# STA 6351, Report.1.10

**Carson Slater** *Baylor University*

### 1.10

Using the coupled Poisson-normal development above, proceed as follows.

**(a) Differentiate the log-likelihood**

$$\ell(\lambda, \mu, \sigma^2) = \log L(\lambda, \mu, \sigma^2 \mid x)$$

**with respect to $\lambda$, and show that the score function is**

$$U_\lambda(\lambda, \mu, \sigma^2) = \sum_{i=1}^{n} \left( -1 + \frac{\mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2]}{\lambda} \right).$$

The log-likelihood for a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$\ell(\lambda, \mu, \sigma^2 \mid \mathbf{x}) = \sum_{i=1}^{n} \log f_X(x_i),$$

where $f_X(x)$ is the marginal density of $X$. For the Poisson-coupled-normal model, $X$ is a compound Poisson random variable, $X = \sum_{k=1}^{Y} Z_k$, where $Y \sim \text{Poisson}(\lambda)$ and $Z_k \sim N(\mu, \sigma^2)$. The density is

$$f_X(x) = \sum_{y=0}^{\infty} P(Y = y) f_{X|Y}(x|y) = \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} f_{X|Y}(x|y).$$

The score function is $U_\lambda(\lambda, \mu, \sigma^2) = \frac{\partial}{\partial\lambda}\ell(\lambda, \mu, \sigma^2 \mid \mathbf{x}) = \sum_{i=1}^{n} \frac{\partial}{\partial\lambda} \log f_X(x_i) = \sum_{i=1}^{n} \frac{1}{f_X(x_i)} \frac{\partial}{\partial\lambda} f_X(x_i)$.

First, we differentiate $f_X(x)$ with respect to $\lambda$:

$$\frac{\partial}{\partial\lambda} f_X(x) = \sum_{y=0}^{\infty} f_{X|Y}(x|y) \frac{\partial}{\partial\lambda} \left[ \frac{e^{-\lambda}\lambda^y}{y!} \right].$$

The derivative of the Poisson probability mass function with respect to $\lambda$ is:

$$\frac{\partial}{\partial\lambda} P(Y = y) = \frac{\partial}{\partial\lambda} \left[ \frac{e^{-\lambda}\lambda^y}{y!} \right] = \frac{1}{y!} \left[ -e^{-\lambda}\lambda^y + y e^{-\lambda}\lambda^{y-1} \right]$$

$$= \frac{e^{-\lambda}\lambda^y}{y!} \left[ -1 + \frac{y}{\lambda} \right] = P(Y = y) \left( \frac{y - \lambda}{\lambda} \right).$$

Substituting this back into the derivative of $f_X(x)$:

$$\frac{\partial}{\partial\lambda} f_X(x) = \sum_{y=0}^{\infty} f_{X|Y}(x|y) P(Y = y) \left( \frac{y - \lambda}{\lambda} \right)$$

$$= \frac{1}{\lambda} \sum_{y=0}^{\infty} (y - \lambda) P(Y = y) f_{X|Y}(x|y).$$

Next, we substitute this into the expression for the score of a single observation $\frac{\partial}{\partial \lambda} \log f_X(x_i)$:

$$\frac{\partial}{\partial \lambda} \log f_X(x_i) = \frac{1}{f_X(x_i)} \frac{\partial}{\partial \lambda} f_X(x_i)$$

$$= \frac{1}{f_X(x_i)} \frac{1}{\lambda} \sum_{y=0}^{\infty} (y - \lambda) P(Y = y) f_{X|Y}(x_i|y)$$

$$= \frac{1}{\lambda} \left[ \sum_{y=0}^{\infty} y \frac{P(Y = y) f_{X|Y}(x_i|y)}{f_X(x_i)} - \lambda \sum_{y=0}^{\infty} \frac{P(Y = y) f_{X|Y}(x_i|y)}{f_X(x_i)} \right].$$

Using Bayes' theorem, the term $\frac{P(Y=y) f_{X|Y}(x_i|y)}{f_X(x_i)}$ is the conditional probability $P(Y = y \mid X = x_i;\, \lambda, \mu, \sigma^2)$. The summation $\sum_{y=0}^{\infty} y P(Y = y \mid X = x_i;\, \lambda, \mu, \sigma^2)$ is the conditional expectation $\mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2]$. The summation $\sum_{y=0}^{\infty} P(Y = y \mid X = x_i;\, \lambda, \mu, \sigma^2)$ is the sum of all conditional probabilities, which equals 1.

Therefore, the score for a single observation is:

$$\frac{\partial}{\partial \lambda} \log f_X(x_i) = \frac{1}{\lambda} \left[ \mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2] - \lambda \cdot 1 \right]$$

$$= \frac{\mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2]}{\lambda} - 1.$$

The score function for the sample is the sum of the individual scores:

$$U_\lambda(\lambda, \mu, \sigma^2) = \sum_{i=1}^{n} \left( \frac{\mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2]}{\lambda} - 1 \right)$$

$$= \sum_{i=1}^{n} \left( -1 + \frac{\mathbb{E}[Y_i \mid X_i = x_i;\, \lambda, \mu, \sigma^2]}{\lambda} \right).$$

**(b) By differentiating the normal log-density inside the mixture, show that the score with respect to $\mu$ can be written as**

$$U_\mu(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0} \sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2) \, \frac{x_i - y\mu}{\sigma^2},$$

**where the conditional probabilities are given by**

$$w_{iy}(\lambda, \mu, \sigma^2) = \Pr(Y_i = y \mid X_i = x_i;\, \lambda, \mu, \sigma^2).$$

The log-likelihood for a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$\ell(\lambda, \mu, \sigma^2 \mid \mathbf{x}) = \sum_{i=1}^{n} \log f_X(x_i).$$

The score function with respect to $\mu$ is $U_\mu(\lambda, \mu, \sigma^2) = \frac{\partial}{\partial \mu}\ell(\lambda, \mu, \sigma^2 \mid \mathbf{x}) = \sum_{i=1}^{n} \frac{\partial}{\partial \mu} \log f_X(x_i)$.

The marginal density of $X$ is $f_X(x) = \sum_{y=0}^{\infty} P(Y = y) f_{X|Y}(x|y)$. For $x = 0$, the density is $f_X(0) = P(Y = 0) = e^{-\lambda}$. This term is **independent of** $\mu$. Thus, $\frac{\partial}{\partial \mu} \log f_X(0) = 0$. The summation for the score function only needs to be over the non-zero observations, i.e., $i : x_i \neq 0$.

$$U_\mu(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0} \frac{\partial}{\partial \mu} \log f_X(x_i).$$

For $x_i \neq 0$, the sum starts at $y = 1$, since $X|Y = 0$ is a point mass at 0.

$$f_X(x_i) = \sum_{y=1}^{\infty} P(Y = y) f_{X|Y}(x_i|y) = \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} f_{X|Y}(x_i|y),$$

where $f_{X|Y}(x_i|y)$ is the density of $N(y\mu, y\sigma^2)$:

$$f_{X|Y}(x_i|y) = \frac{1}{\sqrt{2\pi y\sigma^2}} \exp\left(-\frac{(x_i - y\mu)^2}{2y\sigma^2}\right).$$

We use the general formula for the score: $\frac{\partial}{\partial \mu} \log f_X(x_i) = \frac{1}{f_X(x_i)} \frac{\partial}{\partial \mu} f_X(x_i)$.

First, we differentiate $f_X(x_i)$ with respect to $\mu$. The Poisson term $P(Y = y)$ does not depend on $\mu$:

$$\frac{\partial}{\partial \mu} f_X(x_i) = \sum_{y=1}^{\infty} P(Y = y) \frac{\partial}{\partial \mu} f_{X|Y}(x_i|y).$$

Next, we differentiate the normal log-density $\log f_{X|Y}(x_i|y)$ with respect to $\mu$.

$$\log f_{X|Y}(x_i|y) = -\frac{1}{2}\log(2\pi y\sigma^2) - \frac{(x_i - y\mu)^2}{2y\sigma^2}.$$

The derivative is:

$$\frac{\partial}{\partial \mu} \log f_{X|Y}(x_i|y) = -\frac{1}{2y\sigma^2} \frac{\partial}{\partial \mu}(x_i - y\mu)^2 = -\frac{1}{2y\sigma^2} \cdot 2(x_i - y\mu) \cdot (-y)$$

$$= \frac{x_i - y\mu}{\sigma^2}.$$

This expression is the conditional score $\frac{\partial}{\partial \mu} \log f_{X|Y}(x_i|y)$.

Now we use the identity $\frac{\partial}{\partial \mu} f_{X|Y}(x_i|y) = f_{X|Y}(x_i|y) \cdot \frac{\partial}{\partial \mu} \log f_{X|Y}(x_i|y)$:

$$\frac{\partial}{\partial \mu} f_{X|Y}(x_i|y) = f_{X|Y}(x_i|y) \frac{x_i - y\mu}{\sigma^2}.$$

Substituting this back into $\frac{\partial}{\partial \mu} f_X(x_i)$:

$$\frac{\partial}{\partial \mu} f_X(x_i) = \sum_{y=1}^{\infty} P(Y = y) f_{X|Y}(x_i|y) \frac{x_i - y\mu}{\sigma^2}.$$

Finally, we substitute this into the score for a single observation $x_i$:

$$\frac{\partial}{\partial \mu} \log f_X(x_i) = \frac{1}{f_X(x_i)} \frac{\partial}{\partial \mu} f_X(x_i)$$

$$= \frac{1}{f_X(x_i)} \sum_{y=1}^{\infty} P(Y = y) f_{X|Y}(x_i|y) \frac{x_i - y\mu}{\sigma^2}.$$

We use the definition of the conditional probability:

$$w_{iy}(\lambda, \mu, \sigma^2) = \Pr(Y_i = y \mid X_i = x_i) = \frac{P(Y = y) f_{X|Y}(x_i|y)}{f_X(x_i)}.$$

The score for a single observation $x_i \neq 0$ is:

$$\frac{\partial}{\partial \mu} \log f_X(x_i) = \sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2) \frac{x_i - y\mu}{\sigma^2}.$$

Summing over all $i$ such that $x_i \neq 0$ gives the total score function:

$$U_\mu(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0} \sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2) \, \frac{x_i - y\mu}{\sigma^2}.$$

**(c) Similarly, show that the score with respect to $\sigma^2$ is**

$$U_{\sigma^2}(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0} \sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2) \left( -\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4} \right).$$

The score function with respect to $\sigma^2$ is

$$U_{\sigma^2}(\lambda, \mu, \sigma^2) = \frac{\partial}{\partial \sigma^2} \ell(\lambda, \mu, \sigma^2 \mid \mathbf{x}) = \sum_{i=1}^{n} \frac{\partial}{\partial \sigma^2} \log f_X(x_i).$$

As established in part (b), $\frac{\partial}{\partial \sigma^2} \log f_X(0) = 0$ since $f_X(0) = e^{-\lambda}$ is independent of $\sigma^2$. Thus, the summation is only over the non-zero observations:

$$U_{\sigma^2}(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0} \frac{\partial}{\partial \sigma^2} \log f_X(x_i).$$

For $x_i \neq 0$, the marginal density is $f_X(x_i) = \sum_{y=1}^{\infty} P(Y = y) f_{X|Y}(x_i|y)$. We use the general formula $\frac{\partial}{\partial \sigma^2} \log f_X(x_i) = \frac{1}{f_X(x_i)} \frac{\partial}{\partial \sigma^2} f_X(x_i)$.

First, we differentiate the marginal density with respect to $\sigma^2$. The Poisson term $P(Y = y)$ does not depend on $\sigma^2$:

$$\frac{\partial}{\partial \sigma^2} f_X(x_i) = \sum_{y=1}^{\infty} P(Y = y) \frac{\partial}{\partial \sigma^2} f_{X|Y}(x_i|y).$$

Next, we differentiate the conditional log-density $\log f_{X|Y}(x_i|y)$ with respect to $\sigma^2$, where $f_{X|Y}(x_i|y)$ is the density of $N(y\mu, y\sigma^2)$. It is helpful to treat $v = \sigma^2$ as the variable of interest:

$$\log f_{X|Y}(x_i|y) = -\frac{1}{2}\log(2\pi y) - \frac{1}{2}\log(\sigma^2) - \frac{(x_i - y\mu)^2}{2y\sigma^2}.$$

The derivative with respect to $\sigma^2$ is:

$$\frac{\partial}{\partial\sigma^2}\log f_{X|Y}(x_i|y) = -\frac{1}{2}\frac{\partial}{\partial\sigma^2}\log(\sigma^2) - \frac{(x_i - y\mu)^2}{2y}\frac{\partial}{\partial\sigma^2}(\sigma^2)^{-1}$$

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

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f_{X|Y}(x_i|y) &= -\frac{1}{2\sigma^2} - \frac{(x_i - y\mu)^2}{2y}\left(-\frac{1}{\sigma^4}\right) \\ &= -\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4}.\end{aligned}$$

Using the identity $\frac{\partial}{\partial\sigma^2}f_{X|Y}(x_i|y) = f_{X|Y}(x_i|y)\cdot\frac{\partial}{\partial\sigma^2}\log f_{X|Y}(x_i|y)$, we substitute this back into $\frac{\partial}{\partial\sigma^2}f_X(x_i)$:

$$\frac{\partial}{\partial\sigma^2}f_X(x_i) = \sum_{y=1}^{\infty} P(Y=y) f_{X|Y}(x_i|y)\left(-\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4}\right).$$

Finally, we substitute this into the score for a single observation $x_i$:

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f_X(x_i) &= \frac{1}{f_X(x_i)}\frac{\partial}{\partial\sigma^2}f_X(x_i) \\ &= \frac{1}{f_X(x_i)}\sum_{y=1}^{\infty} P(Y=y) f_{X|Y}(x_i|y)\left(-\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4}\right).\end{aligned}$$

Recognizing the conditional probability $w_{iy}(\lambda, \mu, \sigma^2) = \frac{P(Y=y)f_{X|Y}(x_i|y)}{f_X(x_i)}$, the score for a single observation $x_i \neq 0$ is:

$$\frac{\partial}{\partial\sigma^2}\log f_X(x_i) = \sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2)\left(-\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4}\right).$$

Summing over all $i$ such that $x_i \neq 0$ gives the total score function:

$$U_{\sigma^2}(\lambda, \mu, \sigma^2) = \sum_{i:\, x_i \neq 0}\sum_{y=1}^{\infty} w_{iy}(\lambda, \mu, \sigma^2)\left(-\frac{1}{2\sigma^2} + \frac{(x_i - y\mu)^2}{2y\sigma^4}\right).$$

**(d) Discuss how the score functions depend on the conditional probabilities $w_{iy}$ and explain why closed-form maximum likelihood estimators do not exist, motivating the use of Newton–Raphson or Fisher scoring iterations. Each score component involves the conditional distribution of the latent counts $Y_i$ given $X_i$, and therefore requires the infinite sums weighted by the $w_{iy}$'s, which depend on the parameters. Because of this dependence, closed-form estimators cannot be obtained and iterative methods are required.**

The score functions for the Poisson-coupled-normal model, $\mathbf{U}(\boldsymbol{\theta}) = (U_\lambda, U_\mu, U_{\sigma^2})^\intercal$, all explicitly depend on the conditional probabilities $w_{iy} = \Pr(Y_i = y \mid X_i = x_i;\, \boldsymbol{\theta})$.

1. **Implicit Dependence:** Each score component involves a weighted sum (often an infinite sum) over the latent counts $y$, where the weights are $w_{iy}$. For instance, $U_\lambda \propto \sum_i \mathbb{E}[Y_i \mid X_i]$, and $U_\mu$ and $U_{\sigma^2}$ are weighted averages of the conditional scores $\frac{\partial}{\partial \mu} \log f_{X|Y}$ and $\frac{\partial}{\partial \sigma^2} \log f_{X|Y}$.

2. **Intertwined Parameters:** The conditional probability $w_{iy}$ is a complex ratio involving the likelihood of all parameters $\boldsymbol{\theta} = (\lambda, \mu, \sigma^2)^\intercal$:

$$w_{iy}(\boldsymbol{\theta}) = \frac{P(Y_i = y \mid \lambda) f_{X|Y}(x_i \mid y;\, \mu, \sigma^2)}{f_X(x_i \mid \lambda, \mu, \sigma^2)}.$$

   To find the Maximum Likelihood Estimators (MLEs), the score equations $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ must be solved. Because the conditional probabilities $w_{iy}$ on the right-hand side of these equations are themselves functions of the parameters $\boldsymbol{\theta}$, the resulting system of equations is **non-linear and implicitly defined**.

This implicit and non-linear structure makes it impossible to algebraically isolate the parameters to obtain **closed-form MLEs**.

Since direct solution is impossible, the optimization problem must be solved using **iterative numerical methods**.

- **Newton–Raphson and Fisher Scoring:** These methods, which utilize first and second derivatives (or the Fisher information matrix), provide a structured way to iteratively refine an initial parameter guess $\boldsymbol{\theta}^{(k)}$ towards the root of the score equation $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}(\boldsymbol{\theta}^{(k)})^{-1}\mathbf{U}(\boldsymbol{\theta}^{(k)})$.
- **EM Algorithm:** The structure of the score functions, relying on the conditional expectation of the latent variable $Y$, also perfectly aligns with the **Expectation-Maximization (EM) algorithm**. The EM algorithm often simplifies the optimization by replacing the complex marginal likelihood optimization with simpler conditional expectation (E-step, involving $w_{iy}$) and maximization (M-step) steps.

**(e) To set up the iteration step, let $U(\theta)$ be the score vector and $\mathcal{I}(\theta)$ the Fisher information matrix. The Fisher scoring update in vector form is**

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{I}(\theta^{(t)})^{-1} U(\theta^{(t)}), \qquad \theta = (\lambda, \mu, \sigma^2)^\top.$$

**At each iteration one recomputes the conditional probabilities $w_{iy}$ at the current $\theta^{(t)}$, assembles the scores using the formulas above, builds the Fisher information matrix $\mathcal{I}(\theta^{(t)})$ (for example via the complete-data expectation), and then solves the $3 \times 3$ linear system for the increment $\mathcal{I}(\theta^{(t)})^{-1} U(\theta^{(t)})$.**

The Fisher Scoring algorithm provides a general approach for finding the roots of the score equations, $U(\boldsymbol{\theta}) = \mathbf{0}$. This method is often preferred over the Newton–Raphson method as it uses the expected information, $\mathcal{I}(\boldsymbol{\theta})$, which ensures the matrix is positive definite and improves stability.

At each iteration $t$, the algorithm requires the computation of three main components based on the current parameter estimates $\boldsymbol{\theta}^{(t)} = (\lambda^{(t)}, \mu^{(t)}, (\sigma^2)^{(t)})^\top$:

1. **Conditional Probabilities ($w_{iy}^{(t)}$):** The weights are re-evaluated using the current parameter estimates for all $i$ and $y \geq 1$:

$$w_{iy}^{(t)} = \Pr(Y_i = y \mid X_i = x_i;\, \boldsymbol{\theta}^{(t)}) = \frac{P(Y_i = y \mid \lambda^{(t)}) f_{X|Y}(x_i \mid y;\, \mu^{(t)}, (\sigma^2)^{(t)})}{f_X(x_i \mid \boldsymbol{\theta}^{(t)})}.$$

2. **Score Vector ($U(\boldsymbol{\theta}^{(t)})$):** The $3 \times 1$ score vector is assembled using the component formulas derived in (a), (b), and (c), substituting $w_{iy}^{(t)}$:

$$U(\boldsymbol{\theta}^{(t)}) = \begin{pmatrix} U_\lambda(\boldsymbol{\theta}^{(t)}) \\ U_\mu(\boldsymbol{\theta}^{(t)}) \\ U_{\sigma^2}(\boldsymbol{\theta}^{(t)}) \end{pmatrix},$$

   where the conditional expectation $\mathbb{E}[Y_i \mid X_i = x_i]$ within $U_\lambda$ is computed as $\sum_{y=0}^{\infty} y\, w_{iy}^{(t)}$.

3. **Fisher Information Matrix ($\mathcal{I}(\boldsymbol{\theta}^{(t)})$):** This $3 \times 3$ matrix is typically calculated as the expected value of the negative second partial derivatives (Hessian matrix), $\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$. A practical approach for latent variable models is to use the **Complete-Data Fisher Information** and exploit the EM algorithm's structure, $\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{X}}\left[\frac{\partial^2 \ell_c(\boldsymbol{\theta}|\mathbf{X},\mathbf{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$, where the expectation is taken with respect to the observed data $\mathbf{X}$ and the latent data $\mathbf{Y}$.

The final step is to solve the linear system for the increment $\Delta\boldsymbol{\theta}^{(t)}$:

$$\mathcal{I}(\boldsymbol{\theta}^{(t)}) \cdot \Delta\boldsymbol{\theta}^{(t)} = U(\boldsymbol{\theta}^{(t)}),$$

where $\Delta\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}$. The new parameter vector $\boldsymbol{\theta}^{(t+1)}$ is then calculated:

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

The process is repeated until the change in the parameter vector, $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||$, falls below a specified tolerance, indicating convergence to the MLE.