# STA 6351, Report.1.7

**Carson Slater** *Baylor University*

**1.7**

The Poisson–gated normal model we considered in Report 1.2 provides a convenient laboratory for iterative optimization algorithms. Because the likelihood separates into a discrete Poisson part and a continuous normal part, we can illustrate Newton–Raphson (NR) and Fisher scoring (FS) in a setting that is both realistic and analytically tractable.

Recall that for each unit $i = 1, \dots, n$ the observed data consist of:

- a Poisson count $N_i \sim \mathrm{Pois}(\mu)$, and
- a Normal outcome $X_i \mid N_i \sim N(\alpha N_i, \sigma^2)$ observed only when $N_i > 0$.

Define the indicator

$$\Delta_i = \mathbf{1}\{N_i > 0\},$$

and let $m = \sum_{i=1}^{n} \Delta_i$ be the number of units for which $X_i$ is observed. For those $i$ with $\Delta_i = 1$ we also observe the realized Poisson count $k_i$ (i.e. $N_i = k_i$ for observed units).

We treat the parameters as $\theta = (\mu, \alpha, \sigma^2)$ where $\mu > 0$, $\sigma^2 > 0$.

For a single unit the contribution to the likelihood depends on whether $\Delta_i = 0$ or $1$.

- If $\Delta_i = 0$ (no $X_i$ observed) the contribution is $P(N_i = 0) = e^{-\mu}$.
- If $\Delta_i = 1$ (we observe $N_i = k_i$ and $X_i = x_i$) the joint contribution is

$$P(N_i = k_i)\, f_{X|N}(x_i \mid k_i) = e^{-\mu} \frac{\mu^{k_i}}{k_i!} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{(x_i - \alpha k_i)^2}{2\sigma^2} \Big\}.$$

As shown in Report 1.2, the likelihood combines contributions from all $Y_i$'s and the observed $X_i$'s:

$$L(\mu, \sigma^2 \mid \mathcal{D}) = \prod_{i=1}^{n} \Pr(Y_i = y_i;\, \mu) \;\times \prod_{i:\, \Delta_i = 1} \frac{1}{\sigma}\, \varphi\Big(\frac{x_i - \mu}{\sigma}\Big).$$

**(a) Write the log-likelihood, up to constants, as**

$$\ell(\mu, \alpha, \sigma^2) = -n\mu + \Big(\sum_{i:\Delta_i=1} k_i\Big)\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha) + \text{const.}$$

**For brevity denote $K = \sum_{i:\Delta_i=1} k_i$.**

The likelihood $L(\mu, \alpha, \sigma^2 \mid \mathcal{D})$ is the product of contributions from the $(n-m)$ units with $\Delta_i = 0$ (where $N_i = 0$) and the $m$ units with $\Delta_i = 1$ (where $N_i = k_i$ and $X_i = x_i$ are observed).

$$\begin{aligned}
L(\mu, \alpha, \sigma^2 \mid \mathcal{D}) &= \prod_{i:\Delta_i=0} \underbrace{P(N_i = 0)}_{e^{-\mu}} \times \prod_{i:\Delta_i=1} \underbrace{P(N_i = k_i)\, f_{X|N}(x_i \mid k_i)}_{e^{-\mu}\frac{\mu^{k_i}}{k_i!}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_i-\alpha k_i)^2}{2\sigma^2}\right\}} \\
&= (e^{-\mu})^{n-m} \cdot \prod_{i:\Delta_i=1} \left[e^{-\mu}\frac{\mu^{k_i}}{k_i!} \cdot \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big\{-\frac{(x_i-\alpha k_i)^2}{2\sigma^2}\Big\}\right] \\
&= e^{-n\mu} \cdot \mu^{\sum_{i:\Delta_i=1} k_i} \cdot \left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right)^m \cdot \left(\prod_{i:\Delta_i=1}\frac{1}{k_i!}\right) \cdot \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i:\Delta_i=1}(x_i-\alpha k_i)^2\Big\}
\end{aligned}$$

Taking the natural logarithm of $L$ and using the definitions $K = \sum_{i:\Delta_i=1} k_i$ and $S(\alpha) = \sum_{i:\Delta_i=1}(x_i - \alpha k_i)^2$:

$$\begin{aligned}
\ell(\mu, \alpha, \sigma^2) &= \log L(\mu, \alpha, \sigma^2 \mid \mathcal{D}) \\
&= -n\mu + K\log\mu - \frac{m}{2}\log(2\pi) - \frac{m}{2}\log\sigma^2 + \log\left(\prod_{i:\Delta_i=1}\frac{1}{k_i!}\right) - \frac{1}{2\sigma^2}S(\alpha)
\end{aligned}$$

All terms that do not depend on the parameters $\theta = (\mu, \alpha, \sigma^2)$ are collected into a constant:

$$\text{const} = -\frac{m}{2}\log(2\pi) + \log\left(\prod_{i:\Delta_i=1}\frac{1}{k_i!}\right)$$

The log-likelihood, up to constants, is:

$$\ell(\mu, \alpha, \sigma^2) = -n\mu + \Big(\sum_{i:\Delta_i=1} k_i\Big)\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha) + \text{const.}$$

**(b) Score functions (first derivatives)**

**Differentiate $\ell$ with respect to each parameter.**

- **Score for $\mu$:**

$$U_\mu(\mu, \alpha, \sigma^2) = \frac{\partial \ell}{\partial \mu} = -n + \frac{K}{\mu}.$$

- **Score for $\alpha$:**

$$U_\alpha(\mu, \alpha, \sigma^2) = \frac{\partial \ell}{\partial \alpha} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \alpha}S(\alpha) = \frac{1}{\sigma^2}\sum_{i:\Delta_i=1} k_i(x_i - \alpha k_i).$$

- **Score for $\sigma^2$:**

$$U_{\sigma^2}(\mu, \alpha, \sigma^2) = \frac{\partial \ell}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}S(\alpha).$$

**Set each score to zero to obtain the likelihood equations.**

The log-likelihood, up to constants, for the Poisson–gated normal model is given by:

$$\ell(\mu, \alpha, \sigma^2) = -n\mu + K\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha) + \text{const}$$

where $K = \sum_{i:\Delta_i=1} k_i$ and $S(\alpha) = \sum_{i:\Delta_i=1}(x_i - \alpha k_i)^2$.

The score functions are the first partial derivatives of the log-likelihood with respect to each parameter $\theta = (\mu, \alpha, \sigma^2)$.

*Score for $\mu$*

The score function $U_\mu$ is obtained by differentiating $\ell$ with respect to $\mu$:

$$U_\mu(\mu, \alpha, \sigma^2) = \frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-n\mu + K\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha)\right)$$

$$U_\mu(\mu, \alpha, \sigma^2) = -n(1) + K\left(\frac{1}{\mu}\right) - 0 - 0$$

$$\mathbf{U}_\mu(\mu, \alpha, \boldsymbol{\sigma^2}) = -\mathbf{n} + \frac{\mathbf{K}}{\mu}.$$

*Score for $\alpha$*

The score function $U_\alpha$ is obtained by differentiating $\ell$ with respect to $\alpha$:

$$U_\alpha(\mu, \alpha, \sigma^2) = \frac{\partial \ell}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(-n\mu + K\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha)\right)$$

$$U_\alpha(\mu,\alpha,\sigma^2) = 0 + 0 - 0 - \frac{1}{2\sigma^2}\frac{\partial}{\partial\alpha}\left[\sum_{i:\Delta_i=1}(x_i-\alpha k_i)^2\right]$$

Using the chain rule, $\frac{\partial}{\partial\alpha}(x_i-\alpha k_i)^2 = 2(x_i-\alpha k_i)(-k_i)$:

$$U_\alpha(\mu,\alpha,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i:\Delta_i=1}2(x_i-\alpha k_i)(-k_i)$$

$$\mathbf{U_\alpha(\mu,\alpha,\sigma^2) = \frac{1}{\sigma^2}\sum_{i:\Delta_i=1}k_i(x_i-\alpha k_i).}$$

*Score for $\sigma^2$*

The score function $U_{\sigma^2}$ is obtained by differentiating $\ell$ with respect to $\sigma^2$:

$$U_{\sigma^2}(\mu,\alpha,\sigma^2) = \frac{\partial\ell}{\partial\sigma^2} = \frac{\partial}{\partial\sigma^2}\left(-n\mu + K\log\mu - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\alpha)\right)$$

$$U_{\sigma^2}(\mu,\alpha,\sigma^2) = 0 + 0 - \frac{m}{2}\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}S(\alpha)\frac{\partial}{\partial\sigma^2}(\sigma^{-2})$$

Since $\frac{\partial}{\partial\sigma^2}(\sigma^{-2}) = -(\sigma^2)^{-2} = -\frac{1}{\sigma^4}$:

$$U_{\sigma^2}(\mu,\alpha,\sigma^2) = -\frac{m}{2\sigma^2} - \frac{1}{2}S(\alpha)\left(-\frac{1}{\sigma^4}\right)$$

$$\mathbf{U_{\sigma^2}(\mu,\alpha,\sigma^2) = -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}S(\alpha).}$$

**(c) Observed information (negative Hessian)**

**Compute second derivatives and take their negatives to obtain the observed information matrix $J(\theta) = -\nabla^2\ell(\theta)$.**

- $-\frac{\partial^2\ell}{\partial\mu^2} = \frac{K}{\mu^2}$.
- $-\frac{\partial^2\ell}{\partial\alpha^2} = \frac{1}{\sigma^2}\sum_{i:\Delta_i=1}k_i^2$.
- $-\frac{\partial^2\ell}{\partial(\sigma^2)^2} = \frac{m}{2\sigma^4} - \frac{S(\alpha)}{\sigma^6}$ **(which can also be written by differentiating $U_{\sigma^2}$ and taking the negative).**
- **Mixed second derivatives:**

$$-\frac{\partial^2\ell}{\partial\alpha\partial\mu}=0,\quad -\frac{\partial^2\ell}{\partial\alpha\partial\sigma^2}=-\frac{1}{\sigma^4}\sum_{i:\Delta_i=1}k_i(x_i-\alpha k_i),$$

**and**

$$-\frac{\partial^2\ell}{\partial\mu\partial\sigma^2}=0.$$

**Hence the observed information is a $3\times 3$ matrix with those entries.**

The observed information matrix is $J(\theta)=-\nabla^2\ell(\theta)$, which consists of the negatives of the second partial derivatives of the log-likelihood function. We verify the elements using the score functions from part (b):

$$U_\mu=-n+\frac{K}{\mu}\qquad U_\alpha=\frac{1}{\sigma^2}\sum k_i(x_i-\alpha k_i)\qquad U_{\sigma^2}=-\frac{m}{2\sigma^2}+\frac{S(\alpha)}{2\sigma^4}$$

where $\sum$ denotes $\sum_{i:\Delta_i=1}$.

The diagonal elements are: **For $\mu$:**

$$\frac{\partial^2\ell}{\partial\mu^2}=\frac{\partial}{\partial\mu}(-n+K\mu^{-1})=-K\mu^{-2}\quad\implies -\frac{\partial^2\ell}{\partial\mu^2}=\frac{K}{\mu^2}.$$

**For $\alpha$:**

$$\frac{\partial^2\ell}{\partial\alpha^2}=\frac{\partial}{\partial\alpha}\left[\frac{1}{\sigma^2}\sum(x_ik_i-\alpha k_i^2)\right]=-\frac{1}{\sigma^2}\sum k_i^2\quad\implies -\frac{\partial^2\ell}{\partial\alpha^2}=\frac{1}{\sigma^2}\sum k_i^2.$$

**For $\sigma^2$ ($v=\sigma^2$):**

$$\frac{\partial^2\ell}{\partial v^2}=\frac{\partial}{\partial v}\left(-\frac{m}{2}v^{-1}+\frac{S(\alpha)}{2}v^{-2}\right)=\frac{m}{2v^2}-\frac{S(\alpha)}{v^3}\quad\implies -\frac{\partial^2\ell}{\partial(\sigma^2)^2}=\frac{S(\alpha)}{\sigma^6}-\frac{m}{2\sigma^4}.$$

**The off diagonal elements for $\mu$ and $\alpha$ are:**

$$\frac{\partial^2\ell}{\partial\mu\partial\alpha}=\frac{\partial}{\partial\alpha}U_\mu=0\quad\implies -\frac{\partial^2\ell}{\partial\mu\partial\alpha}=0.$$

**For $\mu$ and $\sigma^2$:**

$$\frac{\partial^2\ell}{\partial\mu\partial\sigma^2}=\frac{\partial}{\partial\sigma^2}U_\mu=0\quad\implies -\frac{\partial^2\ell}{\partial\mu\partial\sigma^2}=0.$$

**For $\alpha$ and $\sigma^2$ ($v=\sigma^2$):**

$$\frac{\partial^2\ell}{\partial\alpha\partial v}=\frac{\partial}{\partial v}\left[v^{-1}\sum k_i(x_i-\alpha k_i)\right]=-\frac{1}{v^2}\sum k_i(x_i-\alpha k_i)\quad\implies -\frac{\partial^2\ell}{\partial\alpha\partial\sigma^2}=\frac{1}{\sigma^4}\sum k_i(x_i-\alpha k_i).$$

The observed information matrix $J(\theta)$ is:

$$J(\theta)=\begin{pmatrix}-\frac{\partial^2\ell}{\partial\mu^2} & -\frac{\partial^2\ell}{\partial\mu\partial\alpha} & -\frac{\partial^2\ell}{\partial\mu\partial\sigma^2}\\ -\frac{\partial^2\ell}{\partial\alpha\partial\mu} & -\frac{\partial^2\ell}{\partial\alpha^2} & -\frac{\partial^2\ell}{\partial\alpha\partial\sigma^2}\\ -\frac{\partial^2\ell}{\partial\sigma^2\partial\mu} & -\frac{\partial^2\ell}{\partial\sigma^2\partial\alpha} & -\frac{\partial^2\ell}{\partial(\sigma^2)^2}\end{pmatrix}$$

Substituting the results:

$$J(\theta) = \begin{pmatrix} \frac{K}{\mu^2} & 0 & 0 \\ 0 & \frac{1}{\sigma^2}\sum k_i^2 & \frac{1}{\sigma^4}\sum k_i(x_i - \alpha k_i) \\ 0 & \frac{1}{\sigma^4}\sum k_i(x_i - \alpha k_i) & \frac{S(\alpha)}{\sigma^6} - \frac{m}{2\sigma^4} \end{pmatrix}$$

The matrix shows that $\mu$ is informationally orthogonal to $(\alpha, \sigma^2)$, but $\alpha$ and $\sigma^2$ are not informationally orthogonal.

**(d) Fisher information (expected negative Hessian)**

**Take expectations (under the model) of the negative second derivatives to obtain the Fisher information $I(\theta) = \mathbb{E}[J(\theta)]$.**

**Key components:**

- $I_{\mu\mu} = \mathbb{E}\left[\frac{K}{\mu^2}\right] = \frac{m_{\text{exp}}}{\mu^2}$ **— in practice, under the model the expected number of observed $X$'s depends on $\mu$, but if we condition on observed $\Delta_i$'s the usual plug-in is $I_{\mu\mu} = K/\mu^2$.**
- $I_{\alpha\alpha} = \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{i:\Delta_i=1} k_i^2\right]$.
- **The cross-terms $I_{\alpha\mu} = 0$ and $I_{\mu,\sigma^2} = 0$ (because the Poisson part and Normal scale separate), while $I_{\alpha,\sigma^2} = 0$ in expectation because $\mathbb{E}[x_i - \alpha k_i] = 0$ conditional on $k_i$.**

**Thus the Fisher information matrix is typically (approximately) block diagonal between $\mu$ and $(\alpha, \sigma^2)$.**

We consider the diagonal elements, For $\mu$ ($I_{\mu\mu}$):

$$I_{\mu\mu} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \mu^2}\right] = \mathbb{E}\left[\frac{K}{\mu^2}\right] = \frac{1}{\mu^2}\mathbb{E}[K] = \frac{n\mu}{\mu^2} = \frac{n}{\mu}$$

**Note:** If treating $K$ as fixed based on observation, $I_{\mu\mu} \approx \frac{K}{\mu^2}$. Using the unconditional expectation $\mathbb{E}[K] = n\mu$ gives $I_{\mu\mu} = n/\mu$.

For $\alpha$ ($I_{\alpha\alpha}$):

$$I_{\alpha\alpha} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \alpha^2}\right] = \mathbb{E}\left[\frac{1}{\sigma^2}\sum_{i:\Delta_i=1} k_i^2\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[\Delta_i N_i^2]$$

Since $\Delta_i = 1$ if $N_i > 0$ and $N_i > 0 \implies \Delta_i = 1$, we have $N_i^2\Delta_i = N_i^2$.

$$I_{\alpha\alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}[N_i^2] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu + \mu^2) = \frac{n(\mu + \mu^2)}{\sigma^2}.$$

**Note:** The provided key component $\frac{1}{\sigma^2}\mathbb{E}\left[\sum_{i:\Delta_i=1} k_i^2\right]$ is equivalent to $\frac{1}{\sigma^2}\sum \mathbb{E}[N_i^2 \mid N_i > 0]P(\Delta_i = 1)$. The result $n(\mu + \mu^2)/\sigma^2$ is derived from the more standard $\mathbb{E}[\sum N_i^2]/\sigma^2$.

For $\sigma^2$ ($I_{\sigma^2\sigma^2}$): We use the fact that $\mathbb{E}[S(\alpha)] = \mathbb{E}[\sum_{i:\Delta_i=1}(X_i - \alpha k_i)^2]$. Since $X_i \mid k_i \sim N(\alpha k_i, \sigma^2)$, we have $\mathbb{E}[(X_i - \alpha k_i)^2 \mid k_i] = \sigma^2$.

$$\mathbb{E}[S(\alpha)] = \mathbb{E}\left[\sum_{i:\Delta_i=1} \mathbb{E}[(X_i - \alpha k_i)^2 \mid k_i]\right] = \mathbb{E}\left[\sum_{i:\Delta_i=1} \sigma^2\right] = \mathbb{E}[m\sigma^2] = \sigma^2\mathbb{E}[m] = np\sigma^2$$

where $p = 1 - e^{-\mu}$ is the probability of observation.

$$I_{\sigma^2\sigma^2} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial(\sigma^2)^2}\right] = \mathbb{E}\left[\frac{S(\alpha)}{\sigma^6} - \frac{m}{2\sigma^4}\right] = \frac{1}{\sigma^6}\mathbb{E}[S(\alpha)] - \frac{1}{2\sigma^4}\mathbb{E}[m]$$

$$I_{\sigma^2\sigma^2} = \frac{1}{\sigma^6}(np\sigma^2) - \frac{1}{2\sigma^4}(np) = \frac{np}{\sigma^4} - \frac{np}{2\sigma^4} = \frac{np}{2\sigma^4}$$

where $p = 1 - e^{-\mu}$.

We consider the off-diagonal elements, For $\mu$ and $\alpha$ ($I_{\mu\alpha}$):

$$I_{\mu\alpha} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial\mu\partial\alpha}\right] = \mathbb{E}[0] = 0. \quad \checkmark$$

For $\mu$ and $\sigma^2$ ($I_{\mu\sigma^2}$):

$$I_{\mu\sigma^2} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial\mu\partial\sigma^2}\right] = \mathbb{E}[0] = 0. \quad \checkmark$$

For $\alpha$ and $\sigma^2$ ($I_{\alpha\sigma^2}$):

$$I_{\alpha\sigma^2} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial\alpha\partial\sigma^2}\right] = \mathbb{E}\left[\frac{1}{\sigma^4}\sum_{i:\Delta_i=1} k_i(x_i - \alpha k_i)\right]$$

By the law of total expectation:

$$\mathbb{E}[k_i(x_i - \alpha k_i)] = \mathbb{E}\left[\mathbb{E}[k_i(x_i - \alpha k_i) \mid k_i]\right] = \mathbb{E}\left[k_i \cdot \mathbb{E}[x_i - \alpha k_i \mid k_i]\right]$$

Since $X_i \mid k_i \sim N(\alpha k_i, \sigma^2)$, $\mathbb{E}[x_i - \alpha k_i \mid k_i] = \alpha k_i - \alpha k_i = 0$.

$$I_{\alpha\sigma^2} = \frac{1}{\sigma^4}\sum \mathbb{E}[0] = 0. \quad \checkmark$$

The Fisher information matrix is:

$$I(\theta) = \begin{pmatrix} I_{\mu\mu} & I_{\mu\alpha} & I_{\mu\sigma^2} \\ I_{\alpha\mu} & I_{\alpha\alpha} & I_{\alpha\sigma^2} \\ I_{\sigma^2\mu} & I_{\sigma^2\alpha} & I_{\sigma^2\sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{n}{\mu} & 0 & 0 \\ 0 & \frac{n(\mu+\mu^2)}{\sigma^2} & 0 \\ 0 & 0 & \frac{n(1-e^{-\mu})}{2\sigma^4} \end{pmatrix}$$

The Fisher information matrix is **block diagonal** with respect to the parameter vector $\theta = (\mu, \alpha, \sigma^2)$, meaning all parameters are informationally orthogonal.

**(e) Newton–Raphson update at iteration $t$**

**Write the parameter vector $\theta = (\mu, \alpha, \sigma^2)$. The Newton–Raphson update is**

$$\theta^{(t+1)} = \theta^{(t)} - [\ -\nabla^2 \ell(\theta^{(t)})\ ]^{-1}\, \nabla \ell(\theta^{(t)}).$$

**Equivalently,**

$$\theta^{(t+1)} = \theta^{(t)} + J(\theta^{(t)})^{-1} U(\theta^{(t)}),$$

**where $U$ is the score vector and $J$ the observed information matrix evaluated at $\theta^{(t)}$. Because of zero (or small) cross-derivatives a practical implementation often updates parameters in blocks (e.g. update $\alpha, \sigma^2$ from the Normal conditional on current $\mu$, and update $\mu$ from the Poisson part).**

The Newton–Raphson iterative scheme for finding the maximum likelihood estimates (MLEs) $\hat{\theta}$ is:

$$\theta^{(t+1)} = \theta^{(t)} - [\ -\nabla^2 \ell(\theta^{(t)})\ ]^{-1}\, \nabla \ell(\theta^{(t)}).$$

Let the **Score Vector** be $U(\theta) = \nabla \ell(\theta)$ and the **Observed Information Matrix** be $J(\theta) = -\nabla^2 \ell(\theta)$. Substituting these definitions yields the equivalent, standard form:

$$\theta^{(t+1)} = \theta^{(t)} + J(\theta^{(t)})^{-1} U(\theta^{(t)}).$$

Given that the Observed Information Matrix $J(\theta)$ for the Poisson-gated normal model is block-diagonal between $\mu$ and $(\alpha, \sigma^2)$, the full $3 \times 3$ inversion can be replaced by two smaller, more stable updates:

1. Update $\mu$ (Poisson Block):

$$\mu^{(t+1)} = \mu^{(t)} + \left[ J_{\mu\mu}(\theta^{(t)}) \right]^{-1} U_\mu(\theta^{(t)})$$

2. Update $(\alpha, \sigma^2)$ (Normal Block):

$$\begin{pmatrix} \alpha^{(t+1)} \\ (\sigma^2)^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ (\sigma^2)^{(t)} \end{pmatrix} + \begin{pmatrix} J_{\alpha\alpha} & J_{\alpha\sigma^2} \\ J_{\sigma^2\alpha} & J_{\sigma^2\sigma^2} \end{pmatrix}^{-1} \begin{pmatrix} U_\alpha \\ U_{\sigma^2} \end{pmatrix}$$

   where all components are evaluated at the current estimate $\theta^{(t)}$.

**(f) Fisher scoring (replace observed Hessian by expected Fisher information)**

**The Fisher scoring iteration replaces the observed information by the Fisher information:**

$$\theta^{(t+1)} = \theta^{(t)} + I(\theta^{(t)})^{-1} U(\theta^{(t)}).$$

**This tends to be more stable when the observed Hessian is noisy or near-singular. When the Fisher information is block diagonal this simplifies to separate updates for the Poisson and Normal parameters.**

The Fisher Scoring iteration is defined by replacing the observed information $J(\theta^{(t)})$ in the Newton–Raphson update with the expected Fisher information $I(\theta^{(t)})$:

$$\theta^{(t+1)} = \theta^{(t)} + I(\theta^{(t)})^{-1} U(\theta^{(t)}).$$

This method typically provides a more stable optimization path than Newton–Raphson, as the Fisher information $I(\theta)$ is non-random (constant for a fixed $\theta$) and globally positive definite.

For the Poisson-gated normal model, the Fisher Information Matrix $I(\theta)$ is block diagonal, $I(\theta) = \text{diag}(I_{\mu\mu}, I_{\alpha\alpha}, I_{\sigma^2\sigma^2})$. This allows the parameter vector $\theta = (\mu, \alpha, \sigma^2)$ to be updated with three separate, simpler steps:

1. Update $\mu$ (Poisson):
$$\mu^{(t+1)} = \mu^{(t)} + \left[ I_{\mu\mu}(\theta^{(t)}) \right]^{-1} U_\mu(\theta^{(t)})$$
2. Update $\alpha$ (Normal):
$$\alpha^{(t+1)} = \alpha^{(t)} + \left[ I_{\alpha\alpha}(\theta^{(t)}) \right]^{-1} U_\alpha(\theta^{(t)})$$
3. Update $\sigma^2$ (Normal):
$$(\sigma^2)^{(t+1)} = (\sigma^2)^{(t)} + \left[ I_{\sigma^2\sigma^2}(\theta^{(t)}) \right]^{-1} U_{\sigma^2}(\theta^{(t)})$$
where the $I$ terms are the expected information calculated in part (d) and the $U$ terms are the scores calculated in part (b), all evaluated at $\theta^{(t)}$.

**(g) Boundary and edge cases**

Discuss briefly:

- If $\mu$ is very small the number of observed $X_i$ (i.e. $m$) may be zero or too small to estimate $\alpha$ and $\sigma^2$. If $m = 0$ the Normal parameters $\alpha, \sigma^2$ are not identifiable from the data — the likelihood does not depend on them.

When the Poisson parameter $\mu$ is very small, the following issues arise for the estimation of the Normal parameters $\alpha$ and $\sigma^2$:

The probability of a positive Poisson count $P(N_i > 0) = 1 - e^{-\mu}$ is low. If $\mu$ is small enough that $N_i = 0$ for all $n$ units, the number of observed Normal outcomes $m$ is zero ($m = 0$). The log-likelihood simplifies to:

$$\ell(\mu, \alpha, \sigma^2) = -n\mu + \text{const.}$$

Since $\ell$ depends only on $\mu$, the parameters $\alpha$ and $\sigma^2$ are **not identifiable** from the data; the observed data provide no information about them. If $m$ is small but greater than zero ($m \geq 1$), the estimation problem is ill-conditioned.

- Estimating $\alpha$ and $\sigma^2$ requires a sufficient number of non-zero observations. With few observations, the estimates are prone to high variance.

- A small $m$ often leads to the $\mathbf{2} \times \mathbf{2}$ lower-right block of the Observed Information Matrix $J(\theta)$ being **near-singular**.

This instability causes iterative optimization algorithms (like NR and FS) to become **unstable** or fail to converge for $\alpha$ and $\sigma^2$.

- If $\sigma^2$ is estimated near zero, care must be taken: the quadratic form $S(\alpha)/(2\sigma^2)$ can dominate and the observed information with respect to $\sigma^2$ becomes large; numerical regularization or parameter transformations (e.g. work on $\log \sigma^2$) improve stability.

When the estimated variance $\hat{\sigma}^2$ approaches zero, the following numerical issues arise:

1. Domination of Likelihood: The term $\frac{1}{2\sigma^2} S(\alpha)$ in the log-likelihood $\ell(\mu, \alpha, \sigma^2)$ becomes extremely large (dominating) or undefined if $\hat{\sigma}^2 \to 0$. This forces the optimizer to find a value of $\alpha$ that minimizes the residual sum of squares $S(\alpha)$ precisely, which is often unstable.
2. Large Information: The Observed Information Matrix entry $J_{\sigma^2\sigma^2}$ becomes large:

$$J_{\sigma^2\sigma^2} = -\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{S(\alpha)}{\sigma^6} - \frac{m}{2\sigma^4} \approx \frac{S(\alpha)}{\sigma^6}.$$

   If $\sigma^2$ is near zero, $J_{\sigma^2\sigma^2}$ explodes. This leads to an **ill-conditioned Hessian** ($\nabla^2 \ell$) or a nearly singular $J(\theta)$, destabilizing iterative methods like Newton–Raphson and Fisher Scoring.
3. Mitigation Strategies: To maintain numerical stability and enforce the constraint $\sigma^2 > 0$:
   - **Parameter Transformation:** Work with $\eta = \log \sigma^2$ or $\log \sigma$, as $\eta$ is unbounded.
   - **Numerical Regularization:** Add a small constant to $\sigma^2$ (e.g., $\sigma^2 + \epsilon$) or apply penalized likelihood methods to stabilize the Hessian.

- Because the Poisson and Normal components partly separate, one can often (and stably) alternate conditional updates: update $\mu$ from the Poisson marginal (score $U_\mu = -n + K/\mu$), and update $\alpha, \sigma^2$ from the Normal regression of $x_i$ on $k_i$ restricted to observed cases.

The structure of the Poisson-gated normal likelihood, where the parameters $\mu$ and $(\alpha, \sigma^2)$ are informationally orthogonal ($I_{\mu\alpha} = I_{\mu\sigma^2} = 0$), makes the model highly suitable for **Alternating Conditional Updates (ACU)**, a form of block coordinate ascent often used for stable estimation.

Procedure: At each iteration, the parameters are updated in two separate blocks:

1. **Update $\mu$ (Poisson Block):** Estimate $\mu$ using only the information from the Poisson component, treating the current values of $(\alpha, \sigma^2)$ as fixed. Since $\mu$ is only coupled to the data through the total observed count $\mathbf{K} = \sum \mathbf{k_i}$ and the sample size $\mathbf{n}$, the update is often simple, based on setting the score $\mathbf{U}_\mu = -\mathbf{n} + \mathbf{K}/\mu$ to zero. This yields an MLE of $\hat{\mu} = K/n$.

2. **Update $\alpha$ and $\sigma^2$ (Normal Block):** Estimate $(\alpha, \sigma^2)$ using only the $\mathbf{m}$ observed pairs $(\mathbf{x_i}, \mathbf{k_i})$, treating the current $\mu$ as fixed. This is equivalent to fitting a simple linear regression model ($\mathbf{x_i} \sim \mathbf{N}(\alpha \mathbf{k_i}, \sigma^{\mathbf{2}})$) to the observed data. The updates are found by solving the joint likelihood equations for $\mathbf{U}_\alpha$ and $\mathbf{U}_{\sigma^2}$.

Advantages:

- **Stability:** ACU is typically more stable than a full matrix inversion (NR or FS) because it avoids simultaneous large steps and is guaranteed to increase the likelihood at every step.
- **Simplicity/Efficiency:** It leverages the separation in the model structure, replacing a complex $3 \times 3$ matrix inversion with simpler, independent optimization steps for each block.

**Appendix**

```
knitr::opts_chunk$set(
  dev = "cairo_pdf",
  fig.width = 5,
  fig.height = 5,
  fig.align = 'center',
  echo = FALSE,
  message = FALSE,
  warning = FALSE,
  error = FALSE,
  results = 'markup'
)

# Load required libraries
library("tidyverse")
library("patchwork")
library("glue")
library("scales", warn.conflicts = FALSE)
library("extrafont")
library("tinytex")
library("knitr")
library("tidyr")
library("latex2exp")
library("MASS")
library("kableExtra")

theme_set(theme_minimal(base_family = "Roboto Condensed"))

conflicted::conflicts_prefer(
  readr::col_factor(),
  purrr::discard(),
  dplyr::lag(),
  readr::parse_date(),
  kableExtra::group_rows(),
  dplyr::select
)
```