective. Thus, the approximate objective becomes:

$$\begin{aligned}\tilde{\mu}_r, \tilde{\sigma}^2_r = \text{argmax}_{\mu_r, \sigma^2_r} & \left[\text{PELBO}^c(q(\boldsymbol{\theta}), \exp(\hat{\ln r}))\right] + \\ & \frac{1}{2}\left[\frac{\partial}{\partial[\ln r]^2}\text{PELBO}^c(q(\boldsymbol{\theta}), \exp(\hat{\ln r}))\right]_{\ln r = \hat{\ln r}} \left([\mu_r - \hat{\ln r}]^2 + \sigma^2_r\right) + \\ & \mu_r + \frac{1}{2}\ln\left(2\pi e \sigma^2_r\right)\end{aligned} \tag{8}$$

Given $\hat{\ln r}$, there is a simple closed-form update to this problem, analogous to Wang and Blei (2013) and thus the variational updates can be derived.[1]

Assuming that $q(\ln r) \sim N(\tilde{\mu}_r, \tilde{\sigma^2}_r)$ and applying the approximations discussed in the main text, the variational updates are shown below:

$$\begin{aligned}\hat{\ln r} &= \text{argmax}_{\ln r}\text{PELBO}^c(q(\boldsymbol{\theta}), \exp(\ln r)) \\ \tilde{\mu}_r &= \tilde{\sigma}^2_r + \hat{\ln r}; \quad \tilde{\sigma}^2_r = \frac{1}{\left[\frac{\partial}{\partial[\ln r]^2}\text{PELBO}^c(q(\boldsymbol{\theta}), \exp(\ln r))\right]_{\ln r = \hat{\ln r}}}\end{aligned}$$

The second derivative of the profiled likelihood with respect to $\ln r$ is shown below where r is short-hand for $\exp(\ln r)$ and ψ_i is short-hand for $E_{q(\boldsymbol{\theta})}[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}]$ and σ^2_i

1. The software also implements the delta method (Wang and Blei 2013), although its performance is virtually identical in simulated experiments and is noticeably slower to fit and thus the Laplace approximation is set as the default.

for $Var_{q(\boldsymbol{\theta})}[\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}]$

$$\sum_{i=1}^{N}\left[\begin{array}{l} r\psi^{(0)}\left(y_i+r\right)+r^2\psi^{(1)}\left(y_i+r\right)-r\ln(2)-r\psi^{(0)}\left(r\right)-r^2\psi^{(1)}\left(r\right)+ \\ r+\frac{1}{2}\cdot r\cdot(\psi_i-\ln r)+ \\ -(y_i+r)\left(\begin{array}{l} -\frac{(\psi_i-\ln r)^2\tanh\left(\frac{1}{2}\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}\right)}{2\left((\psi_i-\ln r)^2+\sigma_i^2\right)^{3/2}}+ \\ \frac{\tanh\left(\frac{1}{2}\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}\right)}{2\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}}+ \\ \frac{(\psi_i-\ln r)^2\mathrm{sech}^2\left(\frac{1}{2}\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}\right)}{4\left((\psi_i-\ln r)^2+\sigma_i^2\right)} \end{array}\right)+ \\ \frac{r(\psi_i-\ln r)\tanh\left(\frac{1}{2}\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}\right)}{\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}}-r\log\left(\cosh\left(\frac{1}{2}\sqrt{(\psi_i-\ln r)^2+\sigma_i^2}\right)\right) \end{array}\right]$$

A downside of this approximate strategy is that the corresponding objective is no longer guaranteed to deterministically increase. Following Wang and Blei (2013), I monitor a surrogate objective although, as it may decrease, convergence should be assessed by looking at the stationarity of the variational parameters.

$$\begin{aligned}&\left[\mathrm{PELBO}^c(q(\boldsymbol{\theta}),\exp(\hat{\ln r}))\right]+\frac{1}{2}\left[\frac{\partial}{\partial[\ln r]^2}\mathrm{PELBO}^c(q(\boldsymbol{\theta}),\exp(\hat{\ln r}))\right]_{\ln r=\hat{\ln r}}\left([\tilde{\mu}_r-\hat{\ln r}]^2+\tilde{\sigma}_r^2\right)\\&E_{q(\boldsymbol{\alpha},\{\boldsymbol{\Sigma}_j\})}[\ln p(\boldsymbol{\alpha},\{\boldsymbol{\Sigma}_j\})]+\tilde{\mu}_r+\frac{1}{2}\ln\left(2\pi e\tilde{\sigma}_r^2\right)\end{aligned} \tag{9}$$

References

Goplerud, Max. 2020. "Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models Using Variational Inference." *Working Paper*, https://mgoplerud.com/papers/Goplerud_MAVB.pdf.

Pillow, Jonathan W., and James G. Scott. 2012. "Fully Bayesian Inference for Neural Models with Negative-Binomial Spiking." In *Neural Information Processing Systems 2012.* http://papers.nips.cc/paper/4567-fully-bayesian-inference-for-neural-models-with-negative-binomial-spiking.

Wang, Chong, and David M Blei. 2013. "Variational Inference in Nonconjugate Models." *Journal of Machine Learning Research* 14:1005–1031.

Zhou, Mingyuan, Lingbo Li, David Dunson, and Lawrence Carin. 2012. "Lognormal and Gamma Mixed Negative Binomial Regression." In *International Conference on Machine Learning.* https://icml.cc/2012/papers/665.pdf.