Statistics 221 Final Project: C++ team

December 13, 2013

1 Compare MLE to SGD

MLE methods such Newton-Raphson compute the gradient of the log-likelihood, $\nabla\ell$ for all observations in one step. While this works well for small samples, it is too computationally expensive to be done with especially large amounts of data. Sakrison's method relies on the assumption that the expectation of the gradient for a single observation is proportional to the gradient for all observations. By the law of large numbers, repeatedly updating using the gradient for single observations will converge to expectations. This makes it possible to do the update step for a single observation at a time and compute the MLE.

The implicit updates use the update step, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + a_t\nabla\ell(\boldsymbol{\theta}_{t+1}; y_t, \boldsymbol{x}_t)$ instead of $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + a_t\nabla\ell(\boldsymbol{\theta}_t; y_t, \boldsymbol{x}_t)$. However the expectation of the gradient is the same for all observations. Therefore $\nabla\ell(\boldsymbol{\theta}_{t+1}; y_t, \boldsymbol{x}_t) = \nabla\ell(\boldsymbol{\theta}_t; y_t, \boldsymbol{x}_t)$. Since the implicit method has the same expectation as Sakrison's method, it must also compute the MLE.

2 GLM Proofs

(a) Show $E(y_t|\boldsymbol{x}_t) = h(\boldsymbol{x}_t^T\boldsymbol{\theta}^*) = b'(\eta_t)$

We start with the moment generating function, which we solve for using

LOTUS

$$
\begin{aligned}
M_Y(t) &= E[e^{tY}] \\
&= \int_Y \exp(ty_k) f(y_k|\eta_k) dy_k \\
&= \int_Y \exp(ty_k) \exp\left(\frac{\eta_t y_t - b(\eta_t)}{\phi}\right) \cdot c(y_t, \phi) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(ty_k + \frac{\eta_t y_t}{\phi} - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(\left(\frac{\eta_t y_t + \phi t}{\phi}\right) y_k - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(\left(\frac{\eta_t y_t + \phi t}{\phi}\right) y_k - \frac{b(\eta_k + \phi t)}{\phi} + \frac{b(\eta_k + \phi t)}{\phi} - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) c(y_t, \phi) \exp\left(\frac{(\eta_t y_t + \phi t) y_k - b(\eta_k + \phi t)}{\phi}\right) dy_k \\
&= \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) \int_Y c(y_t, \phi) \exp\left(\frac{(\eta_t y_t + \phi t) y_k - b(\eta_k + \phi t)}{\phi}\right) dy_k
\end{aligned}
$$

The remaining integral is the conditional pdf of $y|\eta_k + \phi t$ and integrates to 1, which leaves the remaining MGF for Y

$$M_Y(t) = \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right)$$

We then use the MGF to find $E(y_t|\boldsymbol{x}_t)$. We start by finding $M'_Y(t)$

$$
\begin{aligned}
M'_Y(t) &= \frac{\partial}{\partial t}\left[\exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right)\right] \\
&= \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) \cdot \frac{b'(\eta_k + \phi t)\phi}{\phi} \\
&= \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) \cdot b'(\eta_k + \phi t)
\end{aligned}
$$

We then evaluate at $t = 0$ which is the first moment.

$$
\begin{aligned}
E(y_t|\boldsymbol{x}_t) &= M'_Y(0) \\
&= \exp\left(\frac{b(\eta_k + \phi \cdot 0) - b(\eta_t)}{\phi}\right) \cdot b'(\eta_k + \phi \cdot 0) \\
&= \exp\left(\frac{0}{\phi}\right) \cdot b'(\eta_k) \\
&= b'(\eta_k)
\end{aligned}
$$

(b) Show $\text{Var}(y_t|\eta_t) = \phi \cdot h'(\eta_t)$
We find the variance using the moment generating function from part (a).

$$\text{Var}(y_k|\eta_k) = E(y_k^2|\eta_k) - [E(y_k|\eta_k)]^2$$

We can use the value of $E(y_k|\eta_k)$ from part (a). We then find the second moment $E(y_k^2|\eta_k)$ with our MGF.

$$\begin{aligned}
M_Y''(t) =& \frac{\partial}{\partial t}\left[\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi t)\right] \\
=& \exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi t)\cdot b'(\eta_k+\phi t) + \\
& b''(\eta_k+\phi t)\cdot\phi\cdot\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right) \\
=& \exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot\left[b'(\eta_k+\phi t)\right]^2 + b''(\eta_k+\phi t)\cdot\phi\cdot\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)
\end{aligned}$$

Evaluating at 0 gives

$$\begin{aligned}
E(y_k^2|\eta_k) =& M_Y''(0) \\
=& \exp\left(\frac{b(\eta_k+\phi*0)-b(\eta_t)}{\phi}\right)\left[b'(\eta_k+\phi*0)\right]^2 + b''(\eta_k+\phi*0)\phi\exp\left(\frac{b(\eta_k+\phi*0)-b(\eta_t)}{\phi}\right) \\
=& \exp(\frac{0}{\phi})\left[b'(\eta_k)\right]^2 + \phi\cdot b''(\eta_k)\exp(\frac{0}{\phi}) \\
=& \left[b'(\eta_k)\right]^2 + \phi\cdot b''(\eta_k)
\end{aligned}$$

Putting the two parts together gives the solution

$$\begin{aligned}
\text{Var}(y_t|\eta_t) &= E(y_k^2|\eta_k) - [E(y_k|\eta_k)]^2 \\
&= [b'(\eta_k)]^2 + \phi\cdot b''(\eta_k) - [b'(\eta_k)]^2 \\
&= \phi\cdot b''(\eta_k)
\end{aligned}$$

From part (a) we know that $b'(\eta_t) = h(\eta)$ therefore $b''(\eta_t) = h'(\eta_t)$ which gives

$$\text{Var}(y_t|\eta_t) = \phi\cdot h'(\eta_t)$$

(c) Show $\nabla\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) = \frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T\theta))\boldsymbol{x}_t$
To find $\nabla\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)$, we take the partial derivative with respect to $\boldsymbol{\theta}$. The

likelihood is proportional to the pdf.

$$
\begin{aligned}
L(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &\propto \exp\left(\frac{\eta_t y_t - b(\eta_t)}{\phi}\right) \cdot c(y_t, \phi) \\
\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &= \frac{\eta_t y_t - b(\eta_t)}{\phi} \\
\nabla\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)\right] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{\eta_t y_t - b(\eta_t)}{\phi}\right] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{\boldsymbol{x}_t^T\boldsymbol{\theta} y_t - b(\boldsymbol{x}_t^T\boldsymbol{\theta})}{\phi}\right] \\
&= \frac{1}{\phi}(\boldsymbol{x}_t y_t - \boldsymbol{x}_t b'(\boldsymbol{x}_t^T\boldsymbol{\theta})) \\
&= \frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T\boldsymbol{\theta}))\boldsymbol{x}_t
\end{aligned}
$$

(d) Show $\mathcal{J}(\boldsymbol{\theta}) = -E(\nabla\nabla\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)) = \frac{1}{\phi}E(h'(\boldsymbol{x}_t^T\boldsymbol{\theta})\boldsymbol{x}_t\boldsymbol{x}_t')$
To find the Fisher information, we take the negative expectation of the second partial derivative of the log-likelihood with respect to $\boldsymbol{\theta}$. We start with the the solution from part (c).

$$
\begin{aligned}
\mathcal{J}(\boldsymbol{\theta}) &= -E\Big(\nabla\nabla\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)\Big) \\
&= -E\Big(\frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T\boldsymbol{\theta}))\boldsymbol{x}_t\right]\Big) \\
&= -E\Big(-\frac{1}{\phi}\cdot h'(\boldsymbol{x}_t^T\boldsymbol{\theta})\boldsymbol{x}_t\boldsymbol{x}_t^T\Big) \\
&= \frac{1}{\phi}\Big(h'(\boldsymbol{x}_t^T\boldsymbol{\theta})\boldsymbol{x}_t\boldsymbol{x}_t^T\Big)
\end{aligned}
$$