Greedy Cauchy Prior EM

September 10, 2014

Michael Lindon

1 Previously...

When we spoke we were interested in finding the posterior mode of the model space $p(\gamma|Y)$, which we were going to get at by doing EM on $p(\gamma|Y,\phi)$ which had the form

$$p(\gamma|Y,\phi) \propto \left(\frac{|\Lambda_\gamma|}{|\Lambda_\gamma + X^TX|}\right)^{\frac{1}{2}} e^{\frac{\phi}{2}Y^TX(X^TX+\Lambda_\gamma)^{-1}X^TY} p(\gamma),$$

which comes from $\propto p(Y|\phi,\gamma)p(\gamma)$, where $p(Y|\phi,\gamma) = \int p(Y|\beta,\phi)p(\beta|\phi,\gamma)d\beta$ where we assumed a normal prior with a fixed diagonal covariance matrix on the regression coefficients β which corresponds to a ridge regression type problem with fixed ridge penalty. The unknown parameter ϕ was filled in by EM at each iteration. This is a good thing to start with, but it wouldn't excite any statisticians because people prefer more exotic priors on the regression coefficients that have some good selective shrinkage properties. Most of these have some representation as a scale mixture of normals, so people put a prior on the individual λ's so that the penalty terms can vary. The problem with this larger model is that we would need to do an expectation step to fill in the extra λ parameters of the model- but these are locked up in nasty places so doing the expectation is not possible in closed form.

It's easier to find the posterior mode of $\pi(\beta,\gamma,\phi|Y)$ by doing EM on $\pi(\beta,\gamma,\phi|Y,\lambda)$ as the expectations are a lot easier.

2 New Stuff

Consider the statistical model

$$\begin{aligned}
Y|\beta,\phi &\sim N(1\beta_0 + X\beta, \phi^{-1}) \\
p(\beta_i|\phi,\gamma_i,\lambda_i) &= (1-\gamma_i)\delta_0 + \gamma_i N\left(\beta_i|0,\phi^{-1}\lambda_i\right) \\
\lambda_i &\sim Ga\left(\frac{\alpha}{2},\frac{\alpha}{2}\right) \\
p(\gamma) &= \prod_{i=1}^{p} p^{\gamma_i}(1-p)^{1-\gamma_i} \\
p(\beta_0,\phi) &\propto \frac{1}{\phi}.
\end{aligned}$$

This exploits the scale-mixture representation of a student's t distribution. In particular for $\alpha = 1$ this equates to a prior on the regression coefficients that is a mixture of a point mass at zero with prior probability $1-p$ and a Cauchy centered at zero with prior probability p. By allowing the λ_i to vary we get selective shrinkage of the coefficients and the stats people like the heavy tailed priors so that large coefficients are minimally shrunk toward zero. Cauchy has some very good properties. There are lots of exotic priors but the EM for Cauchy is nice.

We seek to find the mode of $p(\beta,\gamma,\phi|Y)$.
In an EM framework we can treat the λ_i's as missing data, so we can minorize by computing $\mathbb{E}_\lambda[p(\beta,\gamma,\phi|Y,\lambda)|\beta^{(t)},\gamma^{(t)},\phi^{(t)}]$ and then maximize along ϕ and then along (β,γ).

3 Expectation

$p(\lambda_i|\beta_i, \gamma_i, \phi) \propto p(\beta_i|\lambda_i, \gamma_i, \phi)p(\lambda_i)$ so the distribution of our "missing data" $\lambda_i|\beta_i, \gamma_i, \phi \sim Ga\left(\frac{1}{2}(\alpha + \gamma_i), \frac{1}{2}(\alpha + \phi\gamma_i\beta_i^2)\right)$, which has expectation $\mathbb{E}[\lambda_i] = (\alpha + \gamma_i)/(\alpha + \phi\gamma_i\beta_i^2)$. So our objective function is

$$Q(\beta, \gamma, \phi||\beta^{(t)}, \gamma^{(t)}, \phi^{(t)}) = \mathbb{E}\left[p(\beta, \gamma, \phi|Y, \lambda)\right] \propto \frac{n-3}{2}\log\phi - \frac{\phi}{2}||Y - X\beta||^2 + \mathbb{E}\left[\sum \log p(\beta_i|\gamma_i, \lambda_i, \phi)\right] + \sum \log p(\gamma_i),$$

where Y now refers to the centered data (the -3 comes from the prior and also the normalizing constant when we integrate out the incercept). When $\gamma_i = 0$ $p(\beta_i|\gamma_i, \lambda_i, \phi)$ is degenerate and does not have a density. I still haven't found a satisfying notation yet, you just have to consider the different cases and avoid writing down a density in the non full rank degenerate case (which is annoying as the dimension changes).

4 Maximization

For ϕ given $(\beta, \gamma)^{(t)}$

$$\phi^{t+1} = \frac{n + \sum \gamma^{(t)} - 3}{||Y - X\beta^{(t)}|| + \sum \gamma_i^{(t)} \mathbb{E}\left[\lambda_i\right] \beta_i^{(t)2}}$$

For (β, γ) given $\phi^{(t+1)}$ note if $\gamma_i = 0$ then $\beta_i = 0$. For the non-zero γ lets construct submatrices and subvectors indexed by γ, then

$$\begin{aligned} Q(\beta, \gamma, \phi^{(t+1)}||\beta^{(t)}, \gamma^{(t)}, \phi^{(t)}) &\propto \frac{n + \sum \gamma_i - 3}{2}\log\phi^{(t+1)} - \frac{\phi^{t+1}}{2}||Y - X_\gamma\beta_\gamma||^2 - \frac{\phi^{(t+1)}}{2}\beta^T\Lambda_\gamma^{(t)}\beta + \sum \log p(\gamma_i) \\ Q(\beta, \gamma, \phi^{(t+1)}||\beta^{(t)}, \gamma^{(t)}, \phi^{(t)}) &\propto \frac{n + \sum \gamma_i - 3}{2}\log\phi^{(t+1)} - \frac{\phi^{t+1}}{2}||\beta_\gamma - (X_\gamma^T X_\gamma + \Lambda_\gamma)^{-1}X_\gamma^T Y||_{(X_\gamma^T X_\gamma + \Lambda_\gamma)} \\ &\quad - \frac{\phi^{(t+1)}}{2}Y^T(I - X_\gamma(X_\gamma^T X_\gamma + \Lambda_\gamma)^{-1}X_\gamma^T)Y + \sum \log p(\gamma_i), \end{aligned}$$

which is achieves a maximum for a given γ of

$$\sup_\beta Q(\beta, \gamma, \phi^{(t+1)}||\beta^{(t)}, \gamma^{(t)}, \phi^{(t)}) \propto \frac{n + \sum \gamma_i - 3}{2}\log\phi - \frac{\phi^{(t+1)}}{2}Y^T(I - X_\gamma(X_\gamma^T X_\gamma + \Lambda_\gamma)^{-1}X_\gamma^T)Y + \sum \log p(\gamma_i),$$

Basically do the greedy bit for the γ vector during which maximize with respect to β for the current γ