A mixture of normal regressions model

Christopher A. Sims
Princeton University
sims@princeton.edu

February 21, 2018

Motivation

- In our "AK" example, the extreme tails of the distribution of residuals seem to be very slowly decreasing, while over the range of about ± 2 standard deviations, the quantile function is not very different from the normal.
- One hypothesis about why this happens: Some observations are drawn from a different distribution.
- Two versions of this
  - Data errors: typos, misreporting, etc.
  - Some people are different.
- People could be risk-takers, more or less able to take advantage of education, inherently talented without need for education, inherently flawed so they earn little regardless of education, etc.

A model to capture this idea

$$y_i \sim \sum_{j=1}^{k} \pi_j \phi(c_j + X_i \beta_j, \sigma_j^2)\ .$$

I.e., there are k types of observations, each type j occurring with probability π_j, and each satisfying a distinct normal linear model.

An "outlier" model might have $k = 2$, with $\sigma_2^2 \gg \sigma_1^2$ and probably small values in the β_j vector. A "raw talent outliers" model might have $k = 3$, with $c_3 \gg c_2 \gg c_1$, for example.

Gibbs sampling for this model

Iterate not only over the k-dimensional parameters c, π, and σ^2, and over the $k \times m$ dimensional parameter β, but also over an N-dimensional parameter ν_i that is the value of j for observation i.

1. Given ν, estimate a separate normal linear regression for each subsample defined by a given value for ν. Draw c, σ^2, and β from the posteriors of these distinct models.

2. Draw π conditional on ν. (This will have the form of a Dirichlet; see below)

3. Given the regression parameters and π, draw each observation's value of ν from its posterior. (How to do that discussed below.)

Drawing π

The full pdf of y_i given the parameters, including ν, is

$$\prod_i \phi(y_i - c_{\nu_i} - X_i\beta_{\nu_i}; \sigma^2_{\nu_i}) \ .$$

Notice that π does not appear. But we need a prior. An infinite-dimensional parameter like ν always requires a prior, even in large samples, and the prior will matter for inference.

Priors

The regression parameters can be assumed to have conjugate priors, so we will assume they are incorporated by dummy observations. (The ν value for these dummy observations has to be held fixed.)

Conditional on π, the probability of $\nu_i = j$ is π_j, while the prior pdf of the π vector itself is Dirichlet, say $\prod_j \pi_j^{\alpha-1}$.

Together, this gives us the additional factor in the posterior kernel

$$\prod_j \pi_j^{\alpha-1+n_j} \,,$$

where n_j is the number of observations with $\nu_i = j$. So drawing from the posterior on π is just drawing from a Dirichlet, and will give something very close to the sample frequencies of ν_i values in a sample with many observations for each j.

Drawing ν

A particular observation i's ν_i value affects the posterior only via the likelihood for observation i and (via its effect on n_j) the joint π, ν prior. Conditional density of ν_i values over $j = 1, \ldots k$ is proportional to

$$\pi_j \phi(y_i - c_j - X_i \beta_j; \sigma_j^2) \, .$$

This is easily calculated and defines, when normalized to sum to one, a multinomial distribution over j, which is easy to draw from.

Big data complications

- With over 300,000 observations and a fairly big regression model, these MCMC iterations could be time-consuming.

- Initially, you will want to just maximize posterior density, and this should probably begin by maximizing over a subsample, say 1000 or 3000 observations. Since this is mainly just a starting point for MCMC, there might be no need to use the full sample for the optimization.

- You might do MCMC over a subsample also, at least to start. This can give you an idea of a reasonable range of starting points for full-sample MCMC chains.

Mixture model complications

- Mixture models can produce weird likelihoods and bad MCMC behavior if the model and prior imply that exact prediction might be possible for some j. It can fit perfectly a small group of observations and make likelihood go to infinity. So a prior density on σ_j^2 that is near zero near $\sigma_j^2 = 0$ is important.

- Mixture models always have a permutation normalization issue. With a $k = 2$ model where one of the models is for outliers, it should be enough to insist on $\sigma_2^2 > \sigma_1^2$.

Imposing normalization during MCMC

- In Gibbs sampling, this requires discarding the complete set of draws for the σ_j^2's (since they are dependent when the ordering is imposed) whenever the ordering is violated.

- With Metropolis-Hastings MCMC, the previous draw is repeated whenever the proposed draw violates the ordering. That is, the proposed draw is subject to the usual accept/reject rule, with draws violating the ordering treated as having 0 posterior density).

Instead of normalization during MCMC sampling

- One can sample without normalizing, then when sampling is complete, map all the draws into their normalized counterparts.

- Of course one could also do this mapping after each draw j, mapping draw $j-1$ to its unnormalized counterpart. This is equivalent to remapping all draws after sampling is finished, since the chain is Markov.

- A non-symmetric prior, favoring, e.g., $\sigma_2^2 > \sigma_1^2$ might be enough to keep all, or nearly all, draws in the favored region, making remapping unnecessary. However, if much remapping is needed, the implications of such a prior for prior beliefs about the model under remapping may be obscure.