FINITE MIXTURE MODELS

AAEC 6564
INSTRUCTOR: KLAUS MOELTNER

Textbooks: Koop (2003), Ch.3; Koop et al. (2007), Ch.15
`Matlab` scripts: `mod10_2CMM_data, mod10_2CMMln, mod10_2CMMchi2, mod10_2CMMmix, mod10_2CMMplots, mod10_2CMM_rearrange, mod10_fmrm_data, mod10_fmrm, mod10_fmrm_data_v2, mod10_fmrm_v2`
`Matlab` functions: `gs_2cmm, gs_fmrm`

Model Outline

This setup is similar to the regime switching model (RSM, aka "Roy model") discussed previously, in that each observation may follow one of G different regression models. Each regression model ("regime") has its own set of coefficients (and potentially its own set of explanatory variables), and its own error variance. However, in contrast to the RSM, regime assignment is not observed and must be estimated probabilistically. That is, in addition to estimating G sets of coefficients and variances, we need to estimated G regime probabilities.

Let these regime probabilities be denoted as π_g, $g = 1 \dots G$, with $\sum_{g=1}^{G} \pi_g = 1$. We will collect these probabilities in vector $\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \pi_2 & \dots \pi_G \end{bmatrix}'$.

The likelihood function for observation i, *unconditional on regime membership* can then be expressed as a mixture-of-normals, that is:

$$\begin{aligned} &p\left(y_i|\mathbf{x}_i, \left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^{G}, \boldsymbol{\pi}\right) = \sum_{g=1}^{G} \pi_g \left((2\pi)^{-1/2} \left(\sigma_g^2\right)^{-1/2} exp\left(-\tfrac{1}{2\sigma_g^2} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g\right)^2 \right) \right) \quad \text{with} \\ &\pi_g \in [0,1] \; \forall g, \quad \sum_{g=1}^{G} \pi_g = 1 \end{aligned} \tag{1}$$

The product of this over all i produces the likelihood for the entire sample. Computation is facilitated by augmenting the model with *component indicator vector* $\mathbf{z}_i$ for each i, given as

$$\mathbf{z}_i = \begin{bmatrix} z_{1i} & z_{2i} & \dots & z_{Gi} \end{bmatrix}', \tag{2}$$

where $z_{gi} = 1$ if i is in the g^{th} regime, and zero otherwise. We can then express the kernel of the *augmented*, that is *membership-conditioned* likelihood compactly as

$$\begin{aligned} &p\left(\mathbf{y}|\mathbf{X}, \left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^{G}, \left\{\mathbf{z}_i\right\}_{i=1}^{n}\right) \propto \\ &\prod_{i=1}^{n}\left[\sum_{g=1}^{G}\left\{\left(\sigma_g^2\right)^{-1/2} exp\left(-\tfrac{1}{2\sigma_g^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g\right)^2\right)\right\} I\left(z_{gi}=1\right)\right] \end{aligned} \tag{3}$$

where $I(.)$ is the usual indicator function that takes a value of one if the condition it describes holds, and a value of zero otherwise.

The hierarchical prior of $\mathbf{z}_i$, given $\boldsymbol{\pi}$, is modeled as a multinomial distribution with $T = 1$ trial, that is $\mathbf{z} \sim mn(1, \boldsymbol{\pi})$, explicitly given as:

$$\begin{aligned} &p\left(\mathbf{z}_i|\boldsymbol{\pi}\right) = \left(\frac{T}{z_{1i}!z_{2i}!\ldots z_{Gi}!}\right)\pi_1^{z_{1i}}\pi_2^{z_{2i}}\ldots\pi_G^{z_{Gi}} = \prod_{g=1}^{G}\pi_g^{z_{gi}} \quad \text{with} \\ &\sum_{g=1}^{G} z_{gi} = T = 1 \end{aligned} \tag{4}$$

Regime probability vector $\boldsymbol{\pi}$, in turn, receives a Dirichlet prior with parameter vector $\boldsymbol{\alpha}$, that is $\boldsymbol{\pi} \sim D(\boldsymbol{\alpha})$, explicitly given as:

$$\begin{aligned} &p\left(\boldsymbol{\pi}\right) = \left(\frac{\Gamma\left(\boldsymbol{\alpha}\right)}{\prod_{g=1}^{G}\Gamma\left(\alpha_g\right)}\right)\prod_{g=1}^{G}\pi_g^{\alpha_g - 1}, \quad \text{where} \\ &\boldsymbol{\alpha} = \sum_{g=1}^{G}\alpha_g, \quad \alpha_g > 0,\ \forall g \end{aligned} \tag{5}$$

Note that for $G = 2$ the Dirichlet reduces to the univariate Beta distribution $B(\alpha_1, \alpha_2)$:

$$\begin{aligned} &p\left(\pi\right) = \left(\frac{\Gamma\left(\alpha_1 + \alpha_2\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_2\right)}\right)\pi^{\alpha_1 - 1}\left(1 - \pi\right)^{\alpha_2 - 1}, \quad \text{where} \\ &\alpha_g > 0,\ g = 1, 2 \end{aligned} \tag{6}$$

Combining all these parts, plus the usual multivariate normal prior for the $\boldsymbol{\beta}_g$'s with (for simplicity) common prior mean $\boldsymbol{\mu}_0$ and common prior variance-covariance matrix $\mathbf{V}_0$, and the standard inverse-gamma prior for the σ_g^2's with (for simplicity) common shape ν_0 and common scale τ_0 leads to the

following augmented posterior kernel:

$$
\begin{aligned}
& p\left(\left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^G, \boldsymbol{\pi}, \{\mathbf{z}_i\}_{i=1}^n | \mathbf{y}, \mathbf{X}\right) \propto \\
& \prod_{g=1}^G exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_g - \boldsymbol{\mu}_0\right)' \mathbf{V}_0^{-1}\left(\boldsymbol{\beta}_g - \boldsymbol{\mu}_0\right)\right) * \\
& \prod_{g=1}^G \left(\sigma_g^2\right)^{(-\nu_0+1)} exp\left(-\frac{\tau_0}{\sigma_g^2}\right) * \\
& \prod_{g=1}^G \pi_g^{\alpha_g - 1} * \\
& \prod_{i=1}^n \left(\prod_{g=1}^G \pi_g^{z_{gi}}\right) * \\
& \prod_{i=1}^n \left(\sum_{g=1}^G \left\{\left(\sigma_g^2\right)^{-1/2} exp\left(-\tfrac{1}{2\sigma_g^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g\right)^2\right)\right\} I\left(z_{gi} = 1\right)\right)
\end{aligned}
\tag{7}
$$

In each iteration of the Gibbs Sampler every observation i will be assigned to one of the G regimes. This determines the current sample size for regime G, say n_g, and the current rows of the data associated with regime g, let's call those $\mathbf{y}_g$ and $\mathbf{X}_g$, respectively. Then, for each regime g the corresponding vector of regression coefficients. $\boldsymbol{\beta}_g$ can be drawn as usual, that is:

$$
\begin{aligned}
& \boldsymbol{\beta}_g | \sigma_g^2, \mathbf{y}_g, \mathbf{X}_g \sim n\left(\boldsymbol{\mu}_{g1}, \mathbf{V}_{g1}\right), \quad \text{with} \\
& \mathbf{V}_{g1} = \left(\mathbf{V}_0^{-1} + \tfrac{1}{\sigma_g^2}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}, \quad \text{and} \\
& \boldsymbol{\mu}_{g1} = \mathbf{V}_{g1}\left(\mathbf{V}_0^{-1}\boldsymbol{\mu}_0 + \tfrac{1}{\sigma_g^2}\mathbf{X}_g'\mathbf{y}_g\right)
\end{aligned}
\tag{8}
$$

Note, that, implicitly, $\boldsymbol{\beta}_g$ is drawn conditional on $\{\mathbf{z}_i\}_{i=1}^n$, since the latter determines the contents of data $\mathbf{y}_g$ and $\mathbf{X}_g$. Similarly, for the regime-specific variances we obtain:

$$
\begin{aligned}
& \sigma_g^2 | \boldsymbol{\beta}_g, \mathbf{y}_g, \mathbf{X}_g \sim ig\left(\nu_{g1}, \tau_{g1}\right), \quad \text{with} \\
& \nu_{g1} = \frac{2\nu_0 + n_g}{2}, \quad \text{and} \\
& \tau_{g1} = \tau_0 + \tfrac{1}{2}\left(\mathbf{y}_g - \mathbf{X}\boldsymbol{\beta}_g\right)'\left(\mathbf{y}_g - \mathbf{X}\boldsymbol{\beta}_g\right)
\end{aligned}
\tag{9}
$$

The relevant posterior kernel for draws of regime probabilities $\boldsymbol{\pi}$ is given as

$$p\left(\boldsymbol{\pi}|\left\{\mathbf{z}_i\right\}_{i=1}^n\right) \propto \prod_{g=1}^G \pi_g^{\alpha_g - 1} \prod_{i=1}^n \left(\prod_{g=1}^G \pi_g^{z_{gi}}\right) = \prod_{g=1}^G \pi_g^{(\alpha_g + n_g - 1)} \tag{10}$$

This points again to a Dirichlet distribution with parameters $\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_G + n_G$, i.e.

$$\boldsymbol{\pi}|\left\{\mathbf{z}_i\right\}_{i=1}^n \sim D\left(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_G + n_G\right) \tag{11}$$

As a final step, we need to draw new assignment vectors $\mathbf{z}_i$ for each observation at each step of the Gibbs Sampler. The relevant posterior kernel for a given $\mathbf{z}_i$ is given by:

$$\begin{aligned} &p\left(\mathbf{z}_i|\left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^G, \boldsymbol{\pi}, \mathbf{y}_i, \mathbf{X}_i\right) \propto \\ &\prod_{g=1}^G \pi_g^{z_{gi}} * \sum_{g=1}^G \left\{\left(\sigma_g^2\right)^{-1/2} exp\left(-\tfrac{1}{2\sigma_g^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g\right)^2\right)\right\} I\left(z_{gi} = 1\right) \end{aligned} \tag{12}$$

Bringing back in the normalizing constant for the normal density in the second line of (12), and replacing the indicator function with an exponent, this can be equivalently written as

$$\begin{aligned} &p\left(\mathbf{z}_i|\left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^G, \boldsymbol{\pi}, \mathbf{y}_i, \mathbf{X}_i\right) \propto \\ &\prod_{g=1}^G \pi_g^{z_{gi}} * \phi\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}_g, \sigma_g^2\right)^{z_{gi}} = \\ &\prod_{g=1}^G \left(\pi_g \phi\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}_g, \sigma_g^2\right)\right)^{z_{gi}}, \end{aligned} \tag{13}$$

where $\phi\left(.\right)$ denotes the normal density. This can be recognized as the kernel of another *multinomial* density with a single trial:

$$\begin{aligned} &\mathbf{z}_i|\left\{\boldsymbol{\beta}_g, \sigma_g^2\right\}_{g=1}^G, \boldsymbol{\pi}, \mathbf{y}_i, \mathbf{X}_i \sim \\ &mn\left(1, \frac{\pi_1\phi\left(1\right)}{\sum_{g=1}^G\left(\pi_g\phi\left(g\right)\right)}, \frac{\pi_2\phi\left(2\right)}{\sum_{g=1}^G\left(\pi_g\phi\left(g\right)\right)}, \ldots, \frac{\pi_G\phi\left(G\right)}{\sum_{g=1}^G\left(\pi_g\phi\left(g\right)\right)}\right), \quad \text{where} \\ &\phi\left(g\right) = \phi\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}_g, \sigma_g^2\right) \end{aligned} \tag{14}$$

Using mixtures-of-normals to approximate unknown densities

As shown in exercise 15.5 in Koop et al. (2007), finite mixture models are quite robust to mis-specifications of the underlying likelihood. In their example, the true data-generating mechanism varies from a log-normal, to a chi-squared, to a two regime ("component") mixture-of-normals (2CMM). In each case, the same 2CMM is used to estimate the model. Comparing the posterior predictive distribution flowing from the model to the true underlying data generating density shows that the 2CMM does a good job recovering the shape of the data, even under mis-specification.

Matlab scripts `mod10_2CMM...` and function `gs_2cmm` replicate KPT's results.

Note: To draw n 1 by G vectors from the multinomial with T trials and 1 by G probability vector $\mathbf{p}$, use Matlab's built-in function `mnrnd(T,p,n)`. If $\mathbf{p}$ is already a n by G matrix, as in our case, use `mnrnd(T,p)`.

To obtain n columns of draws from the Dirichlet with G by 1 parameter vector $\mathbf{a}$ use my own function `dirrnd(a,n)`.

The label-switching issue

As discussed in Geweke (2007) the mixture model as outlined above is *invariant to a permutation of the regime labels*, that is the likelihood function yields the same value irrespective of the order in which we add up the G components, that is the terms $\pi_g \phi\left(y_i, \mathbf{x}_i'\boldsymbol{\beta}_g, \sigma_g^2\right)$. As a result, the posterior simulator (our Gibbs Sampler) "doesn't care" either which regime is "1", "2," etc. So in one iteration the first set of coefficients (which are supposed to be from regime 1) may actually belong to ("true") regime 1, while in the next iteration the first set might actually belong to ("true") regime 2, and so on. The same dilemma holds for regime-specific variances.

In consequence, drawing inference from all "set 1" coefficients produced by the Gibbs Sampler may be meaningless, as it might be based on a mix of coefficients from all regimes. As mentioned in Geweke (2007), if there is a natural ordering in the magnitude of (true) parameters across regimes, as in $\boldsymbol{\beta}_1 < \boldsymbol{\beta}_2$ for all elements of $\boldsymbol{\beta}$, or $\sigma_1^2 < \sigma_2^2$, then, asymptotically this labeling dilemma vanishes, but it may take a very large sample for this to work.

As you can see from the preceding exercise, for our large sample of 10,000 the algorithm de facto switched the regime means, variances, and probabilities compared to the way they were labeled in the true data-generating mechanism. To be specific, when we created the data, we let $\pi_1 = 0.6$ and $\pi_2 = 0.4$, but the Matlab log indicates the opposite. The same switching occurred for the regime means and variances. But at least *within* each regime, we recover the correct mean and variance. Thus, for this example, the label switch occurred immediately, and there was little or no switching back (else our naive posterior means would be way off).

According to Geweke (2007) this entire regime-switching dilemma can be completely ignored if the final construct of interest is invariant to this switching, for example a posterior predictive outcome. In fact, when the mixture model is primarily used as a flexible approximation of an unknown density function, as in the example above, the regimes are by definition devoid of any substantive interpretation, so this label-switching can be ignored. This becomes apparent from our plotted predictive densities, which map out the true underlying sampling density just fine, regardless of

regime labeling.

As an (important) aside, the authors also note that if the state labels have no substantive interpretation, the priors should also be permutation-invariant (which is trivially satisfied if all regimes receive the same prior, as is the case in the above example).

If labeling matters (for example, we hypothesize that the marginal effect of some variable should systematically differ across regimes), and if there is some pre-known natural ordering in one or more of the regime-specific parameters as illustrated above, Geweke (2007) show that this can be simply addressed by running the Gibbs sampler as usual (ignoring any labeling), and then re-labeling the parameter draws *ex-post* to honor the inequality constraints. In our example, for example, we stipulate that $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.5$, so we could, for each draw of the Gibbs Sampler, re-arrange all parameters (mean, variance, probability) that correspond to the smaller variance draw to belong to the *first* regime.

This is shown in script `mod10_2CMMmix_rearrange`.

A full mixtures-of-normals regression model

The previous example was a simplified version of a mixture model compared to our theoretical derivation, as it only included a single mean and variance per regime. As a next step, we will allow the regime means to be linear functions of data and parameters, which corresponds more directly to the structural model above.

Matlab scripts `mod10_fmrm_data` generates simulated data for such a model, with three regimes. To facilitate regime assignments (and thus reduce the labeling problem) we allow the true coefficients in each regime to be increasing in magnitude over regimes, while keeping the regime variances small and equal. Script `mod10_fmrm` implements the sampler. As you can see from the log file, the algorithm generally does a good job recovering the true parameters, though efficiency is less than desirable for several cases. This "slow mixing" problem is a well-known issue for finite mixture model(Frühwirth-Schnatter, 2001; Geweke, 2007).

We can improve things by forcing the bin coefficients even further apart while keeping variances small, as implemented in `mod10_fmrm_data_v2` and `mod10_fmrm_v2`. As is evident from the log file, this sampler has much better mixing properties than the original version. Naturally, in a real world application, we may not be able to make a-priori assumptions on the relative magnitude of regime-specific parameters and/or regime-specific parameters may be similar in magnitude. In such cases the label-switching problem may be a real barrier to inference, other than for predictive constructs.

References

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models, *Journal of the American Statistical Association* **96**: 194–209.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works, *Computational Statistics & Data Analysis* **51**: 3529–3550.

Koop, G. (2003). *Bayesian Econometrics*, Wiley.

Koop, G., Poirier, D. and Tobias, J. (2007). *Bayesian Econometric Methods*, Cambridge University Press.