Mixture models

Missing variable models

Complexity of a model may originate from the fact that some piece of information is *missing*

Example

Arnason–Schwarz model with missing zones
Probit model with missing normal variate

Generic representation

$$f(\mathbf{x}|\theta) = \int_{\mathscr{Z}} g(\mathbf{x}, \mathbf{z}|\theta)\, \mathrm{d}\mathbf{z}$$

Mixture models

Models of *mixtures of distributions*:

$$x \sim f_j \text{ with probability } p_j,$$

for $j = 1, 2, \ldots, k$, with overall density

$$p_1 f_1(x) + \cdots + p_k f_k(x) \ .$$

Usual case: parameterised components

$$\sum_{i=1}^{k} p_i f(x|\theta_i)$$

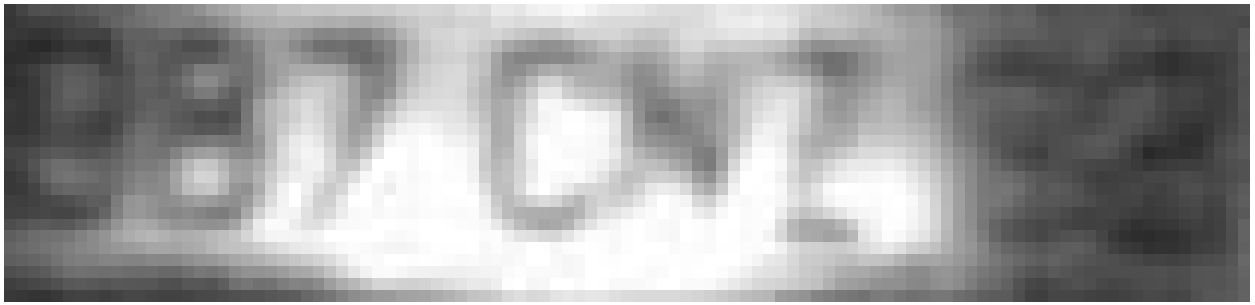
where *weights* p_i's are distinguished from other parameters

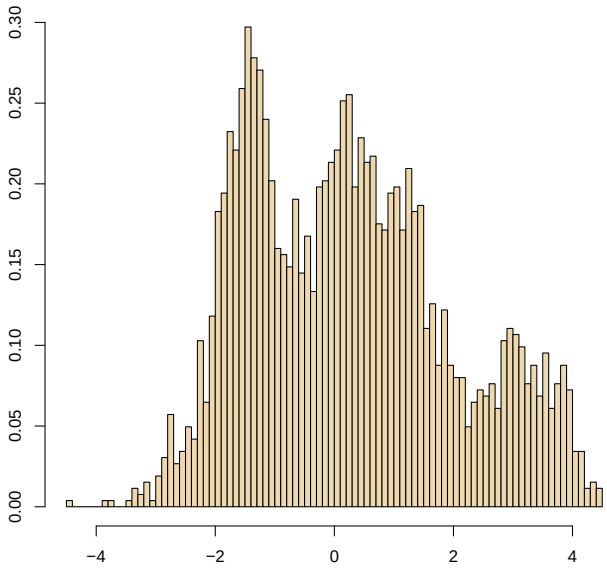
Motivations

- Dataset made of several latent/missing/unobserved strata/subpopulations. Mixture structure due to the missing origin/allocation of each observation to a specific subpopulation/stratum. Inference on either the allocations (clustering) or on the parameters (θ_i, p_i) or on the number of groups
- Semiparametric perspective where mixtures are basis approximations of unknown distributions

License

Dataset derived from license plate image
Grey levels concentrated on 256 values later jittered





Likelihood

For a sample of independent random variables $(x_1, \cdots, x_n)$, likelihood

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\} \ .$$

Expanding this product involves

$$k^n$$

elementary terms: prohibitive to compute in large samples. But likelihood still computable [pointwise] in $\mathrm{O}(kn)$ time.

Normal mean benchmark

Normal mixture

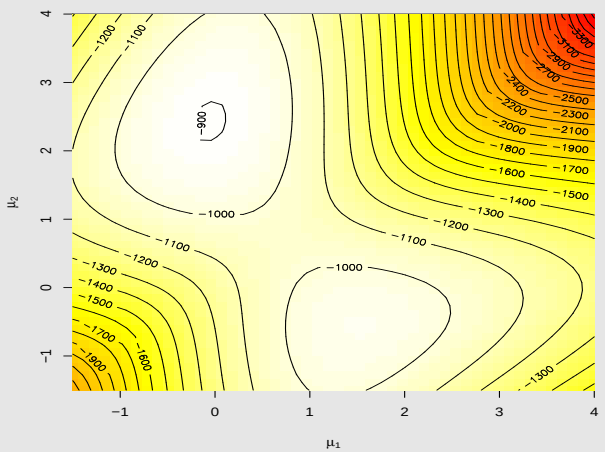
$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1)$$

with only unknown means (2-D representation possible)

Identifiability

Parameters μ_1 and μ_2 identifiable: μ_1 cannot be confused with μ_2 when p is different from 0.5.

Presence of a spurious mode, understood by letting p go to 0.5



Bayesian Inference

For any prior $\pi(\boldsymbol{\theta}, \mathbf{p})$, posterior distribution of $(\boldsymbol{\theta}, \mathbf{p})$ available up to a multiplicative constant

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) \propto \left[\prod_{i=1}^{n} \sum_{j=1}^{k} p_j \, f(x_i|\theta_j)\right] \pi(\boldsymbol{\theta}, \mathbf{p}) \, .$$

at a cost of order $\mathrm{O}(kn)$

Difficulty

Despite this, derivation of posterior characteristics like posterior expectations only possible in an exponential time of order $\mathrm{O}(k^n)$!

Missing variable representation

Associate to each x_i a missing/latent variable z_i that indicates its component:

$$z_i|\mathbf{p} \sim \mathscr{M}_k(p_1, \ldots, p_k)$$

and

$$x_i|z_i, \boldsymbol{\theta} \sim f(\cdot|\theta_{z_i})\,.$$

Completed likelihood

$$\ell(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} p_{z_i}\, f(x_i|\theta_{z_i})\,,$$

and

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) \propto \left[\prod_{i=1}^{n} p_{z_i}\, f(x_i|\theta_{z_i})\right] \pi\left(\boldsymbol{\theta}, \mathbf{p}\right)\,,$$

where $\mathbf{z} = (z_1, \ldots, z_n)$.

Partition sets

Denote by $\mathcal{Z} = \{1, \ldots, k\}^n$ set of the k^n possible vectors $\mathbf{z}$.
$\mathcal{Z}$ decomposed into a partition of sets

$$\mathcal{Z} = \cup_{j=1}^{\mathfrak{r}} \mathcal{Z}_j$$

For a given allocation size vector $(n_1, \ldots, n_k)$, where $n_1 + \ldots + n_k = n$, *partition sets*

$$\mathcal{Z}_j = \left\{ \mathbf{z} : \sum_{i=1}^{n} \mathbb{I}_{z_i=1} = n_1, \ldots, \sum_{i=1}^{n} \mathbb{I}_{z_i=k} = n_k \right\},$$

for all allocations with the given allocation size $(n_1, \ldots, n_k)$ and where labels $j = j(n_1, \ldots, n_k)$ defined by lexicographical ordering on the $(n_1, \ldots, n_k)$'s.

Posterior closed form representations

$$\pi\left(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}\right) = \sum_{i=1}^{\mathfrak{r}} \sum_{\mathbf{z}\in\mathcal{Z}_i} \omega\left(\mathbf{z}\right) \pi\left(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}\right) \ ,$$

where $\omega\left(\mathbf{z}\right)$ represents marginal posterior probability of the allocation $\mathbf{z}$ conditional on $\mathbf{x}$ *[derived by integrating out the parameters* $\boldsymbol{\theta}$ *and* $\mathbf{p}$*]*

Bayes estimator of $(\boldsymbol{\theta}, \mathbf{p})$

$$\sum_{i=1}^{\mathfrak{r}} \sum_{\mathbf{z}\in\mathcal{Z}_i} \omega\left(\mathbf{z}\right) \mathbb{E}^{\pi}\left[\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}\right] \ .$$

© Too costly: 2^n terms

General Gibbs sampling for mixture models

Take advantage of the missing data structure:

Algorithm

- **Initialization:** choose $\mathbf{p}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ arbitrarily
- **Step t.** For $t = 1, \ldots$
  1. Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, \ldots, k)$
  $$\mathbb{P}\left(z_i^{(t)} = j | p_j^{(t-1)}, \theta_j^{(t-1)}, x_i\right) \propto p_j^{(t-1)} f\left(x_i | \theta_j^{(t-1)}\right)$$
  2. Generate $\mathbf{p}^{(t)}$ from $\pi(\mathbf{p}|\mathbf{z}^{(t)})$,
  3. Generate $\boldsymbol{\theta}^{(t)}$ from $\pi(\boldsymbol{\theta}|\mathbf{z}^{(t)}, \mathbf{x})$.

Exponential families

When

$$f(x|\theta) = h(x)\exp(R(\theta)\cdot T(x) > -\psi(\theta))$$

simulation of both $\mathbf{p}$ and $\boldsymbol{\theta}$ usually straightforward:
Conjugate prior on θ_j given byBack to definition

$$\pi_j(\theta) \propto \exp(R(\theta)\cdot\alpha_j - \beta_j\psi(\theta))\,,$$

where $\alpha_j \in \mathbb{R}^k$ and $\beta_j > 0$ are hyperparameters and

$$\mathbf{p} \sim \mathscr{D}(\gamma_1,\ldots,\gamma_k)$$

[Dirichlet distribution]

Gibbs sampling for exponential family mixtures

Algorithm

- **Initialization.** Choose $\mathbf{p}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$,
- **Step t.** For $t = 1, \ldots$
  1. Generate $z_i^{(t)}$ $(i = 1, \ldots, n,\ j = 1, \ldots, k)$ from

$$\mathbb{P}\left(z_i^{(t)} = j | p_j^{(t-1)}, \theta_j^{(t-1)}, x_i\right) \propto p_j^{(t-1)} f\left(x_i | \theta_j^{(t-1)}\right)$$

  2. Compute $n_j^{(t)} = \sum_{i=1}^n \mathbb{I}_{z_i^{(t)}=j}$, $s_j^{(t)} = \sum_{i=1}^n \mathbb{I}_{z_i^{(t)}=j} t(x_i)$
  3. Generate $\mathbf{p}^{(t)}$ from $\mathscr{D}\left(\gamma_1 + n_1, \ldots, \gamma_k + n_k\right)$,
  4. Generate $\theta_j^{(t)}$ $(j = 1, \ldots, k)$ from

$$\pi(\theta_j | \mathbf{z}^{(t)}, \mathbf{x}) \propto \exp\left(R(\theta_j) \cdot (\alpha + s_j^{(t)}) - \psi(\theta_j)(n_j + \beta)\right).$$

Normal mean example

For mixture of two normal distributions with unknown means,

$$p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\theta, \sigma^2) \ ,$$

and a normal prior $\mathscr{N}(\delta, 1/\lambda)$ on μ_1 and μ_2,

Normal mean example (cont'd)

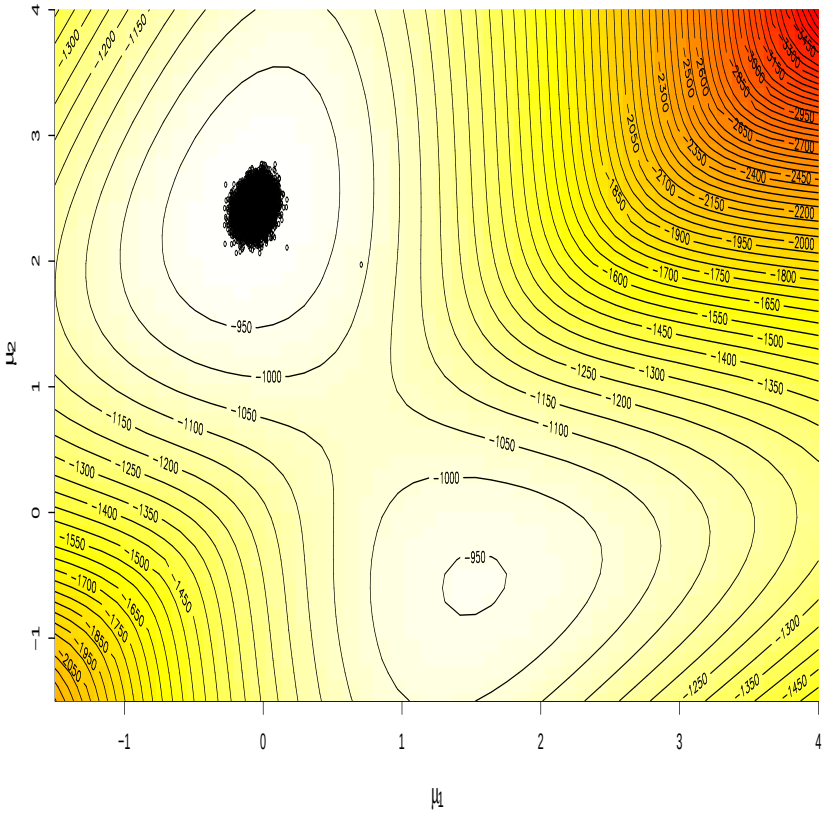
Algorithm

- **Initialization.** Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$,
- **Step t.** For $t = 1, \ldots$
  1. Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from

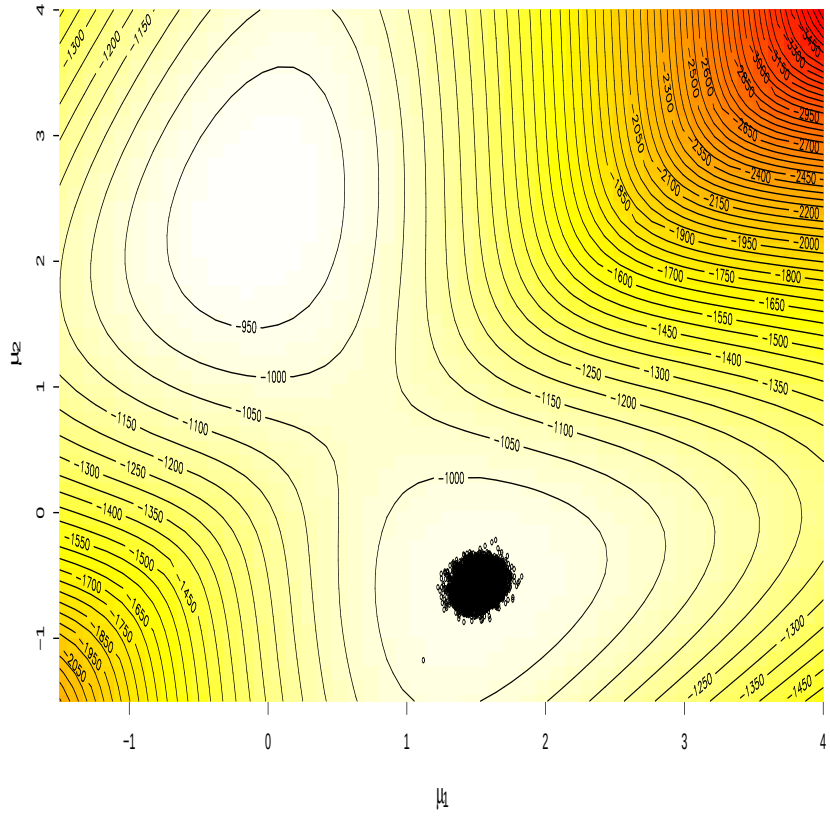
$$\mathbb{P}\left(z_i^{(t)} = 1\right) = 1 - \mathbb{P}\left(z_i^{(t)} = 2\right) \propto p \exp\left(-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right)$$

  2. Compute $n_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j}$ and $(s_j^x)^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j} x_i$
  3. Generate $\mu_j^{(t)}$ $(j = 1, 2)$ from $\mathcal{N}\left(\frac{\lambda\delta + (s_j^x)^{(t)}}{\lambda + n_j^{(t)}}, \frac{1}{\lambda + n_j^{(t)}}\right)$.

Normal mean example (cont'd)



(a) initialised at random

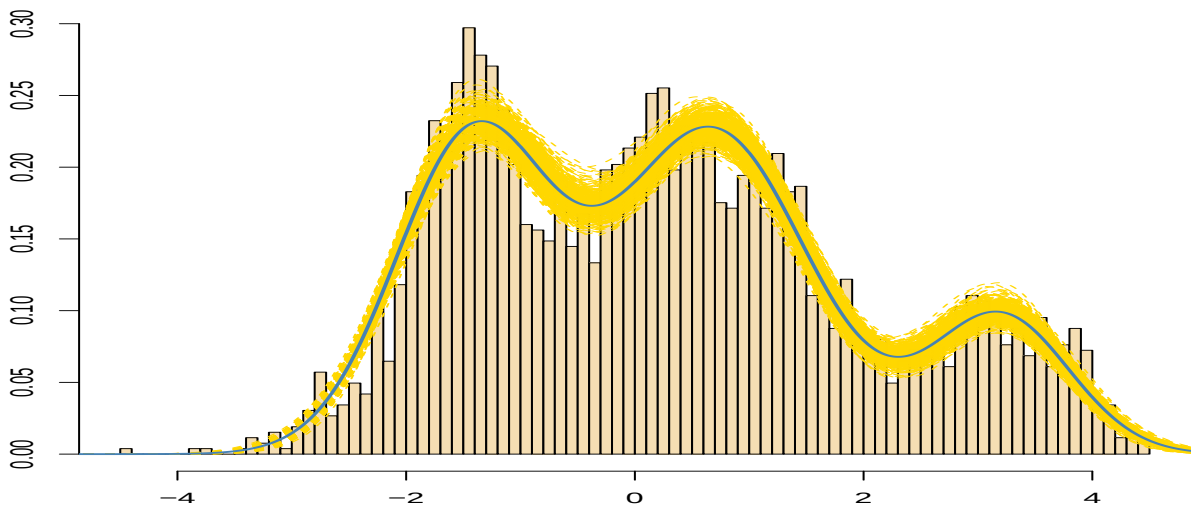


(b) initialised close to the lower mode

License

Consider $k = 3$ components, a $\mathscr{D}_3(1/2, 1/2, 1/2)$ prior for the weights, a $\mathscr{N}(\overline{x}, \hat{\sigma}^2/3)$ prior on the means μ_i and a $\mathscr{G}a(10, \hat{\sigma}^2)$ prior on the precisions σ_i^{-2}, where $\overline{x}$ and $\hat{\sigma}^2$ are the empirical mean and variance of License

[Empirical Bayes]



Metropolis–Hastings alternative

For the Gibbs sampler, completion of $\mathbf{z}$ increases the dimension of the simulation space and reduces the mobility of the parameter chain.

Metropolis–Hastings algorithm available since posterior available in closed form, as long as q provides a correct exploration of the posterior surface, since

$$\frac{\pi(\boldsymbol{\theta}', \mathbf{p}'|\mathbf{x})}{\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})} \frac{q(\boldsymbol{\theta}, \mathbf{p}|\boldsymbol{\theta}', \mathbf{p}')}{q(\boldsymbol{\theta}', \mathbf{p}'|\boldsymbol{\theta}, \mathbf{p})} \wedge 1$$

computable in $\mathbf{O}(kn)$ time

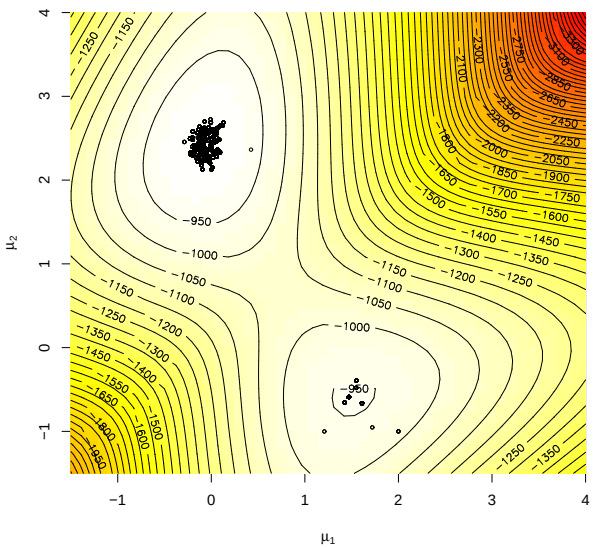
Random walk Metropolis–Hastings

Proposal distribution for the new value

$$\widetilde{\theta}_j = \theta_j^{(t-1)} + u_j \text{ where } u_j \sim \mathcal{N}(0, \zeta^2)$$

In mean mixture case, Gaussian random walk proposal is

$$\widetilde{\mu_1} \sim \mathcal{N}\left(\mu_1^{(t-1)}, \zeta^2\right) \quad \text{and}$$
$$\widetilde{\mu_2} \sim \mathcal{N}\left(\mu_2^{(t-1)}, \zeta^2\right)$$



Random walk Metropolis–Hastings for means

Algorithm

- Initialization:
  Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$
- Iteration t $(t \geq 1)$:
  1. Generate $\widetilde{\mu_1}$ from $\mathcal{N}\left(\mu_1^{(t-1)}, \zeta^2\right)$,
  2. Generate $\widetilde{\mu_2}$ from $\mathcal{N}\left(\mu_2^{(t-1)}, \zeta^2\right)$,
  3. Compute

$$r = \pi\left(\widetilde{\mu_1}, \widetilde{\mu_2}|x\right) / \pi\left(\mu_1^{(t-1)}, \mu_2^{(t-1)}|x\right)$$

  4. Generate $u \sim \mathcal{U}_{[0,1]}$: if $u < r$, then $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = (\widetilde{\mu_1}, \widetilde{\mu_2})$
     else $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = \left(\mu_1^{(t-1)}, \mu_2^{(t-1)}\right)$.

Random walk extensions

Difficulties with **constrained parameters**, like $\mathbf{p}$ such that $\sum_{i=1}^{k} p_k = 1$.

Resolution by overparameterisation

$$p_j = w_j \Big/ \sum_{l=1}^{k} w_l \,, \quad w_j > 0 \,,$$

and proposed move on the w_j's

$$\log(\widetilde{w_j}) = \log(w_j^{(t-1)}) + u_j \text{ where } u_j \sim \mathcal{N}(0, \zeta^2)$$

Watch out for the Jacobian in the $\log$ transform

Identifiability

A mixture model is invariant under permutations of the indices of the components.
E.g., mixtures

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.3,1)$$

and

$$0.7\mathcal{N}(2.3,1) + 0.3\mathcal{N}(0,1)$$

are **exactly** the same!

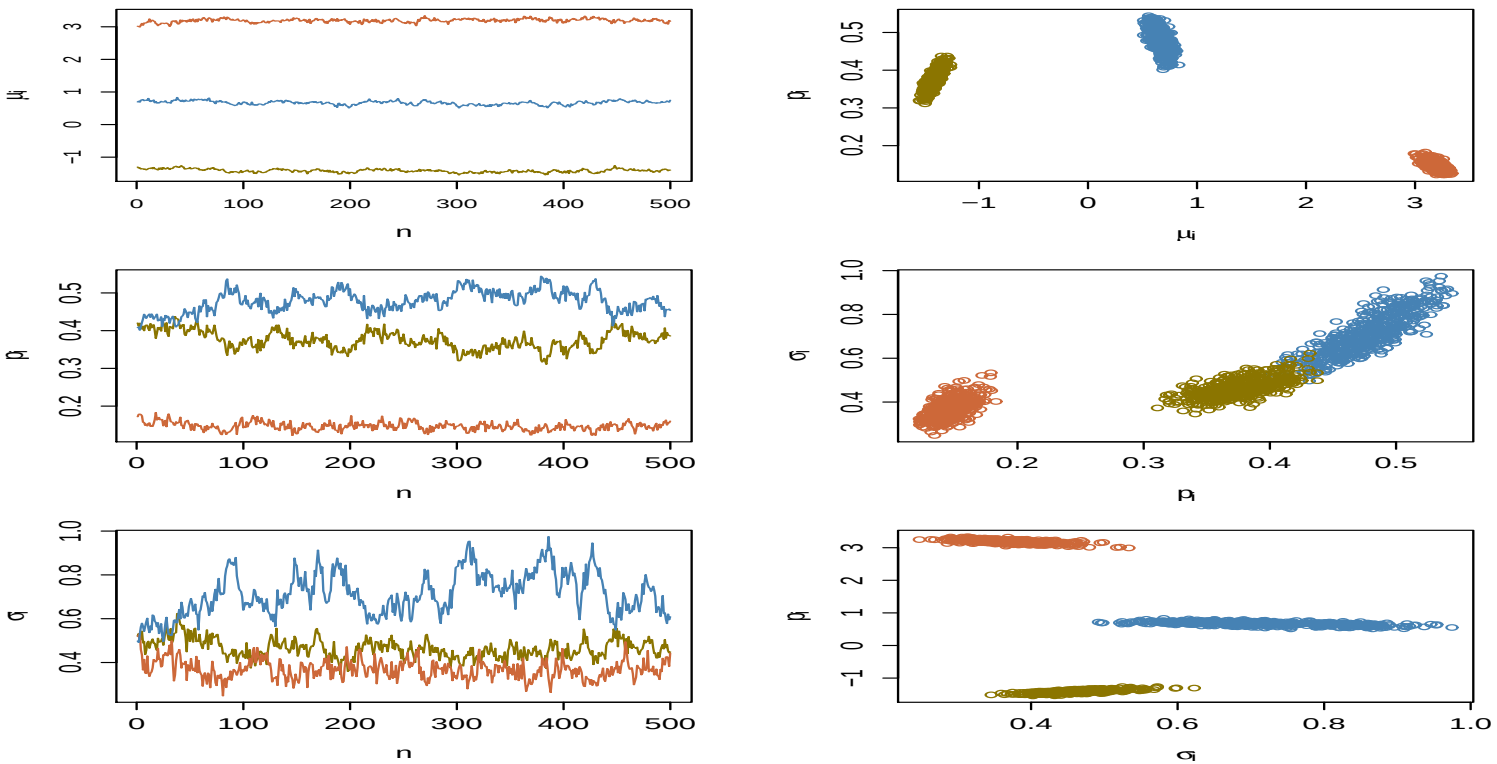
© **The component parameters θ_i are not identifiable marginally since they are exchangeable**

Connected difficulties

1. Number of modes of the likelihood of order $\mathrm{O}(k!)$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\boldsymbol{\theta}, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior expectation of θ_1 equal to posterior expectation of θ_2.

License

Since Gibbs output does not produce exchangeability, the Gibbs sampler has not explored the whole parameter space: it lacks energy to switch simultaneously enough component allocations at once



Label switching paradox

We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.

If we observe it, then we do not know how to estimate the parameters.

If we do not, then we are uncertain about the convergence!!!

Constraints

Usual reply to lack of identifiability: impose constraints like $\mu_1 \leq \ldots \leq \mu_k$ in the prior

Mostly incompatible with the topology of the posterior surface: posterior expectations then depend on the choice of the constraints.

Computational detail

The constraint does not need to be imposed *during* the simulation but can instead be imposed *after* simulation, by reordering the MCMC output according to the constraint. This avoids possible negative effects on convergence.

Relabeling towards the mode

Selection of one of the $k!$ modal regions of the posterior once simulation is over, by computing the approximate MAP

$$(\boldsymbol{\theta}, \mathbf{p})^{(i^*)} \quad \text{with} \quad i^* = \arg \max_{i=1,\ldots,M} \pi \left\{ (\boldsymbol{\theta}, \mathbf{p})^{(i)} | \mathbf{x} \right\}$$

Pivotal Reordering

At iteration $i \in \{1, \ldots, M\}$,

1. Compute the optimal permutation

$$\tau_i = \arg \min_{\tau \in \mathfrak{S}_k} d\left(\tau \left\{ (\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}), (\boldsymbol{\theta}^{(i^*)}, \mathbf{p}^{(i^*)}) \right\} \right)$$

   where $d(\cdot, \cdot)$ distance in the parameter space.

2. Set $(\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}) = \tau_i((\boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}))$.

Re-ban on improper priors

Difficult to use improper priors in the setting of mixtures because independent improper priors,

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{k} \pi_i(\theta_i)\,, \quad \text{with} \quad \int \pi_i(\theta_i)\mathrm{d}\theta_i = \infty$$

end up, for all n's, with the property

$$\int \pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{p} = \infty\,.$$

Reason

There are $(k-1)^n$ terms among the k^n terms in the expansion that allocate *no observation at all* to the i-th component.

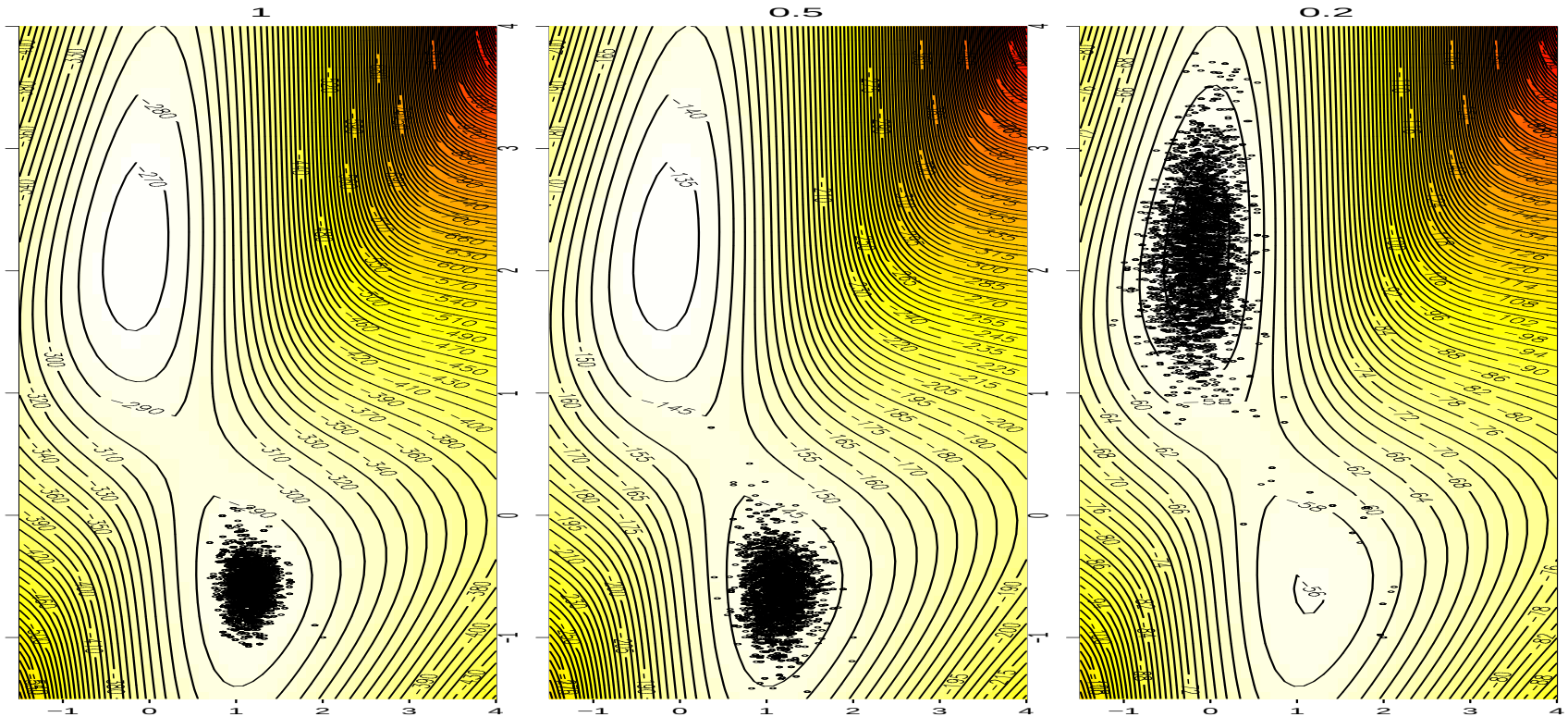
Tempering

Facilitate exploration of π by flattening the target: simulate from $\pi_\alpha(x) \propto \pi(x)^\alpha$ for $\alpha > 0$ large enough

- Determine where the modal regions of π are (possibly with parallel versions using different α's)
- Recycle simulations from $\pi(x)^\alpha$ into simulations from π by importance sampling
- Simple modification of the Metropolis–Hastings algorithm, with new acceptance

$$\left\{ \left(\frac{\pi(\boldsymbol{\theta}', \mathbf{p}'|\mathbf{x})}{\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x})} \right)^\alpha \frac{q(\boldsymbol{\theta}, \mathbf{p}|\boldsymbol{\theta}', \mathbf{p}')}{q(\boldsymbol{\theta}', \mathbf{p}'|\boldsymbol{\theta}, \mathbf{p})} \right\} \wedge 1$$

Tempering with the mean mixture



MCMC for variable dimension models

One of the things we do not know is the number of things we do not know
—P. Green, 1996—

Example

- the number of components in a mixture
- the number of covariates in a regression model
- the number of different capture probabilities in a capture-recapture model
- the number of lags in a time-series model

Variable dimension models

Variable dimension model defined as a collection of models $(k = 1, \ldots, K)$,

$$\mathfrak{M}_k = \{f(\cdot|\theta_k);\ \theta_k \in \Theta_k\}\ ,$$

associated with a collection of priors on the parameters of these models,

$$\pi_k(\theta_k)\,,$$

and a prior distribution on the indices of these models,

$$\{\varrho(k)\,, k = 1, \ldots, K\}\ .$$

Global notation:

$$\pi(\mathfrak{M}_k, \theta_k) = \varrho(k)\,\pi_k(\theta_k)$$

Bayesian inference for variable dimension models

Two perspectives:

1. consider the variable dimension model as a *whole* and estimate quantities meaningful for the whole like predictives

$$\sum_k \Pr(\mathfrak{M}_k|x_1,\ldots,x_n) \int f_k(x|\theta_k)\mathrm{d}x\, \pi_k(\theta_k|x_1,\ldots,x_n)\mathrm{d}\theta\,.$$

   & quantities only meaningful for submodels (like moments of θ_k), computed from $\pi_k(\theta_k|x_1,\ldots,x_n)$. *[Usual setup]*

2. resort to testing by choosing the best submodel via

$$p(\mathfrak{M}_i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}$$

Green's reversible jumps

Computational burden in exploring [possibly infinite] complex parameter space: Green's method set up a proper measure–theoretic framework for designing moves *between* models/spaces $\mathfrak{M}_k/\Theta_k$ of varying dimensions *[no one-to-one correspondence]*

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$] and for all sets A, B *[un-detailed balance]*

Green's reversible kernel

Since Markov kernel $\mathfrak{K}$ necessarily of the form *[either stay at the same value or move to one of the states]*

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y)\mathfrak{q}_m(x, dy) + \omega(x)\mathbb{I}_B(x)$$

where $\mathfrak{q}_m(x, dy)$ transition measure to model $\mathfrak{M}_m$ and $\rho_m(x, y)$ corresponding acceptance probability, only need to consider proposals between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, say.

Green's reversibility constraint

If transition kernels between those models are $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$, formal use of the *detailed balance condition*

$$\pi(d\theta_1)\,\mathfrak{K}_{1\to 2}(\theta_1, d\theta) = \pi(d\theta_2)\,\mathfrak{K}_{2\to 1}(\theta_2, d\theta)\,,$$

- To preserve stationarity, necessary symmetry between moves/proposals from $\mathfrak{M}_1$ to $\mathfrak{M}_2$ and from $\mathfrak{M}_2$ to $\mathfrak{M}_1$

Two-model transitions

How to move from model $\mathfrak{M}_1$ to $\mathfrak{M}_2$, with Markov chain being in state $\theta_1 \in \mathfrak{M}_1$ *[i.e.* $k = 1$*]*?

Most often $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are of different dimensions, e.g.

$$\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1).$$

In that case, need to supplement both spaces Θ_{k_1} and Θ_{k_2} with adequate artificial spaces to create a *one-to-one* mapping between them, most often by augmenting the space of the smaller model.

Two-model completions

E.g., move from $\theta_2 \in \Theta_2$ to Θ_1 chosen to be a *deterministic* transform of θ_2

$$\theta_1 = \Psi_{2\to1}(\theta_2)\,,$$

Reverse proposal expressed as

$$\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$$

where $v_{1\to2}$ r.v. of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to2} \sim \varphi_{1\to2}(v_{1\to2})\,.$$

Two-model acceptance probability

In this case, θ_2 has density *[under stationarity]*

$$\mathfrak{q}_{1\to2}(\theta_2) = \pi_1(\theta_1)\,\varphi_{1\to2}(v_{1\to2}) \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|^{-1},$$

by the Jacobian rule.
To make it $\pi_2(\theta_2)$ we thus need to accept this value with probability

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)}{\pi(\mathfrak{M}_1, \theta_1)\,\varphi_{1\to2}(v_{1\to2})} \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|.$$

- This is restricted to the case when only moves between $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are considered

Interpretation

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ in one-to-one relation.
- reversibility imposes that θ_1 is derived as

$$(\theta_1, v_{1\to 2}) = \Psi_{1\to 2}^{-1}(\theta_2)$$

- appears like a *regular* Metropolis–Hastings move from the couple $(\theta_1, v_{1\to 2})$ to θ_2 when stationary distributions are $\pi(\mathfrak{M}_1, \theta_1) \times \varphi_{1\to 2}(v_{1\to 2})$ and $\pi(\mathfrak{M}_2, \theta_2)$, and when proposal distribution is *deterministic* (??)

Pseudo-deterministic reasoning

Consider the proposals

$$\theta_2 \sim \mathcal{N}(\Psi_{1\to2}(\theta_1, v_{1\to2}), \varepsilon) \qquad \text{and} \qquad \Psi_{1\to2}(\theta_1, v_{1\to2}) \sim \mathcal{N}(\theta_2, \varepsilon)$$

Reciprocal proposal has density

$$\frac{\exp\left\{-(\theta_2 - \Psi_{1\to2}(\theta_1, v_{1\to2}))^2/2\varepsilon\right\}}{\sqrt{2\pi\varepsilon}} \times \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|$$

by the Jacobian rule.
Thus Metropolis–Hastings acceptance probability is

$$1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)}{\pi(\mathfrak{M}_1, \theta_1)\,\varphi_{1\to2}(v_{1\to2})} \left|\frac{\partial\Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})}\right|$$

Does not depend on ε: **Let ε go to 0**

Generic reversible jump acceptance probability

If several models are considered simultaneously, with probability $\varpi_{1\to 2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, as in

$$\mathfrak{K}(x,B) = \sum_{m=1}^{\infty} \int \rho_m(x,y)\mathfrak{q}_m(x,dy) + \omega(x)\mathbb{I}_B(x)$$

acceptance probability of $\theta_2 = \Psi_{1\to 2}(\theta_1, v_{1\to 2})$ is

$$\alpha(\theta_1, v_{1\to 2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2,\theta_2)\,\varpi_{2\to 1}}{\pi(\mathfrak{M}_1,\theta_1)\,\varpi_{1\to 2}\,\varphi_{1\to 2}(v_{1\to 2})} \left|\frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})}\right|$$

while acceptance probability of θ_1 with $(\theta_1, v_{1\to 2}) = \Psi^{-1}_{1\to 2}(\theta_2)$ is

$$\alpha(\theta_1, v_{1\to 2}) = 1 \wedge \frac{\pi(\mathfrak{M}_1,\theta_1)\,\varpi_{1\to 2}\,\varphi_{1\to 2}(v_{1\to 2})}{\pi(\mathfrak{M}_2,\theta_2)\,\varpi_{2\to 1}} \left|\frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})}\right|^{-1}$$

Green's sampler

Algorithm

Iteration t $(t \geq 1)$**:** if $x^{(t)} = (m, \theta^{(m)})$,

1. Select model $\mathfrak{M}_n$ with probability π_{mn}
2. Generate $u_{mn} \sim \varphi_{mn}(u)$ and set $(\theta^{(n)}, v_{nm}) = \Psi_{m \to n}(\theta^{(m)}, u_{mn})$
3. Take $x^{(t+1)} = (n, \theta^{(n)})$ with probability

$$\min\left(\frac{\pi(n, \theta^{(n)})}{\pi(m, \theta^{(m)})} \, \frac{\pi_{nm}\varphi_{nm}(v_{nm})}{\pi_{mn}\varphi_{mn}(u_{mn})} \left|\frac{\partial \Psi_{m \to n}(\theta^{(m)}, u_{mn})}{\partial(\theta^{(m)}, u_{mn})}\right|, 1\right)$$

and take $x^{(t+1)} = x^{(t)}$ otherwise.

Mixture of normal distributions

$$\mathfrak{M}_k = \left\{ (p_{jk}, \mu_{jk}, \sigma_{jk});\ \sum_{j=1}^{k} p_{jk} \mathcal{N}(\mu_{jk}, \sigma_{jk}^2) \right\}$$

Restrict moves from $\mathfrak{M}_k$ to adjacent models, like $\mathfrak{M}_{k+1}$ and $\mathfrak{M}_{k-1}$, with probabilities $\pi_{k(k+1)}$ and $\pi_{k(k-1)}$.

Mixture birth

Take $\Psi_{k\to k+1}$ as a *birth step*: i.e. add a new normal component in the mixture, by generating the parameters of the new component from the prior distribution

$$(\mu_{k+1}, \sigma_{k+1}) \sim \pi(\mu, \sigma) \quad \text{and} \quad p_{k+1} \sim \mathcal{B}e(a_1, a_2 + \ldots + a_k)$$

if $(p_1, \ldots, p_k) \sim \mathcal{M}_k(a_1, \ldots, a_k)$

Jacobian is $(1 - p_{k+1})^{k-1}$

Death step then derived from the reversibility constraint by removing one of the k components at random.

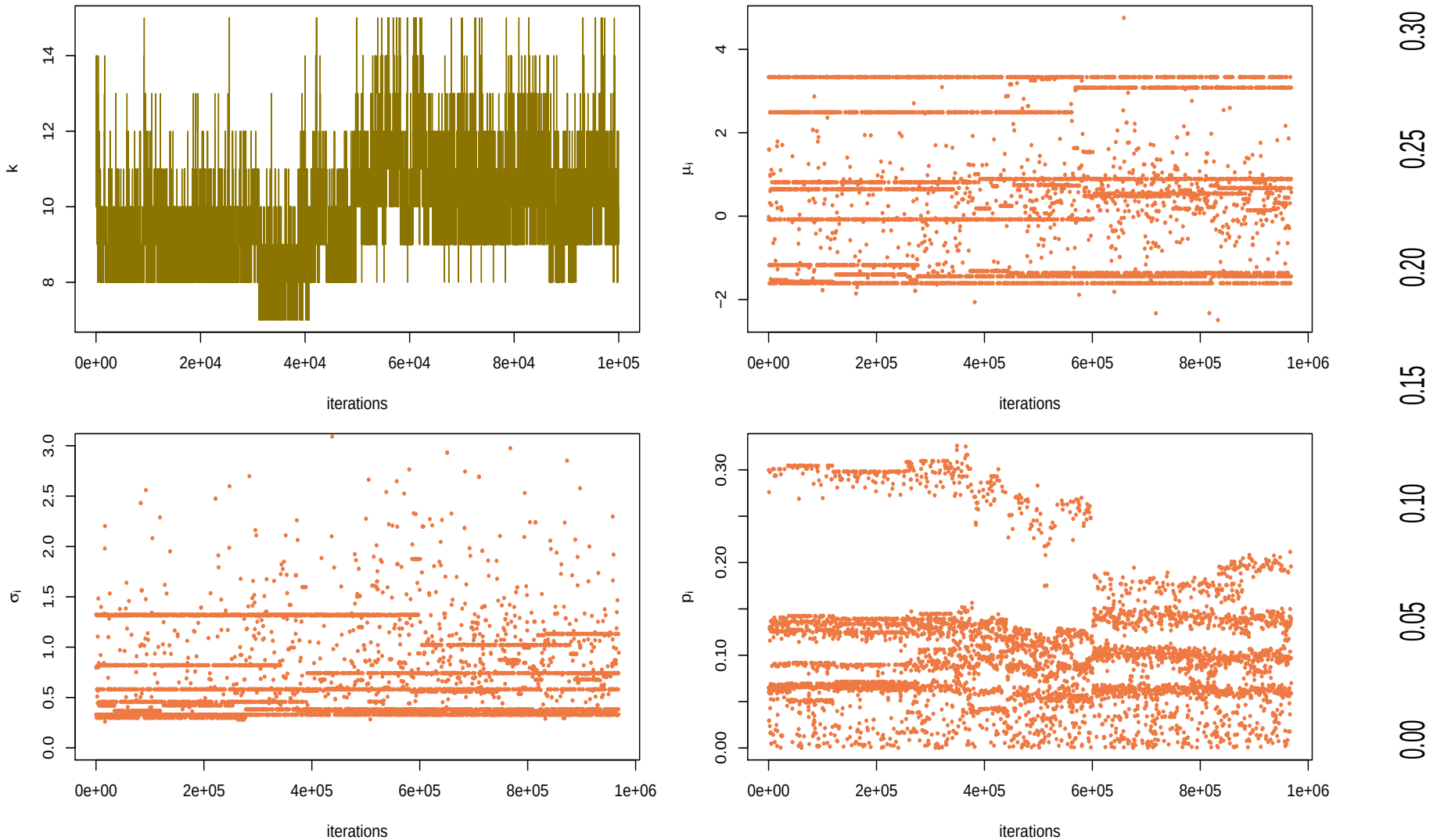
Mixture acceptance probability

Birth acceptance probability

$$\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{(k+1)!}{(k+1)k!}\,\frac{\pi(k+1,\theta_{k+1})}{\pi(k,\theta_k)\,(k+1)\varphi_{k(k+1)}(u_{k(k+1)})},1\right)$$
$$=\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\ell_{k+1}(\theta_{k+1})\,(1-p_{k+1})^{k-1}}{\ell_k(\theta_k)},1\right),$$

where ℓ_k likelihood of the k component mixture model $\mathfrak{M}_k$ and $\varrho(k)$ prior probability of model $\mathfrak{M}_k$.
Combinatorial terms: there are $(k+1)!$ ways of defining a $(k+1)$ component mixture by adding one component, while, given a $(k+1)$ component mixture, there are $(k+1)$ choices for a component to die and then $k!$ associated mixtures for the remaining components.

License



More coordinated moves

Use of local moves that preserve structure of the original model.

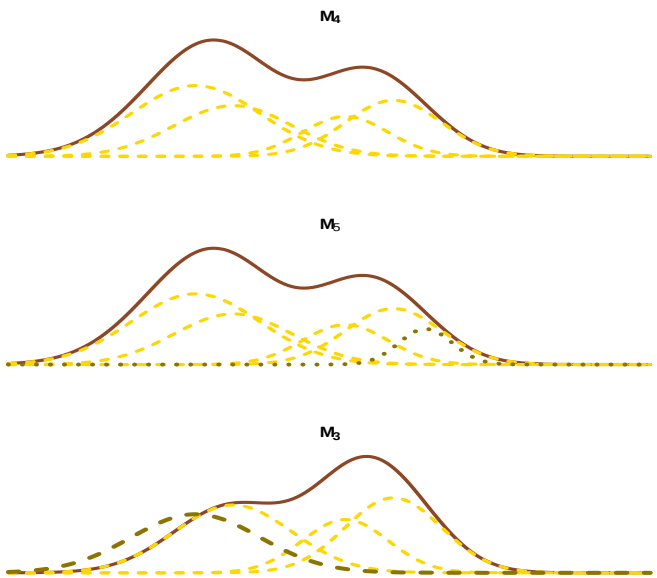
Split move from $\mathfrak{M}_k$ to $\mathfrak{M}_{k+1}$: replaces a random component, say the jth, with two new components, say the jth and the $(j+1)$th, that are *centered* at the earlier jth component. And opposite *merge* move obtained by joining two components together.

Splitting with moment preservation

Split parameters for instance created under a *moment preservation condition:*

$$\begin{aligned} p_{jk} &= p_{j(k+1)} + p_{(j+1)(k+1)}\,, \\ p_{jk}\mu_{jk} &= p_{j(k+1)}\mu_{j(k+1)} + p_{(j+1)(k+1)}\mu_{(j+1)(k+1)}\,, \\ p_{jk}\sigma^2_{jk} &= p_{j(k+1)}\sigma^2_{j(k+1)} + p_{(j+1)(k+1)}\sigma^2_{(j+1)(k+1)}\,. \end{aligned}$$

Opposite *merge* move obtained by reversibility constraint



Splitting details

Generate the auxiliary variable $u_{k(k+1)}$ as

$$u_1, u_3 \sim \mathcal{U}(0,1), u_2 \sim \mathcal{N}(0, \tau^2)$$

and take

$$\begin{array}{rclrcl}
p_{j(k+1)} &=& u_1 p_{jk}\,, & p_{(j+1)(k+1)} &=& (1-u_1)p_{jk}\,,\\
\mu_{j(k+1)} &=& \mu_{jk} + u_2\,, & \mu_{(j+1)(k+1)} &=& \mu_{jk} - \frac{p_{j(k+1)}u_2}{p_{jk}-p_{j)(k+1)}}\,,\\
\sigma^2_{j(k+1)} &=& u_3\sigma^2_{jk}\,, & \sigma_{(j+1)(k+1)} &=& \frac{p_{jk}-p_{j(k+1)}u_3}{p_{jk}-p_{j(k+1)}}\sigma^2_{jk}\,.
\end{array}$$

Jacobian

Corresponding Jacobian

$$\det \begin{pmatrix} u_1 & 1-u_1 & \cdots & \cdots & \cdots & \cdots \\ p_{jk} & -p_{jk} & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 1 & \cdots & \cdots \\ 0 & 0 & 1 & \frac{-p_{j(k+1)}}{p_{jk}-p_{j(k+1)}} & \cdots & \cdots \\ 0 & 0 & 0 & 0 & u_3 & \frac{p_{jk}-p_{j(k+1)}u_3}{p_{jk}-p_{j(k+1)}} \\ 0 & 0 & 0 & 0 & \sigma^2_{jk} & \frac{-p_{j(k+1)}}{p_{jk}-p_{j(k+1)}}\sigma^2_{jk} \end{pmatrix} = \frac{p_{jk}}{(1-u_1)^2}\,\sigma^2_{jk}$$

Acceptance probability

Corresponding split acceptance probability

$$\min\left(\frac{\widetilde{\pi}_{(k+1)k}}{\widetilde{\pi}_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\pi_{k+1}(\theta_{k+1})\ell_{k+1}(\theta_{k+1})}{\pi_k(\theta_k)\ell_k(\theta_k)}\,\frac{p_{jk}}{(1-u_1)^2}\,\sigma_{jk}^2,1\right)$$

where $\widetilde{\pi}_{(k+1)k}$ and $\widetilde{\pi}_{k(k+1)}$ denote split and merge probabilities when in models $\mathfrak{M}_k$ and $\mathfrak{M}_{k+1}$

Factorial terms vanish: for a split move there are k possible choices of the split component and then $(k+1)!$ possible orderings of the θ_{k+1} vector while, for a merge, there are $(k+1)k$ possible choices for the components to be merged and then $k!$ ways of ordering the resulting θ_k.