A note on fitting mean-plus-offset models, with applications to accelerating GLM-EIV and SCEPTRE

Tim B

In this note I consider the problem of efficiently fitting mean-plus-offset models (defined below) for Poisson and negative binomial distributions. The proposed methods could accelerate SCEPTRE and GLM-EIV by several orders of magnitude. I also describe a new initialization procedure for GLM-EIV based on these (and some other) ideas.

1 Mean-plus-offset model

For $i \in \{1, \ldots, n\}$, let Y_i be a random variable with exponential family distribution. Suppose the mean of Y_i is $g^{-1}(\beta + o_i)$, where $\beta \in \mathbb{R}$ is an unknown constant, $\{o_1, \ldots, o_n\}$ are known "offset" terms, and g^{-1} is an inverse link function. The offset terms can be fixed or random; if they are random, we perform estimation and inference conditional on their observed values. Our goal is to estimate β using MLE. We call this model a "mean-plus-offset" model.

1.1 Poisson case

Suppose the Y_is are Poisson-distributed with log link function, i.e.

$$Y_i \sim \text{Pois}(e^{\beta + o_i}).$$

The density f of Y_i is

$$f(y_i; \beta) = \frac{\left[e^{\beta+o_i}\right]^{(y_i)} e^{-\left[e^{\beta+o_i}\right]}}{y_i!} = \frac{e^{y_i\beta} e^{y_i o_i}}{e^{\left[e^{\beta+o_i}\right]} y_i!}.$$

The likelihood of the sample is

$$L(\beta; y) \quad = \quad \frac{e^{y_1\beta} e^{y_1 o_1}}{e^{\left[e^{\beta+o_1}\right]} y_1!} \cdots \frac{e^{y_n\beta} e^{y_n o_n}}{e^{\left[e^{\beta+o_n}\right]} y_n!} \quad = \quad \frac{\left(e^{\beta \sum_{i=1}^n y_i}\right) \left(e^{\sum_{i=1}^n y_i o_i}\right)}{e^{\sum_{i=1}^n e^{\beta+o_i}} \prod_{i=1}^n (y_i!)},$$

and the log-likelihood (up to a constant) is

$$\mathcal{L}(\beta; y) = \beta \sum_{i=1}^n y_i - \sum_{i=1}^n e^{\beta+o_i}.$$

Differentiating and setting equal to zero, we obtain the MLE equation

$$e^{\beta}\left(\sum_{i=1}^{n} e^{o_i}\right) = \sum_{i=1}^{n} y_i. \tag{1}$$

The MLE $\hat{\beta}^{(\text{Pois})}$ is therefore

$$\hat{\beta}^{(\text{Pois})} = \log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e^{o_i}}\right).$$

Suppose now that we have a weighted log-likelihood

$$\mathcal{L}(\beta; y) = \beta \sum_{i=1}^{n} T_i y_i - \sum_{i=1}^{n} T_i e^{\beta + o_i},$$

where $T_1, \ldots, T_n \in [0, 1]$ are weights. Repeating the above process, we find that the weighted MLE $\hat{\beta}^{(\text{Weighted Pois})}$ is

$$\hat{\beta}^{(\text{Weighted Pois})} = \log\left(\frac{\sum_{i=1}^{n} T_i y_i}{\sum_{i=1}^{n} T_i e^{o_i}}\right).$$

Thus, we can calculate $\hat{\beta}^{(\text{pois})}$ and $\hat{\beta}^{(\text{Weighted Pois})}$ analytically. We could instead estimate β by fitting a (possibly weighted) a GLM, but this is much ($\approx 500\times$) slower, as GLMs use an iterative fitting procedure.

1.2 Negative binomial case

Suppose the Y_is are negative binomially distributed with fixed size r and log link function, i.e.,

$$Y_i \sim \text{NB}_r(e^{\beta + o_i}).$$

The density f of Y_i is

$$\begin{aligned} f(y_i; \beta) &= \binom{y_i + r - 1}{y_i} \left(\frac{e^{\beta} e^{o_i}}{e^{\beta} e^{o_i} + r}\right)^{y_i} \left(\frac{r}{e^{\beta} e^{o_i} + r}\right)^{r} \\ &= \binom{y_i + r - 1}{y_i} \left(\frac{e^{y_i \beta} e^{y_i o_i}}{[e^{\beta} e^{o_i} + r]^{y_i}}\right) \left(\frac{r^r}{[e^{\beta} e^{o_i} + r]^r}\right). \end{aligned}$$

The likelihood across the samples is

$$L(\beta; y) = \prod_{i=1}^{n} \binom{y_i + r - 1}{y_i} \frac{e^{(\beta \sum y_i)} e^{(\sum y_i o_i)}}{\prod_i \left[(e^\beta e^{o_i} + r)^{y_i}\right]} \frac{r^{rn}}{\left[\prod_i e^\beta e^{o_i} + r\right]^r},$$

and the log-likelihood (up to a constant) is

$$\begin{aligned}\mathcal{L}(\beta; y) &= \beta \sum_i y_i - \sum_i y_i \log\left(e^\beta e^{o_i} + r\right) - r \sum_{i=1}^{n} \log(e^\beta e^{o_i} + r) \\ &= \beta \sum_i y_i - \sum_i (y_i + r) \log(e^\beta e^{o_i} + r).\end{aligned}$$

Differentiating in β and setting equal to zero, we obtain the MLE equation

$$e^\beta \sum_i \frac{(y_i + r)e^{o_i}}{e^\beta e^{o_i} + r} = \sum_i y_i. \tag{2}$$

1.2.1 Asymptotically exact solution

We cannot solve for β in (2) analytically. However, we can derive an asymptotically exact solution. Assume that o_i is a random variable. Then by the law of total expectation,

$$\mathbb{E}\left[\frac{(y_i + r)e^{o_i}}{e^{\beta+o_i} + r}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{(y_i + r)e^{o_i}}{e^{\beta+o_i} + r} | o_i\right]\right] = \mathbb{E}\left[\frac{(e^{\beta+o_i} + r)e^{o_i}}{e^{\beta+o_i} + r}\right] = \mathbb{E}[e^{o_i}],$$

because $\mathbb{E}[y_i] = e^{\beta+o_i}$ given fixed o_i. Dividing by n on both sides of (2), we have that

$$e^\beta \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i + r)e^{o_i}}{e^\beta e^{o_i} + r}\right) = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Finally, taking the limit in n and solving for β, we obtain

$$\hat{\beta}^{(\mathrm{NB})} \xrightarrow{P} \log\left(\frac{\mathbb{E}[y_i]}{\mathbb{E}[e^{o_i}]}\right). \tag{3}$$

But the Poisson MLE $\hat{\beta}^{(\mathrm{Pois})}$ converges to the same value (under random o_i):

$$\hat{\beta}^{(\mathrm{Pois})} = \log\left(\frac{(1/n)\sum_{i=1}^{n} y_i}{(1/n)\sum_{i=1}^{n} e^{o_i}}\right) \xrightarrow{P} \log\left(\frac{\mathbb{E}[y_i]}{\mathbb{E}[e^{o_i}]}\right).$$

Therefore, for large n, we can approximate $\hat{\beta}^{(\mathrm{NB})}$ by $\hat{\beta}^{(\mathrm{Pois})}$, which is fast to compute. The weighted case is similar.

1.2.2 Fisher information

We can compute the Fisher information of the negative binomial mean-plus-offset model. The second derivative of the log-likelihood is

$$\frac{d^2\mathcal{L}(\beta|y)}{d\beta^2} = -\sum_{i=1}^{n} \frac{re^{o_i+\beta}(r+y_i)}{(e^{o_i+\beta}+r)^2}.$$

Therefore, the Fisher information is

$$I_n(\beta) \;=\; -\mathbb{E}\left[\frac{d^2\mathcal{L}(\beta|y)}{d\beta^2}\right] \;=\; \sum_{i=1}^{n} \frac{re^{\beta+o_i}(r+e^{\beta+o_i})}{(e^{\beta+o_i}+r)^2} \;=\; \sum_{i=1}^{n} \frac{re^{\beta+o_i}}{r+e^{\beta+o_i}}.$$

Interestingly, the Fisher information for β depends on r, while the approximate MLE does not. The z-score for $\hat{\beta}^{(\mathrm{NB})}$, computable analytically, is

$$z = \hat{\beta}/\sqrt{1/I_n(\hat{\beta})}. \tag{4}$$

2 Application to SCEPTRE

For a given gene-gRNA pair, SCEPTRE fits $B = 500$ negative binomial mean-plus-offset models. Currently, SCEPTRE does this by fitting 500 univariate GLMs of expression onto thresholded perturbation (using distillations as offsets). As we consider developing faster versions of SCEPTRE, the z-score test statistic introduced here might be an attractive alternative to a simple linear test statistic.

3 Application to GLM-EIV, and a new initialization procedure

We can apply some of these ideas to accelerate GLM-EIV by way of a new initialization procedure. Consider an mRNA model with intercept β_0^m, perturbation coefficient β_{pert}^m, and vector technical factor coefficients $\beta_{\mathrm{tech}}^m \in \mathbb{R}^p$. Additionally, consider a gRNA model with analogous terms $\beta_0^g, \beta_{\mathrm{pert}}^g$, and β_{tech}^g. Let π be the marginal perturbation probability, and let n be the number of cells. Recall that the technical factors (e.g., batch, library size, etc.) are observed, while the perturbation indicator is unobserved.

A key observation is that, in practice, n is large ($\approx 200,000$) and π is small (< 0.01). Therefore, we can obtain good estimates of $\beta_0^m, \beta_{\text{tech}}^m, \beta_0^g$, and β_{tech}^g by regressing mRNA and gRNA counts onto the technical factors, even if the effect size of the unobserved perturbation ($\beta_{\text{pert}}^m, \beta_{\text{pert}}^g$) is large. These estimates serve as good pilot estimates for the EM algorithm. This observation motivates the following procedure:

1. Regress mRNA counts onto an intercept term and the technical factors to obtain estimates $\hat{\beta}_0^{m,\text{pilot}}$ and $\hat{\beta}_{\text{tech}}^{m,\text{pilot}}$ of β_o^m and β_{tech}^m. Do the same for the gRNA counts to obtain estimates $\hat{\beta}_0^{g,\text{pilot}}$ and $\hat{\beta}_{\text{tech}}^{g,\text{pilot}}$.

2. Extract the fitted values f^m, f^g from the mRNA model and gRNA model.

3. Run a reduced GLM-EIV on a simplified mRNA model consisting of offsets $\log(f^m)$ and a simplified gRNA model consisting of offsets $\log(f^g)$ (with no intercept or technical factor term in either model). Use $K \approx 15$ random restarts. Obtain pilot estimates $\hat{\beta}_{\text{pert}}^{m,\text{pilot}}, \hat{\beta}_{\text{pert}}^{g,\text{pilot}}, \hat{\pi}^{\text{pilot}}$.

4. Using the pilot estimates

$$\hat{\pi}^{\text{pilot}}, \hat{\beta}_0^{m,\text{pilot}}, \hat{\beta}_{\text{pert}}^{m,\text{pilot}}, \hat{\beta}_{\text{tech}}^{m,\text{pilot}}, \hat{\beta}_0^{g,\text{pilot}}, \hat{\beta}_{\text{pert}}^{g,\text{pilot}}, \hat{\beta}_{\text{tech}}^{g,\text{pilot}}$$

   as a starting location, run GLM-EIV on the full dataset once.

3.1 Speed

Relative to a naive strategy of repeated random parameter initialization, this algorithm is fast. Step 1 can be performed as a single precomputation (similar to SCEPTRE). Step 3 is a univariate EM algorithm. The E step only involves calculating membership probabilities, and the M step reduces to fitting mean-plus-offset models for both the mRNA and gRNA distributions, for which there exist analytic formulas. Finally, step 4 – running GLM-EIV on the full data – is fast because the pilot estimates are close to the global optimum.

Assuming 15 random restarts and 20 iterations per restart, a naive EM algorithm might require fitting $15 \times 20 \times 2 = 600$ GLMs per gRNA-gene pair. The above EM algorithm, by contrast, might involve fitting 10 GLMs per gene-gRNA pair (after pre-computations), making the algorithm feasible to apply at scale. (These numbers should be confirmed empirically.)

3.2 Obtaining good pilot estimates for intercepts and technical factors

In practice the perturbation probability π is expected to be less than 1%. Denoting the ith mRNA count by m_i and the ith technical factor vector by z_i, we can write the likelihood of mRNA regression model fitted in step 1 by

$$\sum_{i=1}^{n} l(m_i, z_i) = \sum_{i:p_i=1} l(m_i, z_i) + \sum_{i:p_i=0} l(m_i, z_i) \approx \sum_{i:p_i=0} l(m_i, z_i).$$

That is, the likelihood of the (technically incorrectly specified) GLM is approximately equal to the likelihood of a correctly specified GLM in which we condition on $p_i = 0$ before fitting the model. This observation (maybe under an additional assumption on the dependence of the covariates) suggests that we can obtain good estimates of β_o^m and β_{tech}^m without observing p_i.

When using a negative binomial model, we must estimate the dispersion parameter, which we will do in step 1. We could use reasoning similar to the above to argue that our dispersion estimates are good. It would be worthwhile to flesh these ideas out in more theoretical detail.

Method name	Method class	Robust to gene expression misspecification	Adjusts for confounders
Monocle regression	Parametric	No	Yes
Virtual FACS	Nonparametric	Yes	No
SCEPTRE	Conditional resampling	Yes	Yes