Advanced Simulation - Lecture 10

Patrick Rebeschini

February 14th, 2018

Outline

- Often we have various possible models for the same dataset.
- Sometimes there's an infinity of possible models!
- How to choose between models?

Green (1995), *Reversible Jump Markov chain Monte Carlo and Bayesian model determination.*

Motivation: Bayesian model choice

- Assume we have a collection of models $\mathcal{M}_k$ for $k \in \mathcal{K}$.
- With data we can learn parameters given each model $\mathcal{M}_k$, but we can also learn about the models.
- Put a prior on models $\mathcal{M}_k$. Within each model, prior $p(\theta_k \mid \mathcal{M}_k)$ on the parameters.
- Joint posterior distribution of interest:

$$\pi(\mathcal{M}_k, \theta_k \mid y) = \pi(\mathcal{M}_k \mid y)\pi(\theta_k \mid y, \mathcal{M}_k)$$

which is defined on

$$\cup_{k \in \mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k \equiv \cup_{k \in \mathcal{K}} \{k\} \times \Theta_k.$$

Polynomial regression

- Data $(x_i, y_i)_{i=1}^n$ where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$.
- Polynomial regression model

$$\mathcal{M}_k : y = \underbrace{\sum_{j=0}^{k} \beta_j x^j}_{=f(x;\beta)} + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

- If k is too large then

$$f\left(x; \widehat{\beta}\right) = \sum_{j=0}^{k} \widehat{\beta}_j x^j$$

where $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k\right)$ is the MLE, will overfit.

Polynomial regression

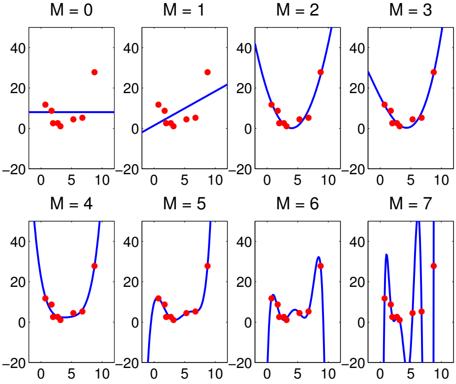

Figure: As order of the model $M = k$ increases, we overfit.

Bayesian polynomial regression

- We select $k \in \{0, ..., M_{\max}\}$ and

$$\mathbb{P}(\mathcal{M}_k) = p_k = \frac{1}{M_{\max} + 1}$$

with $\Theta_k = \mathbb{R}^{k+1} \times \mathbb{R}^+$

$$p_k\left(\beta, \sigma^2\right) = \mathcal{N}\left(\beta; 0, \sigma^2 I_{k+1}\right) \mathcal{IG}\left(\sigma^2; 1, 1\right).$$

- In this case, we have analytic expression for

$$p_k(y_{1:n}) = \int_{\Theta_k} p_k\left(\beta, \sigma^2\right) \prod_{i=1}^{n} \mathcal{N}\left(y_i; f(x_i; \beta), \sigma^2\right) d\beta d\sigma^2.$$

- Bayesian model selection automatically prevents overfitting.

Bayesian Polynomial regression

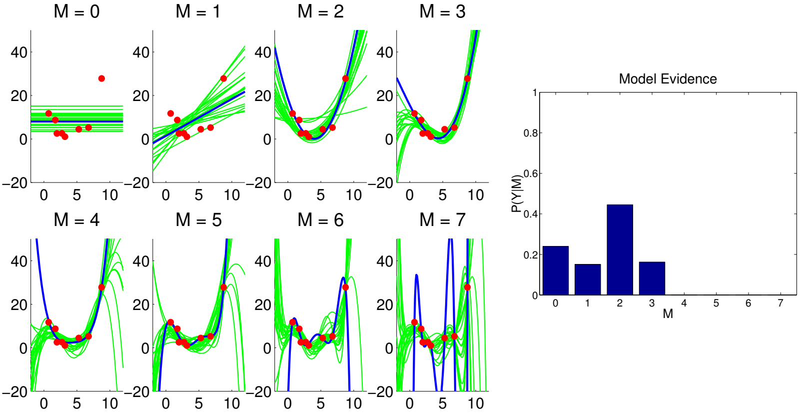

Figure: $f(x;\beta)$ for random draws from $p_M(\beta | y_{1:n})$ and evidence $p_M(y_{1:n})$.

Motivation: mixture models

- Assume the observations $Y_1, \ldots, Y_n$ come from

$$\sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \sigma_k^2)$$

with $\sum_{k=1}^{K} p_k = 1$. For any fixed K, the parameters to infer are $(p_1, \ldots, p_{K-1}, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$ of dimension $3K - 1$.

- But what about inference on K?
- We can put a prior on K, e.g. a Poisson distribution.
- How do we get the posterior?

Sampling in transdimensional spaces

- Consider a collection of models $\mathcal{M}_k$, for $k \in \mathcal{K} \subset \mathbb{N}$.
- We want to design a Markov chain taking values in $\cup_{k \in \mathcal{K}} \{k\} \times \Theta_k$, with the correct joint posterior.
- Reversible jump MCMC is a generalized Metropolis-Hastings using a mixture of kernels.
- For each k, standard MH kernel from $\{k\} \times \Theta_k$ to $\{k\} \times \Theta_k$, i.e. standard within-model moves.
- How to move from $\{k\} \times \Theta_k$ to $\{k'\} \times \Theta_{k'}$?

Transdimensional moves

We can propose k' from $q(k' \mid k)$. Then we need to propose a move from Θ_k to $\Theta_{k'}$, of dimension d_k and $d_{k'}$.

- **dimension matching**: extend the spaces with auxiliary variables.

- Introduce $u_{k \to k'}$ and $u_{k' \to k}$ with distributions $\varphi_{k \to k'}$ and $\varphi_{k' \to k}$ respectively, and such that

$$d_k + \dim(u_{k \to k'}) = d_{k'} + \dim(u_{k' \to k}).$$

Transdimensional moves

- Given θ_k, we sample $u_{k\to k'} \sim \varphi_{k\to k'}$ and then apply a deterministic mapping to get

$$(\theta_{k'}, u_{k'\to k}) = G_{k\to k'}(\theta_k, u_{k\to k'}).$$

- The distributions φ are arbitrary and $G_{k\to k'}$ has to be a diffeomorphism.

- We now have our proposal from Θ_k to $\Theta_{k'}$. With what probability do we accept it?

Transdimensional moves

- Mimicking Metropolis-Hastings, given x we propose a point x' and accept or not with probability $\alpha(x \to x')$.
- We want P to be such that, for all A, B:

$$\int_{x,x' \in A \times B} \pi(dx) P(x \to dx') = \int_{x,x' \in A \times B} \pi(dx') P(x' \to dx)$$

or equivalently

$$\int_{x,x' \in A \times B} \pi(dx) q(x \to dx') \alpha(x \to x') = \int_{x,x' \in A \times B} \pi(dx') q(x' \to dx) \alpha(x' \to x)$$

Transdimensional moves

- Subtle point: $\pi(dx)P(x, dx')$ does not necessarily admit a density with respect to a standard measure.

- We cannot write e.g.

$$\pi(x)P(x, dx') = \pi(x)P(x, x')dxdx'$$

- However $\pi(dx)q(x, dx')$ can be assumed to be dominated and we write

$$\pi(x)q(x, dx') = \pi(x)q(x, x')dxdx'$$

Transdimensional moves

- First term is:

$$\int_{x,x' \in A \times B} \pi(x) q(x \to x') \alpha(x \to x') dx dx'$$

- Suppose we propose x' by sampling $u \sim \varphi$ and then taking $(x', u') = G(x, u)$ deterministically. We write $x'(x, u)$ and $u'(x, u)$.
- The expression becomes

$$\int_{x,x'(x,u) \in A \times B} \pi(x) \varphi(u) \alpha(x \to x'(x, u)) dx du$$

- What is the reverse transition from x' to x? Sample $u' \sim \varphi'$ and take $(x, u) = G^{-1}(x', u')$.

Transdimensional moves

- Second term was:

$$\int_{x,x' \in A \times B} \pi(x') q(x' \to x) \alpha(x' \to x) dx dx'$$

- It becomes, with $(x, u) = G^{-1}(x', u')$:

$$\int_{x(x',u'),x' \in A \times B} \pi(x') \varphi'(u') \alpha(x' \to x(x', u')) dx' du'$$

Let us do a change of variable to get an integral with respect to $dxdu$ instead of $dx'du'$:

$$\int_{.} \pi(x'(x,u)) \varphi'(u'(x,u)) \alpha(x'(x,u) \to x) \left| \frac{\partial G(x,u)}{\partial (x,u)} \right| dxdu$$

Transdimensional moves

- We see that the integrals are equal if

$$\pi(x)\varphi(u)\alpha(x \to x'(x,u))$$
$$= \pi(x'(x,u))\varphi'(u'(x,u))\alpha(x'(x,u) \to x)\left|\frac{\partial G(x,u)}{\partial(x,u)}\right|$$

- Thus we can see a valid choice of $\alpha(x \to x')$ in :

$$\alpha(x \to x') = \min\left(1, \frac{\pi(x')\varphi'(u')}{\pi(x)\varphi(u)}\left|\frac{\partial G(x,u)}{\partial(x,u)}\right|\right)$$

Transdimensional moves

We can now answer the initial question:

- How to move from $\{k\} \times \Theta_k$ to some other $\{k'\} \times \Theta_{k'}$? We start from some (k, θ_k).
- Sample $k' \sim q(k \to k')$, then sample $u_{k \to k'}$ from $\varphi_{k \to k'}$.
- Compute deterministically $(\theta_{k'}, u_{k' \to k}) = G_{k \to k'}(\theta_k, u_{k \to k'})$.
- Compute

  $$\alpha_{k \to k'} = \min\left(1, \frac{\pi(\theta_{k'})\varphi_{k' \to k}(u_{k' \to k})}{\pi(\theta_k)\varphi_{k \to k'}(u_{k \to k'})} \frac{q(k' \to k)}{q(k \to k')} J_{k \to k'}(\theta_k, u_{k \to k'})\right)$$

  where

  $$J_{k \to k'}(\theta_k, u_{k \to k'}) = \left|\frac{\partial G_{k \to k'}(\theta_k, u_{k \to k'})}{\partial(\theta_k, u_{k \to k'})}\right|.$$

Reversible Jump algorithm

- Starting with $\left(k^{(0)}, \theta^{(0)}\right)$ iterate for $t = 1, 2, 3, ...$

- With probability β, set $k^{(t)} = k^{(t-1)}$ and do one step of $K_{k^{(t)}}$ leaving $\pi(\theta_{k^{(t)}} \mid y, \mathcal{M}_{k^{(t)}})$ invariant.

- With probability $1 - \beta$, propose $k' \sim q(k' \mid k^{(t-1)})$.
  - Draw a random variable $u_{k^{(t-1)} \to k'} \sim \varphi_{k^{(t-1)} \to k'}$.
  - Apply the deterministic mapping $G_{k^{(t-1)} \to k'}$ to get θ', u'.
  - With "between-models" acceptance probability $a(\theta^{(t-1)} \to \theta')$:
    accept, i.e. set $\theta^{(t)} = \theta', k^{(t)} = k'$,
    otherwise reject, i.e. set $\theta^{(t)} = \theta^{(t-1)}, k^{(t)} = k^{(t-1)}$.

Toy example

- Two models, uniform prior on models $p(\mathcal{M}_1) = p(\mathcal{M}_2) = \frac{1}{2}$.
- In model $\mathcal{M}_1$, $\theta \in \mathbb{R}$ and we can evaluate pointwise

$$\text{posterior}_1(\theta) \propto p(\theta \mid \mathcal{M}_1)\mathcal{L}(\theta \mid \mathcal{M}_1) = \exp\left(-\frac{1}{2}\left(\theta\right)^2\right)$$

- In model $\mathcal{M}_2$, $\theta \in \mathbb{R}^2$ and we can evaluate pointwise

$$\text{posterior}_2(\theta) \propto p(\theta \mid \mathcal{M}_2)\mathcal{L}(\theta \mid \mathcal{M}_2) = \exp\left(-\frac{1}{2}\left(\theta_1\right)^2 - \frac{1}{2}\left(\theta_2\right)^2\right)$$

Toy situation

- In terms of model comparison, we should find

$$\begin{aligned}\frac{p(\mathcal{M}_2 \mid y)}{p(\mathcal{M}_1 \mid y)} &= \frac{p(y \mid \mathcal{M}_2)p(\mathcal{M}_2)}{p(y \mid \mathcal{M}_1)p(\mathcal{M}_1)} \\ &= \frac{\int_{\mathbb{R}^2} p(\theta \mid \mathcal{M}_2)\mathcal{L}(\theta \mid \mathcal{M}_2)d\theta}{\int_{\mathbb{R}} p(\theta \mid \mathcal{M}_1)\mathcal{L}(\theta \mid \mathcal{M}_1)d\theta} \times \frac{\frac{1}{2}}{\frac{1}{2}} \\ &= \frac{2\pi}{\sqrt{2\pi}} \\ &= \sqrt{2\pi} \approx 2.5066\end{aligned}$$

- In terms of parameters, in model $\mathcal{M}_1$, $\theta \sim \mathcal{N}(0,1)$ and in model $\mathcal{M}_2$, $\theta \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$.

Reversible Jump algorithm

We need to construct various Markov kernels.

- A Markov kernel "within $\mathcal{M}_k$" for each model $\mathcal{M}_k$.

  Toy example: introduce a Metropolis Hastings with random walk proposal, of variance σ^2 for model $\mathcal{M}_1$ and Σ for model $\mathcal{M}_2$.

- A Markov kernel to move between models, i.e. for each pair k, proposing k' and proposing to move parameters of $\mathcal{M}_k$ to parameters of $\mathcal{M}_{k'}$.

  Toy example: introduce K_{12} moving a parameter $\theta \in \mathbb{R}$ to a parameter $(\theta_1, \theta_2) \in \mathbb{R}^2$, and introduce K_{21} moving a parameter $(\theta_1, \theta_2) \in \mathbb{R}^2$ to a parameter $\theta \in \mathbb{R}$.

Toy example

For K_{12} do the following.

- Sample u from $\mathcal{C}(0,1)$, a standard Cauchy (*dimension matching*).
- Map deterministically $(\theta_1, \theta_2) = G_{1\to 2}(\theta, u) = (\theta, u)$, with Jacobian equal to 1.
- Compute

$$\alpha_{1\to 2} = \min\left(1, \frac{\exp(-0.5\theta^2 - 0.5u^2)}{\exp(-0.5\theta^2)\mathcal{C}(u;0,1)}\right)$$

  Indeed the Jacobian is equal to 1, the priors on $\mathcal{M}_1$ and $\mathcal{M}_2$ are identical, and $q(k' \mid k) = q(k \mid k')$.

- Accept θ_1, θ_2 or stay at θ.

Toy example

For K_{21} do the following.

- Map deterministically $(\theta, u) = G_{2\to 1}(\theta_1, \theta_2) = (\theta_1, \theta_2)$, with Jacobian equal to 1.

- Compute

$$\alpha_{2\to 1} = \min\left(1, \frac{\exp(-0.5\theta_1^2)\mathcal{C}(\theta_2; 0, 1)}{\exp(-0.5\theta_1^2 - 0.5\theta_2^2)}\right)$$

  Indeed the Jacobian is equal to 1, the priors on $\mathcal{M}_1$ and $\mathcal{M}_2$ are identical, and $q(k' \mid k) = q(k \mid k')$.

- Accept θ or stay at (θ_1, θ_2).

Reversible Jump algorithm

- Introduce a probability of performing a "between-model" move at each step, say $\beta \in [0, 1]$.

- Given the current state of the chain k_t, θ_t at time t:

  - with probability β, between-model move: draw (k_{t+1}, θ_{t+1}) by drawing $k' \sim q(k' \mid k)$, dimension matching, deterministic mapping, RJ acceptance ratio...

  - with probability $1 - \beta$, within-model move: standard Metropolis-Hastings in the current model.

Results

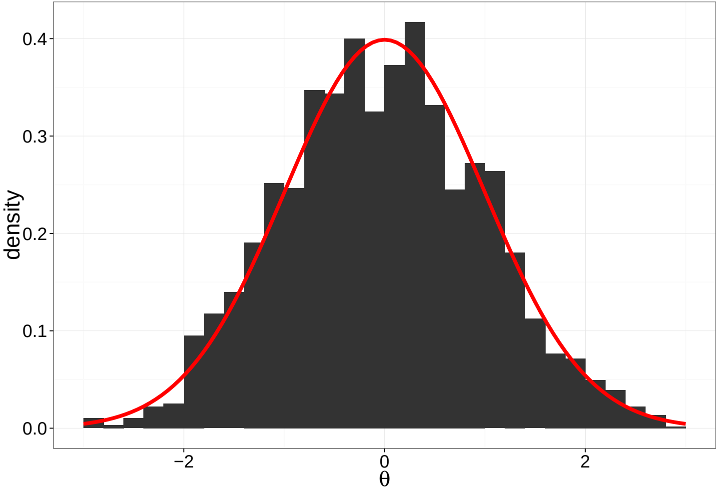

Figure: Parameter θ in model $\mathcal{M}_1$.

Results

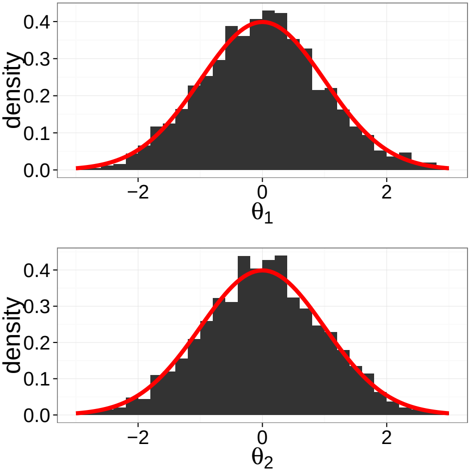

Figure: Parameter (θ_1, θ_2) in model $\mathcal{M}_2$.

Results

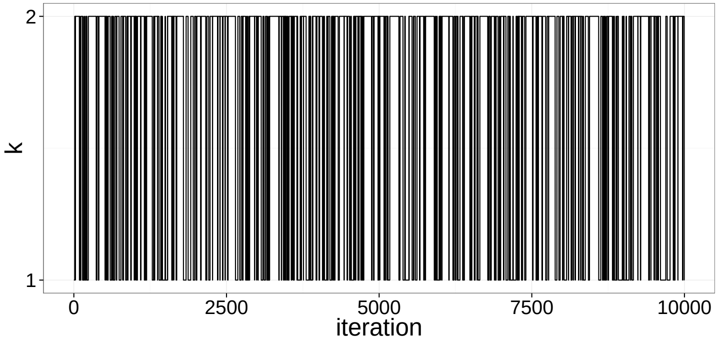

Figure: Model index k along iterations. Probability of accepting model jumps: $\approx 43.6\%$. The number of visits in $\mathcal{M}_2$ divided by the number of visits in $\mathcal{M}_1$ equals ≈ 2.39, approximating the Bayes factor of ≈ 2.51.

Results

If instead of $\mathcal{C}(0,1)$ we use $\mathcal{N}(3,1)$ for the dimension matching variable.

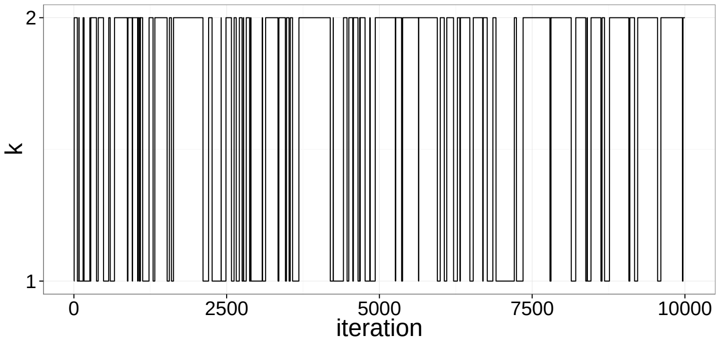

Figure: Model index k along iterations. Probability of accepting model jumps: $\approx 12.2\%$. Bayes factor approximated by ≈ 2.21.

Results

If instead of $\mathcal{C}(0, 1)$ we use $\mathcal{N}(5, 1)$ for the dimension matching variable.

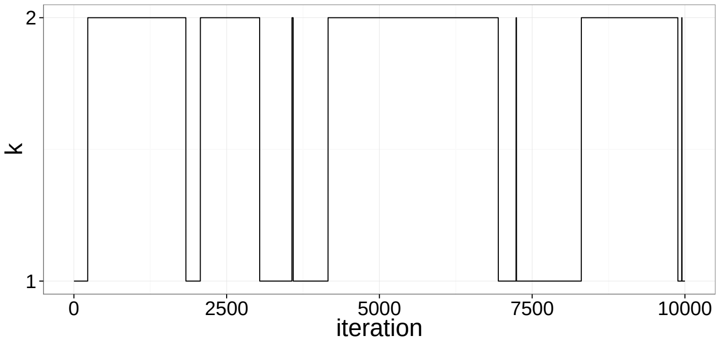

Figure: Model index k along iterations. Probability of accepting model jumps: $\approx 1.43\%$. Bayes factor approximated by ≈ 2.31 (not so bad!).

Reversible Jump algorithm: conclusion

- Probably the most ambitious MCMC algorithm, aiming at parameter estimation and model choice in one run.

- In general it's hard to design auxiliary variables for dimension matching and deterministic mappings such that the acceptance rate of between-model moves is decent.

- Transdimensional samplers constitute an on-going research area, see for instance:
  Annealed Importance Sampling Reversible Jump MCMC Algorithms, by Karagiannis and Andrieu, 2013.