Estimation of Predicate Truth Probability

Feb 16, 2005

1 Notation

A	=	a true/false-valued predicate
N_{A}	=	number of times A is reached during one run of a program (complete data) (N for short)
M_{A}	=	number of times A is sampled during one run (sampled data) (M for short)
X_{A}	=	number of times A is true during one run (full data) (X for short)
Y_{A}	=	number of times A is sampled and is true during one run (sampled data) (Y for short)
α	:=	P(A is true at a single observation)
ρ	:=	sampling rate

2 The Problem

We concentrate on the problem of estimating $P(X > 0; y, m)$.

3 Solution Idea

First, note that $\alpha = P$(A is true at single observation) has an unbiased estimator:

$$\hat{\alpha} = \frac{\text{total number of times A is true (summed across all (failed) runs)}}{\text{total number of times A is observed (summed across all (failed) runs)}} \tag{1}$$

Given N, X and M each has a binomial distribution with parameters α and ρ, respectively. Y has a hypergeometric distribution given N, X, and M. The problem lies in putting a probability distribution on N, the number of times a predicate is reached during a run of the program. Since there is likely to be no proper conjugate prior distribution for N, we will have to model it using some other discrete distribution, say, a Poisson with a spike at zero, or a Poisson with a spike train, or a small multinomial. (These are the three different kinds of distributions of N that we observed in moss data.)

Figure 1 contains our proposed graphical model for this problem. Let's first look at the Poisson-plus-spike-at-zero prior for N.

$$P(n; \lambda, \gamma) = \gamma \frac{\lambda^n e^{-\lambda}}{n!} + (1 - \gamma) * \delta(n = 0) \tag{2}$$

$$P(x|n; \alpha) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x} \tag{3}$$

$$P(m|n; \rho) = \binom{n}{m} \rho^m (1 - \rho)^{n-m} \tag{4}$$

$$P(y|x, m, n) = \frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}} \tag{5}$$

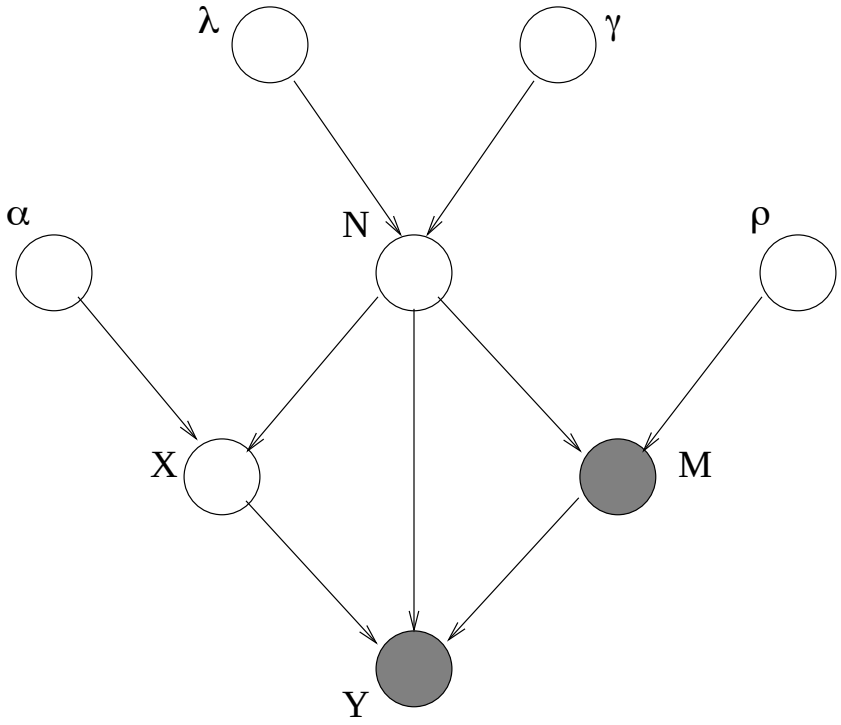

Figure 1: Graphical model for predicate truth probability estimation.

A note about notation, the variables after the semicolon are taken to be fixed parameters, whereas those that come before are taken to be random variables. For now we'll assume that λ and γ are fixed hyper-parameters. But keep in mind that in the Bayesian setting, they are taken to be random variables. Furthermore, following the empirical Bayes methodology, we'll set their estimate to be the mode of the posterior distributions $P(\lambda|m, y)$ and $P(\gamma|m, y)$.

It turns out that for this choice of $P(N)$, we can perform the necessary marginalization analytically and obtain a closed form solution for the posterior distributions. In the following derivations, we'll suppress the semicolon notation for simplicity. All parameters are taken to be known constants.

$$P(n, m, x, y) = P(n)\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\binom{n}{m}\rho^m(1-\rho)^{n-m}\frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}} \tag{6}$$

$$= P(n)\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\rho^m(1-\rho)^{n-m}\binom{x}{y}\binom{n-x}{m-y} \tag{7}$$

4 Derivation Details

Marginalizing out X, we get

$$
\begin{aligned}
P(n,m,y) &= \sum_x P(n,m,x,y) = P(n) \sum_{x=y}^{n-(m-y)} P(m,x,y|n) && (8)\\
&= P(n) \sum_{x=y}^{n-(m-y)} \frac{n!}{x!(n-x)!} \frac{x!}{y!(x-y)!} \frac{(n-x)!}{(m-y)!(n-x-(m-y))!} \alpha^x (1-\alpha)^{n-x} \rho^m (1-\rho)^{n-m} && (9)\\
&= P(n)\rho^m (1-\rho)^{n-m} \sum_{x=y}^{n-(m-y)} \frac{n!}{m!(n-m)!} \frac{m!}{y!(m-y)!} \frac{(n-m)!}{(x-y)!((n-m)-(x-y))!} \alpha^x (1-\alpha)^{n-x} && (10)\\
&= P(n)\binom{n}{m}\rho^m (1-\rho)^{n-m} \binom{m}{y}\alpha^y (1-\alpha)^{m-y} \sum_{x=y}^{n-(m-y)} \binom{n-m}{x-y} \alpha^{x-y} (1-\alpha)^{n-x-(m-y)} && (11)\\
&= P(n)\cdot\binom{n}{m}\rho^m (1-\rho)^{n-m} \cdot \binom{m}{y}\alpha^y (1-\alpha)^{m-y} \cdot 1 && (12)\\
&= P(n)P(m|n)P(y|m). && (13)
\end{aligned}
$$

The terms under the summation in Eqn. (11) form a binomial distribution of $(x-y)$ given $(n-m)$ and α (i.e., the distribution of the non-observed truth counts); it sums to one.

This tells us the intuitive result that, without regard to X (the number of times the predicate is true during the run), the number of times the predicate is observed to be true (Y) has a binomial distribution given the number of times it is observed (M), and the truth rate α. This derivation holds regardless of the form of the prior on N.

Marginalizing further, and this time plugging in the prior for N, we have

$$
\begin{aligned}
P(m,y) &= P(y|m) \sum_{n=m}^{\infty} P(n)P(m|n) && (14)\\
&= P(y|m) \sum_{n=m}^{\infty} \frac{n!}{m!(n-m)!} \rho^m (1-\rho)^{n-m} \cdot \left(\gamma \frac{\lambda^n}{n!} e^{-\lambda} + (1-\gamma)\delta(n=0)\right) && (15)\\
&= P(y|m) \left(\gamma \frac{\rho^m \lambda^m}{m!} e^{-\lambda\rho} \sum_{n=m}^{\infty} \frac{\lambda^{n-m}(1-\rho)^{n-m}}{(n-m)!} e^{-\lambda(1-\rho)} + (1-\gamma)\delta(m=0)\right) && (16)\\
&= P(y|m) \left(\gamma \frac{(\lambda\rho)^m}{m!} e^{-\lambda\rho} + (1-\gamma)\delta(m=0)\right) && (17)\\
&= P(y|m)P(m|\lambda\rho,\gamma). && (18)
\end{aligned}
$$

The terms under the summation is the Poisson distribution of $(n-m)$ with rate $\lambda(1-\rho)$, and hence sums to one. The summation over the spike at zero translates to a spike at zero for M. This tells the intuitive result that, marginalizing out a Poisson random variable that is stacked together with a binomial random variable, we get a Poisson random variable with a slower rate that is the product of the original Poisson rate and the binomial heads probability.

We now deal with the problem of estimating λ and γ given $\mathbf{m} := \{m_i\}$ and $\mathbf{y} := \{y_i\}$ from run i. Notice that the hyperparameters λ and γ are independent from Y given M. Hence we need to calculate $P(\lambda,\gamma|\mathbf{m})$.

We now need a prior distribution for λ and γ. γ is the mixture probability of the Poisson distribution and the spike at zero. For simplicity, we'll assume the non-informative uniform prior on the unit interval. λ is the location of the mean of the Poisson distribution. In our data, for most predicates the mean is concentrated near zero, but sometimes it does vary. We can assume a uniform prior on the interval $[a,b]$, where a and b may be determined by eyeballing the data.

Let $\mathcal{N} := \{i : m_i > 0\}$ and $\mathcal{Z} := \{i : m_i = 0\}$, we have

$$\begin{aligned} P(\mathbf{m}|\lambda,\gamma) &= \prod_i P(m_i|\lambda,\gamma) && (19)\\ &= \prod_i \left[\gamma\frac{(\lambda\rho)^{m_i}}{m_i!}e^{-\lambda\rho} + (1-\gamma)\delta(m_i=0)\right] && (20)\\ &= \prod_{i\in\mathcal{N}} \gamma\frac{(\lambda\rho)^{m_i}}{m_i!}e^{-\lambda\rho}\cdot(1-(1-e^{-\lambda\rho})\gamma)^{|\mathcal{Z}|} && (21)\\ &= \gamma^{|\mathcal{N}|}e^{-\lambda\rho|\mathcal{N}|}\frac{(\lambda\rho)^S}{\prod_i m_i!}(1-(1-e^{-\lambda\rho})\gamma)^{|\mathcal{Z}|}. && (22)\end{aligned}$$

where S in the the last equation denotes the sum of all (non-zero) m_i's.

To continue our quest,

$$\begin{aligned} P(\mathbf{m},\lambda,\gamma) &= P(\lambda;a,b)P(\gamma)P(\mathbf{m}|\lambda,\gamma) && (23)\\ P(\lambda;a,b) &= \frac{1}{b-a} && (24)\\ P(\gamma) &= 1 && (25)\\ P(\lambda,\gamma|\mathbf{m}) &= \frac{P(\mathbf{m},\lambda,\gamma)}{P(\mathbf{m})} && (26)\\ &= \frac{P(\mathbf{m}|\lambda,\gamma)/(b-a)}{\int_a^b\int_0^1 P(\mathbf{m},\lambda,\gamma)d\gamma d\lambda}. && (27)\end{aligned}$$

The integral in the denominator is difficult to compute. We can estimate it using Monte Carlo sampling techniques. But if we only want to set λ and γ to their MAP/ML estimates (the MAP estimate is equal to the ML estimate since λ and γ have uniform priors), then we don't even need the denominator. Simply look for the maximizer of the joint probability in the numerator.

$$(\hat{\lambda},\hat{\gamma}) = \text{argmax}_{\lambda,\gamma}\prod_{i\in\mathcal{N}}\gamma\frac{(\lambda\rho)^{m_i}}{m_i!}e^{-\lambda\rho}\cdot(1-(1-e^{-\lambda\rho})\gamma)^{|\mathcal{Z}|}, \qquad \lambda\in[a,b],\gamma\in[0,1] \tag{28}$$

See Appendix A for details.

5 Calculating $P(X > 0|m, y)$

After estimating $\hat{\lambda}$ and $\hat{\gamma}$, the marginal of X are easily computed, following exactly the same steps as for $P(m, n)$ above.

$$\begin{aligned} P(x;\hat{\lambda},\hat{\gamma}) &= \sum_{n=x}^{\infty} P(x|n)P(n;\hat{\lambda},\hat{\gamma}) && (29)\\ &= \hat{\gamma}\frac{(\hat{\lambda}\alpha)^x}{x!} + (1-\hat{\gamma})\delta(x=0). && (30)\end{aligned}$$

But what we really want is $P(X = 0|m, y; \hat{\lambda}, \hat{\gamma})$, from which we can get $P(X > 0|m, y) = 1 - P(X = 0|m, y)$. Thus we need to calculate the joint probability $P(X = 0, m, y)$, which we can then divide by $P(m, y)$ from Eqn. (17) to get the answer.

First, notice that if $y > 0$, then $P(X = 0, m, y) = 0$, hence $P(X > 0|m, y) = 1$. So we need only concern ourselves with the case where $y = 0$.

$$
\begin{aligned}
P(X = 0, m, y = 0) &= \sum_{n=m}^{\infty} P(n)P(m|n)P(x|n)P(y|x, m, n) && (31)\\
&= \sum_{n=m}^{\infty} P(n)P(m|n)\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}} && (32)\\
&= \sum_{n=m}^{\infty} P(n)P(m|n)(1-\alpha)^n \qquad \text{since } x = y = 0 && (33)\\
&= \sum_{n=m}^{\infty} \frac{n!}{m!(n-m)!}\rho^m(1-\rho)^{n-m}(1-\alpha)^n\left(\hat{\gamma}\frac{\hat{\lambda}^n}{n!}e^{-\hat{\lambda}} + (1-\hat{\gamma})\delta(n=0)\right) && (34)\\
&= \hat{\gamma}\frac{(\hat{\lambda}\rho(1-\alpha))^m}{m!}e^{-\hat{\lambda}}\sum_{n=m}^{\infty}\frac{(\hat{\lambda}(1-\rho)(1-\alpha))^{n-m}}{(n-m)!} + (1-\hat{\gamma})\delta(m=0) && (35)\\
&= \hat{\gamma}\frac{(\hat{\lambda}\rho(1-\alpha))^m}{m!}e^{-\hat{\lambda}(1-(1-\alpha)(1-\rho))} + (1-\hat{\gamma})\delta(m=0). && (36)
\end{aligned}
$$

From Eqn. (17), we obtain

$$
P(m, y = 0; \hat{\lambda}, \hat{\gamma}) = (1-\alpha)^m\hat{\gamma}\frac{(\hat{\lambda}\rho)^m}{m!}e^{-\hat{\lambda}\rho} + (1-\hat{\gamma})\delta(m=0). \quad (37)
$$

Combining Eqns. (36) and (37), we have

$$
\begin{aligned}
P(X = 0|m, y = 0; \hat{\lambda}, \hat{\gamma}) &= \frac{\hat{\gamma}\frac{(\hat{\lambda}\rho(1-\alpha))^m}{m!}e^{-\hat{\lambda}(1-(1-\alpha)(1-\rho))} + (1-\hat{\gamma})\delta(m=0)}{(1-\alpha)^m\hat{\gamma}\frac{(\hat{\lambda}\rho)^m}{m!}e^{-\hat{\lambda}\rho} + (1-\hat{\gamma})\delta(m=0)} && (38)\\
&= \begin{cases} \frac{\hat{\gamma}e^{-\hat{\lambda}\rho-\hat{\lambda}\alpha(1-\rho)}+(1-\hat{\gamma})}{\hat{\gamma}e^{-\hat{\lambda}\rho}+(1-\hat{\gamma})} & \text{if } m = 0 \\ e^{-\hat{\lambda}\alpha(1-\rho)} & \text{if } m > 0 \end{cases} && (39)
\end{aligned}
$$

6 Calculating $P(X > 0|n, m, y)$

Preliminary experiments show that the above model does not perform very well in practice. There are many cases of "over-estimate" of the posterior truth probability, i.e., $X = 0$ but $P(X > 0; m, y = 0) >> 0$. Our hypothesis is that the modeling of $P(X|N)$ is too simplistic. It is mostlikely the case that, in correspondence to underlying usage modes, X is a mixture of several different probabilities given N. To isolate the cause of the over estimate in the posterior truth probability, we replace our probabilistic model of N with its empirical distribution. In other words, we calculate $P(X|N, M, Y)$ and see how well that matches the real data.

From Eqn. (7), we have the full joint probability distribution

$$
P(n, x, m, y) = P(n)\text{Bin}(x|n, \alpha)\text{Bin}(m|n, \rho)\frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}}. \quad (40)
$$

From Eqn. (13), we get the marginal of N, M, and Y

$$
P(n, m, y) = P(n)\text{Bin}(m|n, \rho)\text{Bin}(y|m, \alpha). \quad (41)
$$

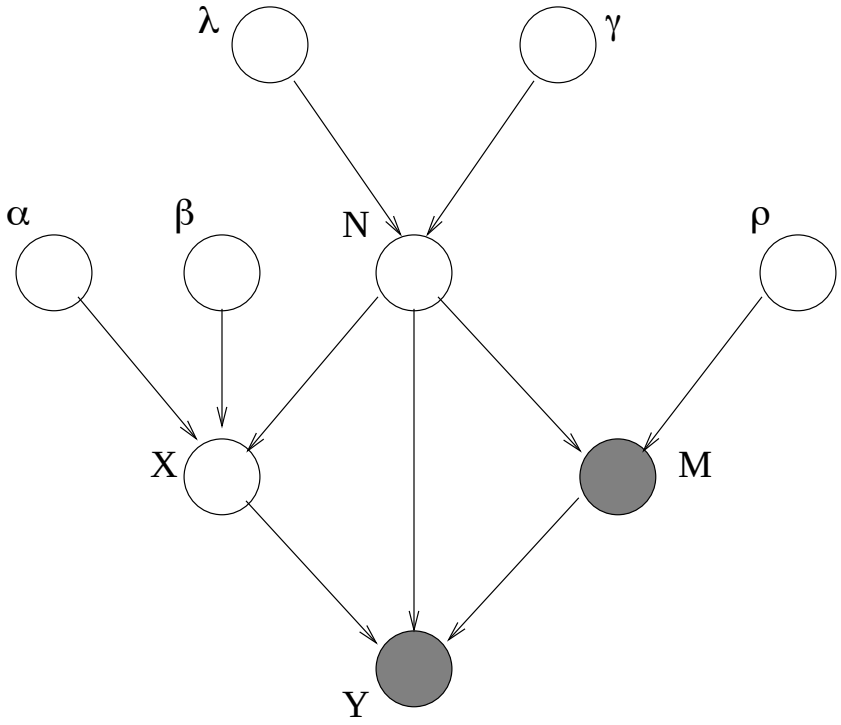

Figure 2: Revised model.

Combining the two, we get

$$\begin{aligned} P(x|n,m,y) &= \frac{\text{Bin}(x|n,\alpha)\binom{x}{y}\binom{n-x}{m-y}}{\text{Bin}(y|m,\alpha)\binom{n}{m}} && (42)\\ &= \binom{n-m}{x-y}\alpha^{x-y}(1-\alpha)^{n-x-m+y} && (43)\\ &= \text{Bin}(x-y|n-m,\alpha). && (44) \end{aligned}$$

As before, $P(X > 0|n, m, y > 0) = 1$, so we only need to worry about $P(X > 0|n, m, y = 0)$.

$$P(X > 0|n, m, y = 0) = (1 - \alpha)^{n-m} \quad (45)$$

7 Alternate Model for $P(X|N)$

In the previous section, we investigate the posterior truth probabilities given the ground truth N. The results are still unsatisfying. We observe the same over-estimation effect. Our main suspect is the binomial assumption of X given N. A plot of the full data histogram of X/N for various predicates of interest reveals spikes at 0 and 1, as well as approximately Gaussian bumps in between.

The Gaussian bump weakly supports the binomial model, but the spikes at zero and one reveals "sticky" modes where the predicate is always false or always true within a run. Thus we modify the model $P(X|N)$ to be a mixture

$$P(x|n, \alpha, \boldsymbol{\beta}) = \beta_1 \text{Bin}(x|n, \alpha) + \beta_2 \delta(x = 0) + \beta_3 \delta(x = n) \quad (46)$$

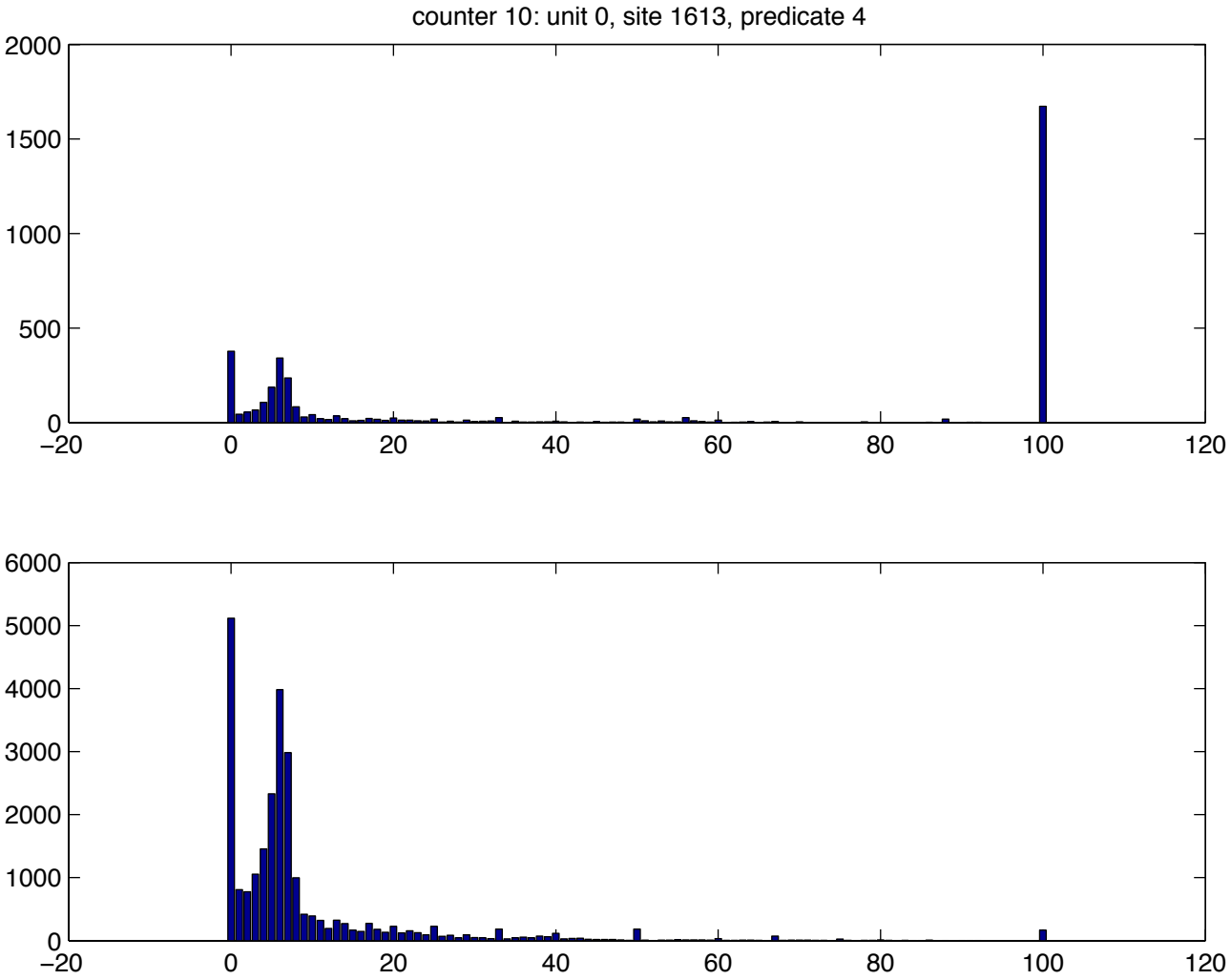

Figure 3: A histogram plot of X/N for predicate $(0, 1613, 4)$, the tenth most important predicate as ranked by lower bound of increase score (in "preds.txt"). Histograms like this is typical. The top plot is for failed runs, and the bottom for successful runs. The x-axis is on a scale of $X/N * 100$, so a marking of 100 really means 1.

Here are the marginal distributions under the new model.

$$\begin{aligned}
P(n,m,y|\rho,\alpha) &= \beta_1 P(n)\mathrm{Bin}(m|n,\rho)\mathrm{Bin}(y|m,\alpha) \\
&\quad +\beta_2 \sum_{x=y}^{n-(m-y)} P(n)\delta(x=0)\binom{n}{m}\rho^m(1-\rho)^{n-m}\frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}} \\
&\quad +\beta_3 \sum_{x=y}^{n-(m-y)} P(n)\delta(x=n)\binom{n}{m}\rho^m(1-\rho)^{n-m}\frac{\binom{x}{y}\binom{n-x}{m-y}}{\binom{n}{m}} \qquad (47)\\
&= \beta_1 P(n)\mathrm{Bin}(m|n,\rho)\mathrm{Bin}(y|m,\alpha) \\
&\quad +\beta_2 P(n)\delta(y=0)\mathrm{Bin}(m|n,\rho) \\
&\quad +\beta_3 P(n)\delta(y=m)\mathrm{Bin}(m|n,\rho) \qquad (48)\\
&= P(n)\mathrm{Bin}(m|n,\rho)\left[\beta_1\mathrm{Bin}(y|m,\alpha)+\beta_2\delta(y=0)+\beta_3\delta(y=m)\right] \qquad (49)\\
&=: P(n)\mathrm{Bin}(m|n,\rho)P(y|m,\alpha,\boldsymbol{\beta}). \qquad (50)
\end{aligned}$$

And, same as before, marginalizing out N induces a distribution on M that is a mixture of a Poisson and a delta function at zero (c.f. Eqn. (17)).

$$\begin{aligned}
P(m,y|\alpha,\boldsymbol{\beta},\lambda,\gamma) &= \sum_{n=m}^{\infty} P(n,m,y|\alpha,\boldsymbol{\beta},\lambda,\gamma,\rho) \qquad (51)\\
&= P(y|m,\alpha,\boldsymbol{\beta})\sum_{n=m}^{\infty} P(n|\lambda,\gamma)\mathrm{Bin}(m|n,\rho) \qquad (52)\\
&= P(y|m,\alpha,\boldsymbol{\beta})\left(\gamma\frac{(\lambda\rho)^m}{m!}e^{-\lambda\rho}+(1-\gamma)\delta(m=0)\right) \qquad (53)\\
&=: P(y|m,\alpha,\boldsymbol{\beta})\mathrm{PoiMix}(m|\lambda\rho,\gamma). \qquad (54)
\end{aligned}$$

It follows that the new hyperparameter posterior probability is proportional to

$$P(\alpha,\boldsymbol{\beta},\lambda,\gamma|\mathbf{m},\mathbf{y},\rho) \quad\propto\quad P(\alpha)P(\boldsymbol{\beta})P(\gamma)P(\lambda)P(\mathbf{y}|\mathbf{m},\alpha,\boldsymbol{\beta})P(\mathbf{m}|\lambda,\gamma,\rho) \qquad (55)$$

We use conjugate priors for the parameters. Thus, α and γ are beta-distributed, $\boldsymbol{\beta}$ has a Dirichlet prior distribution, and λ has a Gamma distribution.

$$\begin{aligned}
P(\alpha) &\sim \mathrm{Beta}(s,t) = \frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)}(1-\alpha)^{s-1}\alpha^{t-1} \qquad (56)\\
P(\boldsymbol{\beta}) &\sim \mathrm{Dir}(c_1,c_2,c_3) = \frac{\Gamma(\sum_i c_i)}{\prod_i \Gamma(c_i)}\beta_1^{c_1-1}\beta_2^{c_2-1}\beta_3^{c_3-1} \qquad (57)\\
P(\gamma) &\sim \mathrm{Beta}(j,k) = \frac{\Gamma(j+k)}{\Gamma(j)\Gamma(k)}(1-\gamma)^{j-1}\gamma^{k-1} \qquad (58)\\
P(\lambda) &\sim \mathrm{Gam}(u,v) = \frac{\lambda^{u-1}e^{-\lambda/v}}{\Gamma(u)v^u}. \qquad (59)
\end{aligned}$$

The hyperparameters can be set so that the mean and variance of the parameters roughly equal the empirical

means and variances.

$$
\begin{aligned}
t &= (\text{total \# truth counts}) + 1 && (60)\\
s &= (\text{total \# false counts}) + 1 = (\text{total \# obs}) - (\text{total \# true}) + 1 && (61)\\
c_1 &= (\text{\# runs } m > y > 0) + 1 && (62)\\
c_2 &= (\text{\# runs } m > y = 0) + 1 && (63)\\
c_3 &= (\text{\# runs } y = m > 0) + 1 && (64)\\
j &= (\text{\# runs } m = 0) + 1 && (65)\\
k &= (\text{\# runs } m > 0) + 1 && (66)\\
u &= \frac{(\text{average \# obs per run})^2}{(\text{variance of \# obs per run})} && (67)\\
v &= \frac{(\text{variance of \# obs per run})}{(\text{average \# obs per run})} && (68)
\end{aligned}
$$

The probability $P(\mathbf{m}|\gamma,\lambda,\rho)$ is as before (c.f. Eqn. (22)). The probability $P(\mathbf{y}|\mathbf{m},\alpha,\boldsymbol{\beta})$ can be written as follows:

$$
\begin{aligned}
P(\mathbf{y}|\mathbf{m},\alpha,\boldsymbol{\beta}) &= \prod_i \left[\beta_1 \mathrm{Bin}(y_i|m_i,\alpha) + \beta_2\delta(y_i=0) + \beta_3\delta(y_i=m_i)\right] && (69)\\
&= \prod_{i\in\mathcal{A}} \beta_1 \binom{m_i}{y_i} \alpha^{y_i}(1-\alpha)^{m_i-y_i} && (70)\\
&\quad \prod_{i\in\mathcal{B}} \left[\beta_1(1-\alpha)^{m_i} + \beta_2\right] && (71)\\
&\quad \prod_{i\in\mathcal{C}} \left[\beta_1\alpha^{m_i} + \beta_3\right], && (72)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{A} &:= \text{runs where } m > y > 0 && (73)\\
\mathcal{B} &:= \text{runs where } m > y = 0 && (74)\\
\mathcal{C} &:= \text{runs where } m = y > 0. && (75)
\end{aligned}
$$

Note that the sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are the same as the ones used in setting the hyperparameters c_1, c_2, and c_3.

Let's return to Eqn. (55). Since $(\alpha,\boldsymbol{\beta})$ and (γ,λ) are conditionally independent given M and Y, the posterior distribution of the hyperparameters can be factored into two parts.

$$
\begin{aligned}
P(\alpha,\boldsymbol{\beta},\lambda,\gamma|\mathbf{m},\mathbf{y},\rho) &\propto P(\alpha)P(\boldsymbol{\beta})P(\gamma)P(\lambda)P(\mathbf{y}|\mathbf{m},\alpha,\boldsymbol{\beta})P(\mathbf{m}|\lambda,\gamma,\rho) && (76)\\
&\propto P(\alpha,\boldsymbol{\beta},\mathbf{y}|\mathbf{m})P(\lambda,\gamma,\mathbf{m}|\rho) && (77)\\
&= P\cdot Q && (78)\\
\text{where } P &:= P(\alpha)P(\boldsymbol{\beta})P(\mathbf{y}|\mathbf{m},\alpha,\boldsymbol{\beta}) && (79)\\
Q &:= P(\lambda)P(\gamma)P(\mathbf{m}|\gamma,\lambda,\rho) && (80)
\end{aligned}
$$

As before, we employ the Empirical Bayes paradigm and replace the hyperparameters with their MAP estimates. We obtain the estimates by maximizing the Bayesian likelihood using Newton-Raphson. The fact that the likelihood is factorizable means that we can estimate $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ separately from $\hat{\gamma}$ and $\hat{\lambda}$. See Appendix B for details.

8 Calculating $P(X > 0|m, y)$ and $P(X > 0|n, m, y)$ in Model 2

As usual, what we really need are the posterior probabilities $P(X > 0|m, y)$ and $P(X > 0|n, m, y)$. Of course, when $y > 0$, the probabilities are 1. Hence we're only interested in the case when $y = 0$. For cleanliness of notation, we drop the $\hat{}$ notation for the MAP estimates of parameters. Just keep in mind that all hyperparameters $(\alpha, \boldsymbol{\beta}, \gamma, \lambda)$ are now fixed to be their MAP values.

Let's start with $P(X > 0|n, m, y)$ first.

$$\begin{aligned}
P(n, m, y = 0, x = 0) &= P(n)P(x = 0|n)P(m|n)P(y = 0|x = 0, m, n) && (81)\\
&= P(n)P(m|n)\left[\beta_1 \text{Bin}(x = 0|n, \alpha) + \beta_2 + \beta_3\delta(x = n)\right] && (82)\\
&= P(n)\text{Bin}(m|n)[\beta_1(1 - \alpha)^n + \beta_2 + \beta_3\delta(n = 0)], && (83)\\
P(x = 0|n, m, y = 0) &= \frac{P(n, m, y = 0, x = 0)}{P(n, m, y = 0)} && (84)\\
&= \frac{\beta_1(1 - \alpha)^n + \beta_2 + \beta_3\delta(n = 0)}{\beta_1 \text{Bin}(y|m, \alpha) + \beta_2\delta(y = 0) + \beta_3\delta(y = m)} && (85)\\
&= \frac{\beta_1(1 - \alpha)^n + \beta_2 + \beta_3\delta(n = 0)}{\beta_1(1 - \alpha)^m + \beta_2 + \beta_3\delta(m = 0)} && (86)\\
&= \begin{cases} 1, & \text{if } m = n = 0 \\ \beta_1(1 - \alpha)^n + \beta_2, & \text{if } m = 0,\ n > 0 \\ \frac{\beta_1(1-\alpha)^n + \beta_2}{\beta_1(1-\alpha)^m + \beta_2}, & \text{if } m > 0,\ n > 0 \end{cases}. && (87)
\end{aligned}$$

In Eqn. (84) above, we made use of the formula for $P(n, m, y)$ from Eqn. (49) in Section 7.

Next, we calculate $P(X > 0|m, y = 0)$.

$$\begin{aligned}
P(x = 0, m, y = 0) &= \sum_{n=m}^{\infty} P(n)P(m|n)P(x|n)P(y|m, n, x) && (88)\\
&= \sum_{n=m}^{\infty} P(n)P(m|n)[\beta_1 \text{Bin}(x|n, \alpha) + \beta_2\delta(x = 0) + \beta_3\delta(x = n)] \cdot 1 && (89)\\
&= \sum_{n=m}^{\infty} P(n)P(m|n)[\beta_1(1 - \alpha)^n + \beta_2 + \beta_3\delta(n = 0)] && (90)\\
&= \beta_1 P_1(X = 0, m, y = 0) + \beta_2 \text{PoiMix}(m|\lambda\rho, \gamma) + \beta_2\delta(m = 0)P(n = 0), && (91)
\end{aligned}$$

where $P_1(X = 0, m, y = 0)$ is the probability $P(X = 0, m, y = 0)$ under Model 1 (c.f. Eqn. (36)).

From Eqn. (54), we have

$$\begin{aligned}
P(m, y = 0) &= P(y|m, \alpha, \boldsymbol{\beta})\text{PoiMix}(m|\lambda\rho, \gamma) && (92)\\
&= [\beta_1(1 - \alpha)^m + \beta_2 + \beta_3\delta(m = 0)] \cdot [\gamma\frac{(\lambda\rho)^m}{m!}e^{-\lambda\rho} + (1 - \gamma)\delta(m = 0)]. && (93)
\end{aligned}$$

Hence, combining Eqns. (36), (37), (91), and (93), we get

$$P(X = 0|m, y = 0) = \begin{cases} \frac{\beta_1(1-\alpha)^m e^{-\lambda\alpha(1-\rho)} + \beta_2}{\beta_1(1-\alpha)^m + \beta_2}, & \text{if } m > 0, \\ \frac{(\beta_1 e^{-\lambda\alpha(1-\rho)} + \beta_2 + \beta_3 e^{-\lambda(1-\rho)})\gamma e^{-\lambda\rho} + (1-\gamma)}{\gamma e^{-\lambda\rho} + (1-\gamma)}, & \text{if } m = 0. \end{cases} \quad (94)$$

Appendix A: Hyperparameter Estimates in Model 1

We estimate $\hat{\lambda}_{MAP}$ and $\hat{\gamma}_{MAP}$ using Newton's method. Let

$$\begin{aligned}
A &:= (1 + (e^{-\lambda\rho} - 1)\gamma) && (95)\\
L &:= \log P(\mathbf{m}|\lambda, \gamma) = |\mathcal{N}| \log\gamma - \lambda\rho|\mathcal{N}| + S\log\lambda + |\mathcal{Z}|\log A + const. && (96)
\end{aligned}$$

Then, (c.f. Eqn. (28))

$$(\hat{\lambda}, \hat{\gamma}) = \text{argmax}_{\lambda,\gamma} L, \qquad \lambda \in [a, b], \gamma \in [0, 1].$$

Let

$$B := \frac{\partial A}{\partial \lambda} = -\gamma\rho e^{-\lambda\rho} \tag{97}$$

$$C := \frac{\partial A}{\partial \gamma} = e^{-\lambda\rho} - 1 \tag{98}$$

$$D := \frac{\partial^2 A}{\partial \lambda^2} = \gamma\rho^2 e^{-\lambda\rho} \tag{99}$$

$$E := \frac{\partial^2 A}{\partial \lambda \partial \gamma} = -\rho e^{-\lambda\rho} \tag{100}$$

$$F := \frac{\partial^2 A}{\partial \gamma^2} = 0. \tag{101}$$

We have

$$\frac{\partial L}{\partial \lambda} = -\rho|\mathcal{N}| + \frac{S}{\lambda} + \frac{|\mathcal{Z}|}{A} \cdot B \tag{102}$$

$$\frac{\partial L}{\partial \gamma} = \frac{|\mathcal{N}|}{\gamma} + \frac{|\mathcal{Z}|}{A} \cdot C \tag{103}$$

$$\frac{\partial^2 L}{\partial \lambda^2} = -\frac{S}{\lambda^2} - \frac{|\mathcal{Z}|}{A^2} B^2 + \frac{|\mathcal{Z}|}{A} D \tag{104}$$

$$\frac{\partial^2 L}{\partial \gamma^2} = -\frac{|\mathcal{N}|}{\gamma^2} - \frac{|\mathcal{Z}|}{A^2} C^2 \tag{105}$$

$$\frac{\partial^2 L}{\partial \lambda \partial \gamma} = -\frac{|\mathcal{Z}|}{A^2} BC + \frac{|\mathcal{Z}|}{A} E \tag{106}$$

The Newton-Raphson update equations are

$$\begin{bmatrix} \lambda_{n+1} \\ \gamma_{n+1} \end{bmatrix} = \begin{bmatrix} \lambda_n \\ \gamma_n \end{bmatrix} - (\nabla^2 L)^{-1} \cdot \nabla L, \tag{107}$$

$$\text{where} \qquad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial \lambda} \\ \frac{\partial L}{\partial \gamma} \end{bmatrix}, \quad \nabla^2 L = \begin{bmatrix} \frac{\partial^2 L}{\partial \lambda^2} & \frac{\partial^2 L}{\partial \lambda \partial \gamma} \\ \frac{\partial^2 L}{\partial \lambda \partial \gamma} & \frac{\partial^2 L}{\partial \gamma^2} \end{bmatrix}. \tag{108}$$

Also recall the matrix inverse formula for 2x2 matrices:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad A^{-1} = \frac{1}{ad - cb} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Appendix B: Hyperparameter Estimates in Model 2

In this section, we use Newton-Raphson to obtain the MAP estimate of the hyperparameters in model 2. This requires calculating the gradient and the Hessian. As explained in section 7, the posterior probability of the hyperparameters given observations $\mathbf{m}$ and $\mathbf{y}$ is factorable, which simplies the calculation of the cross-terms in the Hessian.

Continuing from Eqn. (78), we have

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\lambda}) = \text{argmax}_{\alpha,\boldsymbol{\beta},\gamma,\lambda} \log P(\alpha, \boldsymbol{\beta}, \gamma, \lambda, \mathbf{m}, \mathbf{y}|\rho) \tag{109}$$

$$= \text{argmax}_{\alpha,\boldsymbol{\beta},\gamma,\lambda} \log P + \log Q \tag{110}$$

$$= (\text{argmax}_{\alpha,\boldsymbol{\beta}} \log P, \text{argmax}_{\gamma,\lambda} \log Q), \tag{111}$$

where

$$\begin{aligned}
\log P(\mathbf{y},\alpha,\boldsymbol{\beta}|\mathbf{m}) &= \log\frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)}+\log\frac{\Gamma(\sum_i c_i)}{\prod_i\Gamma(c_i)} && (112)\\
&\quad +(s-1)\log(1-\alpha)+(t-1)\log\alpha && (113)\\
&\quad +(c_1-1)\log\beta_1+(c_2-1)\log\beta_2+(c_3-1)\log\beta_3 && (114)\\
&\quad +\sum_{i\in\mathcal{A}}\left[\log\beta_1+\log\binom{m_i}{y_i}+y_i\log\alpha+(m_i-y_i)\log(1-\alpha)\right] && (115)\\
&\quad +\sum_{i\in\mathcal{B}}\log[\beta_1(1-\alpha)^{m_i}+\beta_2] && (116)\\
&\quad +\sum_{i\in\mathcal{C}}\log[\beta_1\alpha^{m_i}+\beta_3], && (117)
\end{aligned}$$

and

$$\begin{aligned}
\log Q &= \log\frac{\lambda^{u-1}e^{-\lambda/v}}{\Gamma(u)v^u}+\log\frac{\Gamma(j+k)}{\Gamma(j)\Gamma(k)}(1-\gamma)^{j-1}\gamma^{k-1} && (118)\\
&\quad +\log\gamma^{|\mathcal{N}|}e^{-\lambda\rho|\mathcal{N}|}\frac{(\lambda\rho)^S}{\prod_i m_i!}(1-(1-e^{-\lambda\rho})\gamma)^{|\mathcal{Z}|} && (119)\\
&= (u-1)\log\lambda-\frac{\lambda}{v}+(j-1)\log(1-\gamma)+(k-1)\log\gamma && (120)\\
&\quad +|\mathcal{N}|\log\gamma-\lambda\rho|\mathcal{N}|+S\log\lambda+|\mathcal{Z}|\log(1-(1-e^{-\lambda\rho})\gamma) && (121)\\
&\quad +const. && (122)
\end{aligned}$$

Let's first optimize $\log P$ for $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$. Let $B_i := \beta_1(1-\alpha)^{m_i}+\beta_2$ and $C_i := \beta_1\alpha^{m_i}+\beta_3$. In addition, we use the softmax representation for the β's:

$$\begin{aligned}
\beta_i &= \frac{e^{t_i}}{T}, && (123)\\
\text{where}\qquad T &= \sum_j e^{t_j}. && (124)
\end{aligned}$$

We first calculate the partial derivatives of β_i, B_i, and C_i.

$$\begin{aligned}
d_{ij} := \frac{\partial\beta_i}{\partial t_j} &= \frac{Te^{t_i}\delta(i=j)-e^{t_i}e^{t_j}}{T^2} && (125)\\
D_{ijk} := \frac{\partial^2\beta_i}{\partial t_j\partial t_k} &= \frac{T^2e^{t_i}\delta(i=j=k)-Te^{t_i}(e^{t_j}\delta(i=k)+e^{t_j}\delta(j=k)+e^{t_k}\delta(i=j))+2e^{t_i}e^{t_j}e^{t_k}}{T^3} && (126)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial B_i}{\partial t_j} &= d_{1j}(1-\alpha)^{m_i}+d_{2j} && (127)\\
\frac{\partial B_i}{\partial\alpha} &= -\beta_1 m_i(1-\alpha)^{m_i-1} && (128)\\
\frac{\partial^2 B_i}{\partial t_j\partial t_k} &= D_{1jk}(1-\alpha)^{m_i}+D_{2jk} && (129)\\
\frac{\partial^2 B_i}{\partial t_j\partial\alpha} &= -d_{1j}m_i(1-\alpha)^{m_i-1} && (130)\\
\frac{\partial^2 B_i}{\partial\alpha^2} &= \beta_1 m_i(m_i-1)(1-\alpha)^{m_i-2} && (131)
\end{aligned}$$

$$\frac{\partial C_i}{\partial t_j} = d_{1j}\alpha^{m_i} + d_{3j} \tag{132}$$

$$\frac{\partial C_i}{\partial \alpha} = \beta_1 m_i \alpha^{m_i - 1} \tag{133}$$

$$\frac{\partial^2 C_i}{\partial t_j \partial t_k} = D_{1jk}\alpha^{m_i} + D_{3jk} \tag{134}$$

$$\frac{\partial^2 C_i}{\partial t_j \partial \alpha} = d_{1j} m_i \alpha^{m_i - 1} \tag{135}$$

$$\frac{\partial^2 C_i}{\partial \alpha^2} = \beta_1 m_i (m_i - 1)\alpha^{m_i - 2} \tag{136}$$

Thus, the first-order derivatives are

$$\frac{\partial \log P}{\partial \alpha} = -\frac{s-1}{1-\alpha} + \frac{t-1}{\alpha} + \sum_{i \in \mathcal{A}} \left[\frac{y_i}{\alpha} - \frac{m_i - y_i}{1-\alpha} \right] + \sum_{i \in \mathcal{B}} B_i^{-1} \frac{\partial B_i}{\partial \alpha} + \sum_{i \in \mathcal{C}} C_i^{-1} \frac{\partial C_i}{\partial \alpha}, \tag{137}$$

$$\frac{\partial \log P}{\partial t_j} = \sum_{i=1,2,3} \frac{c_i - 1}{\beta_i} d_{ij} + \sum_{i \in \mathcal{A}} \beta_1^{-1} d_{1j} + \sum_{i \in \mathcal{B}} B_i^{-1} \frac{\partial B_i}{\partial t_j} + \sum_{i \in \mathcal{C}} C_i^{-1} \frac{\partial C_i}{\partial t_j}. \tag{138}$$

The second-order derivatives are

$$\begin{aligned} \frac{\partial^2 \log P}{\partial \alpha^2} = & -\frac{s-1}{(1-\alpha)^2} - \frac{t-1}{\alpha^2} + \sum_{i \in \mathcal{A}} \left[-\frac{y_i}{\alpha^2} - \frac{m_i - y_i}{(1-\alpha)^2} \right] \\ & + \sum_{i \in \mathcal{B}} -B_i^{-2} \left(\frac{\partial B_i}{\partial \alpha} \right)^2 + B_i^{-1} \frac{\partial^2 B_i}{\partial \alpha^2} \\ & + \sum_{i \in \mathcal{C}} -C_i^{-2} \left(\frac{\partial C_i}{\partial \alpha} \right)^2 + C_i^{-1} \frac{\partial^2 C_i}{\partial \alpha^2} \end{aligned} \tag{139}$$

$$\begin{aligned} \frac{\partial^2 \log P}{\partial t_j \partial t_k} = & \sum_{i=1,2,3} -\frac{c_i - 1}{\beta_i^2} d_{ik} d_{ij} + \frac{c_i - 1}{\beta_i} D_{ijk} \\ & + \sum_{i \in \mathcal{A}} -\beta_1^{-2} d_{1k} d_{1j} + \beta_1^{-1} D_{1jk} \\ & + \sum_{i \in \mathcal{B}} -B_i^{-2} \frac{\partial B_i}{\partial t_k} \frac{\partial B_i}{\partial t_j} + B_i^{-1} \frac{\partial^2 B_i}{\partial t_k \partial t_j} \\ & + \sum_{i \in \mathcal{C}} -C_i^{-2} \frac{\partial C_i}{\partial t_k} \frac{\partial C_i}{\partial t_j} + C_i^{-1} \frac{\partial^2 C_i}{\partial t_k \partial t_j} \end{aligned} \tag{140}$$

$$\begin{aligned} \frac{\partial^2 \log P}{\partial t_j \partial \alpha} = & \sum_{i \in \mathcal{B}} -B_i^{-2} \frac{\partial B_i}{\partial t_j} \frac{\partial B_i}{\partial \alpha} + B_i^{-1} \frac{\partial^2 B_i}{\partial t_j \partial \alpha} \\ & + \sum_{i \in \mathcal{C}} -C_i^{-2} \frac{\partial C_i}{\partial t_j} \frac{\partial C_i}{\partial \alpha} + C_i^{-1} \frac{\partial^2 C_i}{\partial t_j \partial \alpha}. \end{aligned} \tag{141}$$

Now let's turn to $\log Q$ for estimating $\hat{\lambda}$ and $\hat{\gamma}$. The derivatives would be exactly the same as those derived in Appendix A, except for the fact that we're now using conjugate priors with specific values for

hyperparameters. Nevertheless, the new derivatives are very similar to the old ones given in Eqns. (102-106).

$$\frac{\partial \log Q}{\partial \lambda} = \frac{u-1}{\lambda} - \frac{1}{v} + \frac{\partial L}{\partial \lambda} \tag{142}$$

$$\frac{\partial \log Q}{\partial \gamma} = -\frac{j-1}{1-\gamma} + \frac{k-1}{\gamma} + \frac{\partial L}{\partial \gamma} \tag{143}$$

$$\frac{\partial^2 \log Q}{\partial \lambda^2} = -\frac{u-1}{\lambda^2} + \frac{\partial^2 L}{\partial \lambda^2} \tag{144}$$

$$\frac{\partial^2 \log Q}{\partial \gamma^2} = -\frac{j-1}{(1-\gamma)^2} - \frac{k-1}{\gamma^2} + \frac{\partial^2 L}{\partial \gamma^2} \tag{145}$$

$$\frac{\partial^2 \log Q}{\partial \lambda \partial \gamma} = \frac{\partial^2 L}{\partial \lambda \partial \gamma} \tag{146}$$