

# Greedy Cauchy Prior EM

September 8, 2014

Michael Lindon

Consider the statistical model

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

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  $\lambda_i$  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  $\lambda_i$ '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  $\phi$  and then along  $(\beta, \gamma)$ .

## 1 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

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

where  $Y$  now refers to the centered data (the  $-3$  comes from the prior and also the normalizing constant when we integrate out the intercept). 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).

## 2 Maximization

For  $\phi$  given  $(\beta, \gamma)^{(t)}$

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

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

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

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

$$- \frac{\phi^{(t+1)}}{2} Y^T (I - X_\gamma (X_\gamma^T X_\gamma + \Lambda_\gamma)^{-1} X_\gamma^T) Y + \sum \log p(\gamma_i),$$

which achieves a maximum for a given  $\gamma$  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  $\gamma$  vector during which maximize with respect to  $\beta$  for the current  $\gamma$