MÉTODOS COMPUTACIONAIS
em
INFERÊNCIA BAYESIANA

Dani Gamerman
Instituto de Matemática
UFRJ

dani@dme.ufrj.br

Minicurso do 5o. Encontro Bayesiano
preparado com a colaboração de Mariane B. Alves

Disponível em http://acd.ufrj.br/~dani/teaching/5ebeb.ps

Baseado no artigo MCMC in Econometrics
a aparecer no Vol. 1 da revista Economia

Dezembro - 1999

Conteúdo

1. Motivação

2. Inferência Bayesiana

3. Métodos aproximados de inferência

4. Teoria de Monte Carlo com cadeias de Markov

5. Esquemas MCMC comuns

6. Aspectos computacionais

7. Modelos com estrutura estocástica complexa

8. Aplicações

1. Motivação

Exemplo 1: Efeito de uma campanha publicitária sobre as vendas de um produto

Y: Percentual de pessoas que se recordam de ter visto o anúncio
X: Gasto com o anúncio

Relação entre gasto e alcance do anúncio:

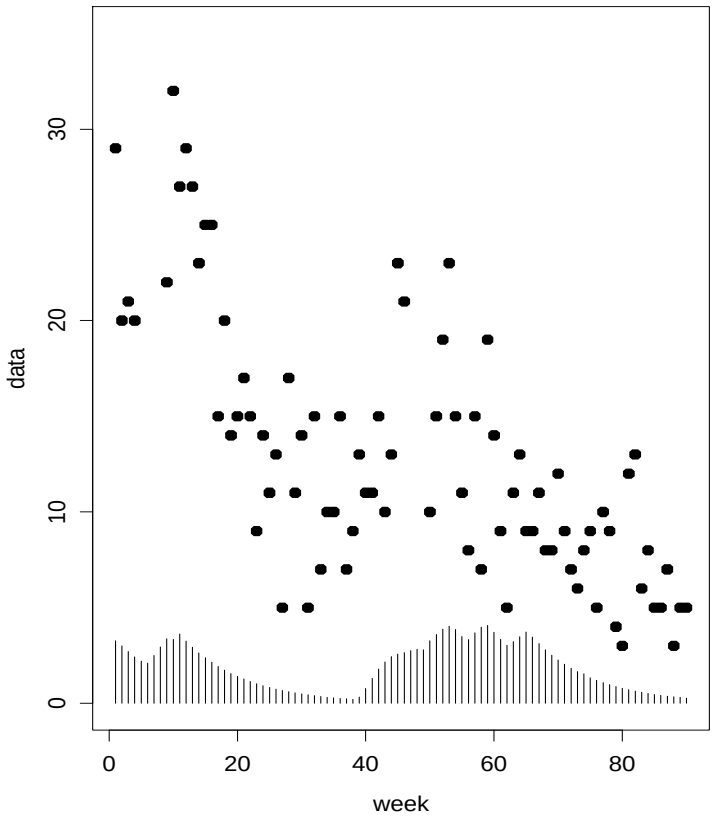
$$Y = \alpha + \beta X + e\,, \text{ onde } e \sim N(0, \sigma^2)$$

▷Quantidades observáveis: $\mathbf{W} = (X, Y)$
▷Quantidades não observáveis: $\theta = (\alpha, \beta, \sigma^2)$

Fontes de informação sobre as quantidades desconhecidas:

$*$ Experiência de marketing $p(\theta)$
$*$ Resultados de pesquisas, quando os percentuais de pessoas que se recordam do anúncio passam a ser conhecidos $p(\mathbf{W}|\theta)$

Dados sobre campanhas publicitárias



Objetivos:

1. Obter informação sobre as quantidades desconhecidas combinando as duas fontes de informação

2. Calcular, com base na informação obtida, valores médios e intervalos de probabilidade para as quantidades desconhecidas

3. Prever resultados futuros das pesquisas.

Exemplo 2: Avaliação da produção industrial

Sejam:
Y: Índice de produção industrial
X: Número de dias trabalhados

Relação entre produção e número de dias trabalhados:

$$Y = \alpha + \beta X + e\,, \text{ onde } e \sim N(0, \sigma^2)$$

▷Quantidades observáveis: $\mathbf{W} = (X, Y)$
▷Quantidades não observáveis: $\theta = (\alpha, \beta, \sigma^2)$

2. Inferência Bayesiana

Sejam:

- y: observações em um dado problema, com densidade $p(y|\theta)$

- θ: quantidades desconhecidas usadas para especificar a distribuição de y

- $p(\theta)$: densidade a priori para θ.

A densidade a posteriori de θ, dado y é obtida através do teorema de Bayes:

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

$p(y|\theta)$: verossimilhança de θ, também denotada por $l(\theta)$

No exemplo 1:
Priori→ Experiência de marketing
Verossimilhança→ Resultados de pesquisas, os percentuais de pessoas que se recordam do anúncio passam a ser conhecidos
Posteriori→ Distribuição de probabilidade dos parâmetros desconhecidos, à luz da experiência de marketing e dos resultados das pesquisas

Problema: $p(\theta|y)$ pode ser complexa e não permitir extração de informações tais como médias, medianas e intervalos de probabilidade.

Sistemas com estrutura estocástica complexa
$\Rightarrow$ Não é possível fazer simulação de forma direta
$\Rightarrow$ Utilizar métodos de aproximação

3. Métodos Aproximados de Inferência

Várias técnicas de aproximação desenvolvidas (anos 80), baseadas em preceitos analíticos e determinísticos

Difícil uso em espaços paramétricos com dimensões maiores que 10

Sejam:
π: distribuição a posteriori para o vetor de parâmetros θ
l: função de verossimilhança de θ
p: distribuição a priori para θ

1. Aproximação de π pela normal

   Obtida a partir da expansão de Taylor do $\log \pi$ em torno da moda

   MLG: moda obtida a partir do algoritmo MQPI (Mínimos Quadrados Ponderados Iterativos)

   Modelo de regressão normal é construído a cada iteração do algoritmo

2. Aproximação de Laplace

   $E_\pi[g(\theta)] = \int g(\theta)\pi(\theta)d\theta = \frac{\int g(\theta)l(\theta)p(\theta)d\theta}{\int \theta p(\theta)d\theta}$

   Aproximam-se integrandos por formas quadráticas no log

   Erros de ordem $\mathcal{O}(n^{-1})$ anulam-se $\rightarrow$ erro fica $O(n^{-2})$

   Densidades marginais podem ser obtidas, mas custo computacional é alto

3. Aproximação via quadratura Gaussiana

   Aproxima integrandos da forma $e^{-x^2/2}h(x)$

   Se h for uma função polinomial $\Rightarrow$ aproximação é exata

   Aproximações baseadas na normal: reparametrizações podem ser úteis

4. Métodos baseados em simulação estocástica

   Inferência baseada em amostras da posteriori π

   Amostra é sempre substituto parcial da informação de uma densidade

   Métodos analíticos: erros diminuem com o aumento do número n de observações

   Métodos baseados em simulação: erros dependem do número de valores gerados

Considere amostra $\theta_1, ..., \theta_n$ de $\pi(\theta)$ onde $\theta = (\theta_1, ..., \theta_p)$

Amostra marginal de θ_i: $\theta_{i1}, ..., \theta_{in}$

Amostra de $g(\theta)$: $g_1 = g(\theta_1), ..., g_n = g(\theta_n)$

Muito mais simples que a obtenção da posteriori de $g(\theta)$

De posse de uma amostra é possível obter:

1. estimativa pontuais. Ex: $E_\pi(\theta_i) \doteq \frac{1}{n} \sum_{j=1}^{n} \theta_{ij}$

2. intervalos de confiança

3. reconstrução de densidades marginais

Monte Carlo Simples

Uma estimativa de $E_\pi[g(X)]$ pode ser construída por

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X^{(i)}) \qquad \text{onde } X^{(i)} \sim \pi$$

A Lei Forte dos Grandes números garante que

$$Pr\left(\lim_{n\to\infty} \bar{g}_n = E_\pi[g(X)]\right) = 1$$

ou seja, $\bar{g}_n \overset{n\to\infty}{\longrightarrow} E_\pi[g(X)]$

A incerteza associada à estimativa é $O(n^{-1}) \Rightarrow$ Pode-se controlar a acurácia da estimação através do controle do tamanho da amostra simulada

Exemplo: Predição de valores futuros z.

Densidade preditiva:

$p(z|y) = \int p(z|\theta, y)p(\theta|y)d\theta = E_{\theta|y}[g(\theta)]$

onde $g(\theta) = p(z|\theta, y)$

▷ Utilizar um método aproximado para obtenção de $p(\theta|y)$

▷ Tomar uma amostra $\theta_1, \theta_2, \cdots, \theta_n$ de $p(\theta|y)$

▷ Avaliar $p(z|\theta, y)$

▷ Usar Monte Carlo simples para estimar o valor esperado:

$\widehat{p}(z|y) = \frac{1}{n}\sum_{i=1}^{n} p(z|\theta_i, y)$

Pela Lei Forte dos Grandes Números:
$\widehat{p}(z|y) \overset{n\to\infty}{\longrightarrow} p(z|y), \quad \forall z$

Métodos de reamostragem

Descritos aqui como técnicas que propõe geração em duas etapas:

1a.) são feitas gerações de uma distribuição aproximadora.

2a.) é exercido um mecanismo de correção para redirecionar a amostra.

Mecanismo de correção é tipicamente estocástico

densidade de interesse é π e
densidade auxiliar é q

Principais métodos:

a) método da rejeição

b) método da reamostragem ponderada

c) método da rejeição adaptativo

Método da rejeição

Método gera valores da densidade proporcional a π

Condição matemática sobre q:
$\exists A < \infty$ tal que $\pi(x) < Aq(x)$, $\forall x$.

O método gera $x \sim \pi$ em 2 etapas:

1a.) gera x de q e $u \sim U[0, 1]$ independentes

2a.) aceita x como um valor de π sse
$A\,u\,q(x) < \pi(x)$

Densidade q deve ser o mais próxima possível de π.

Obtenção de A deve ser feita com muito cuidado
$\rightarrow$ envelopamento deve ser total

Isso pode forçar valores altos de A: método fica ineficiente

Prob (aceitação) $\downarrow A$

Envelopamento inferior por $s \leq \pi$ pode otimizar geração
se computação de π for custosa (espremedura)

Método da reamostragem ponderada

Também chamado de amostragem-reamostragem por importância
em analogia com integração por Monte Carlo

Problema do método da rejeição:
Ném sempre é fácil obter constante A

MRP também gera valores de q (fácil de amostrar)

mas não precisa ser um envelope para π

Desvantagem do método: valores apenas aproximadamente distribuidos segundo π

Supondo amostra $x_1, ..., x_n$ de q, constroem-se pesos

$$w_i = \frac{\pi(x_i)/q(x_i)}{\sum_{j=1}^{n} \pi(x_j)/q(x_j)}, \quad i = 1, ..., n$$

Reamostra-se da distr discreta em $\{x_1, ..., x_n\}$ com prob. $w_1, ..., w_n$

Amostra resultante tem distribuição aproximadamente π

Método gera assintoticamente de π

MRP não precisa conhecimento completo de π

Reamostra pode ter qualquer tamanho

Reamostra deve ser tomada com reposição

Como no método da rejeição, $q \simeq \pi$

4. Teoria de Monte Carlo com Cadeias de Markov

Cadeia de Markov $\rightarrow$ Seqüência de variáveis aleatórias $\{X_j : j \in \mathcal{N}^+\}$ tal que,

$$p(x_j|x_{j_{-1}}, ..., x_1) = p(x_j|x_{j_{-1}}) \, , \forall j.$$

$S \rightarrow$ Espaço de estados: conjunto de possíveis valores de X.

Distribuição conjunta da cadeia de Markov:

$$p(x_1, x_2, ...) = \prod_{j=1}^{\infty} p(x_j|x_{j-1}) \qquad (1)$$

com valor inicial x_0.

$Q_j \rightarrow$ **Matriz de transição na iteração** j

$$q_{j,ik} = p(X_j = s_k|X_{j-1} = s_i), S = \{s_1, s_2, ...\}.$$

Cadeia homogênea: $Q_j = Q$, para todo j

Distribuição condicional completa de X_j

$$p(X_j|X_1,\cdots,X_{j-1},X_{j+1},\cdots) = p(X_j|X_{-j})$$

de (1) tem-se que:

$$p(x_j|x_{-j}) \propto p(x_{j+1}|x_j)p(x_j|x_{j-1}).$$

$\Rightarrow (X_j|X_{-j}) \sim (X_j|X_{j+1},X_{j-1})$:

apenas os vizinhos afetam a variável de uma cadeia.

Cadeia irredutível: Para todo par $(s_i, s_j) \in S$, a cadeia se movimenta do estado s_i para o estado s_j em um número finito de iterações com probabilidade positiva.

Estado positivo recorrente: o número de iterações para que a cadeia retorne ao mesmo estado tem média finita

Cadeia positiva recorrente: Todos os estados são positivos recorrentes.

Período: Um estado tem período d se a cadeia só pode retornar a este após um número de passos que é múltiplo de d.

Cadeia aperiódica: todos os seus estados são aperiódicos.

Ergodicidade = Aperiodicidade + Recorrência Positiva

Resultados Limite

(i) Existência de distribuição limite

Cadeia de Markov homogênea, irredutível, positiva recorrente e aperiódica $\Rightarrow$

Existe distribuição limite:

$X_n \to \pi$ quando $n \to \infty$

π é dada por

$$\pi(s) = \lim_{n \to \infty} Q^n(u, s), \ \forall u, s \in S.$$

O comportamento assintótico da cadeia depende apenas da matriz de transição Q.

π deve satisfazer:

$$\pi(s) = \sum_{u \in S} \pi(u) Q(u, s) \, , \, \forall u, s \in S. \qquad (2)$$

π é chamada distribuição de equilíbrio.

(2) é condição necessária para obter a distribuição limite, mas não é condição suficiente.

Reversibilidade: Uma cadeia de Markov é reversível se satisfaz:

$$\pi(s) Q(s, u) = \pi(u) Q(u, s) \, , \quad \text{for all } u, s \in S.$$

(ii) Teorema Ergódico

Para qualquer função $g(X)$, a média ergódica

$$\bar{g}_n = \frac{1}{n} \sum_{j=1}^{n} g(X_j) \quad \stackrel{n \to \infty}{\longrightarrow} \quad E_\pi[g(X)]$$

Médias amostrais convergem para médias teóricas sob a distribuição limite π

A acurácia das médias amostrais $\bar{g}_n$ é dada por

$$Var(\bar{g}_n) = c \, \frac{Var_\pi[g(X)]}{n},$$

onde $c = 1$ para amostragem independente e

$$c = 1 + 2 \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \rho_j \qquad (3)$$

para médias ergódicas de uma cadeia onde ρ_j é a correlação de ordem j, $j = 1, ..., n-1$ da cadeia

Para amostras indepententes, $\rho_j = 0, \forall j$

$\Rightarrow$ Para n fixo: Quanto maior a autocorrelação da cadeia, menos confiáveis são as médias ergódicas.

Com teorema ergódico, aproxima-se $E_\pi[g(X)]$ via Monte Carlo usando cadeias de Markov (MCMC)

5. Esquemas MCMC comuns

a. Amostrador de Gibbs (GS)

▷ Geman e Geman (1984): Reconstrução de imagens
▷ Gelfand e Smith (1990): Aplicações estatísticas

Sejam:

- X: variável d-dimensional de interesse

- X_1, X_2, ... , X_r: r blocos d_i-dimensionais de X, onde $d_i \geq 1$, $i = 1, ..., r$ e $d_1 + \cdots + d_r = d$

- π: Distribuição conjunta de X

- π_i: Distribuição condicional completa de X_i: $\pi_i(x_i) = \pi(x_i|x_{-i})$, $i = 1, ..., r$.

Esquema iterativo: Escolhe-se um bloco X_i e este é atualizado a partir de sua condicional completa π_i.

Algoritmo:

1. Inicie o contador de iterações em $j = 1$ e estabeleça os valores iniciais $x_1^{(0)}, \cdots, x_r^{(0)}$

2. Obtenha um novo valor $x^{(j)}$, a partir de $x^{(j-1)}$ usando gerações sucessivas:

$$
\begin{aligned}
x_1^{(j)} &\sim \pi(x_1 | x_{-1}^{(j-1)}) \\
x_2^{(j)} &\sim \pi(x_2 | x_{-2}^{(j-1)}) \\
&\vdots \\
x_r^{(j)} &\sim \pi(x_r | x_{-r}^{(j-1)})
\end{aligned}
$$

3. Incremente o contador de iterações $j = j + 1$ e volte ao passo (b) até que a convergência tenha sido atingida.

Após convergência, cada vetor $x^{(j)}$ representa um ponto amostral de π.

Valores da cadeia dependem apenas dos valores da iteração anterior $\Rightarrow$ a cadeia é Markoviana.

O esquema é plausível se a amostragem das r distribuições condicionais é simples.

b. Metropolis-Hastings (MH)

▷ Metropolis et al. (1953): Propriedades físico-químicas de substâncias.
▷ Hastings (1970): Extensão do método.

Cadeia de Markov com movimentos de x para y com uma proposta de núcleo de transição q.

Probabilidade de aceitação dos movimentos: $\alpha(x, y) = \min\{1, w(x, y)\}$

onde a razão de teste $w(x, y)$ é dada por

$$w(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}. \qquad (4)$$

O núcleo de transição da cadeia é

$$q(x, y)\alpha(x, y), \text{ para } y \neq x$$

Obs: a cadeia é reversível.

Algoritmo:

1. Inicie o contador de iterações em $j = 1$ e estabeleça o valor inicial $x^{(0)}$

2. Gere um valor $y \sim q(x^{(j-1)}, .)$ e um valor $u \sim U(0, 1)$ independente de y

3. Calcule o valor da probabilidade de aceitação, $\alpha(x, y) = \min\{1, w(x, y)\}$.

$$\text{Faça } x^{(j)} = \begin{cases} y & \text{se } u \leq \alpha(x^{(j-1)}, y) \\ x & \text{se } u > \alpha(x^{(j-1)}, y) \end{cases}$$

4. Incremente o contador de iterações: $j = j + 1$ e volte ao passo (b) até que a convergência tenha sido atingida.

Casos Particulares:

- Cadeia Passeio Aleatório:
  $q(x, y) = \phi(|y - x|)$

  Uma escolha usual para ϕ é a distribuição normal.

$$\Rightarrow \begin{cases} q(x, y) = q(y, x) \\ w(x, y) = \pi(y)/\pi(x) \end{cases}$$

- Cadeia Independente:

  $q(x, y) = \phi(y)$

  onde ϕ é uma densidade que aproxime π e com simulação simples.

  $\Rightarrow w(x, y) = \frac{\pi(y)/\phi(y)}{\pi(x)/\phi(x)}$

  Se ϕ= priori:

  $\Rightarrow w(x, y) = l(y)/l(x)$

c. Metropolis-Hastings em Blocos

▷ Combinação dos esquemas GS e MH.

X é dividido em blocos $X_1, \cdots, X_r$

Esquema iterativo:

Amostragem de cada bloco é feita de acordo com o algoritmo MH

x_i é atualizado para y_i segundo o núcleo de transição q_i

Probabilidade de aceitação: $\alpha_i(x_i, y_i)$

Razão de teste: $w_i(x_i, y_i)$, dada por (4), substituindo-se π e q por π_i e q_i, $i = 1, ..., r$.

Observações:

1. Cadeia é reversível e distribuição limite é π

2. Se π_i é de fácil geração, pode-se tomar $q_i = \pi_i$ e, nesse caso, $\alpha_i = 1$

6. Aspectos computacionais

a. Diagnósticos de Convergência

(i) Monitoração informal de convergência

Gelfand e Smith (1990) sugeriram técnicas gráficas

Considere m cadeias paralelas e componente qq de θ

Após j grande, construir histograma dos m valores gerados

Após $j + k$, construir outro histograma dos m valores e comparar

Histogramas parecidos: indício de convergência

k pequeno: influência da correlação inerente à cadeia de Markov

k grande: desnecessário

Tipicamente, valores entre 10 e 50 iterações são apropriados.

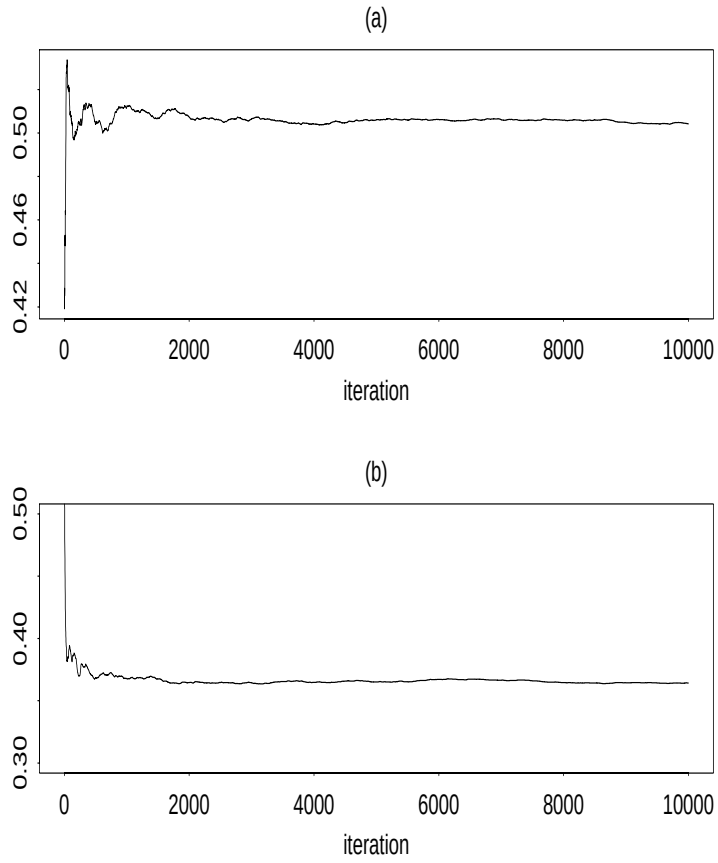
Similarmente, pode-se usar uma versão gráfica do teorema ergódico

Técnicas gráficas devem ser usados com muita cautela

Técnicas gráficas podem ser ilusórias

Além disso, problemas de metaestabilidade

Trajetória das médias ergódicas



(ii) Geweke (1992)

▷ Idéia: Após convergência, o comportamento deve ser estacionário.

⇒ Amostras do início e do final da cadeia após estacionariedade devem ser similares.

$\bar{g}_B$ and $\bar{g}_E$: médias de sub-amostras do início e do final da cadeia após *convergência*.

Sob a hipótese de estacionariedade:

$$\begin{aligned} E(\bar{g}_B - \bar{g}_E) &= 0 \text{ e} \\ Var(\bar{g}_B - \bar{g}_E) &= Var(\bar{g}_B) + Var(\bar{g}_E) \\ &= c\ Var_\pi[g(X)](n_B^{-1} + n_E^{-1}) \end{aligned}$$

onde

n_B e n_E: tamanhos das sub-amostras.

$c = 1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\rho_j$

Var_π é estimada a partir dos valores da cadeia.

A diferença padronizada entre as médias amostrais tem distribuição aproximadamente $N(0,1)$.

$\Rightarrow$ Rejeita-se a estacionariedade se a diferença padronizada entre as médias amostrais é grande.

A não rejeição da estacionariedade não implica necessariamente obtenção de convergência $\Rightarrow$ testar diferentes funções g.

Escolhas usuais para g: alguns componentes de X e $\log \pi(X)$.

(iii) Gelman and Rubin (1992)

▷ Idéia: Rodar algumas cadeias em paralelo e avaliar se seus comportamentos são semelhantes, para evitar ótimos locais.

Sugestão: Trabalhar com fator de redução potencial de escala, dado por:

$F = \sqrt{V/W}$, onde $V = (1 - 1/n)W + (1/n)B$

W: Variância dentro das cadeias

B: Variância entre as cadeias

Em geral, $F > 1$. Valores *aceitáveis* são menores que 1,2.

Prática usual: Gerar um pequeno número de cadeias paralelas e empregar os critérios de Geweke (propriedades médias das cadeias) e de Gelman e Rubin (variabilidade das cadeias).

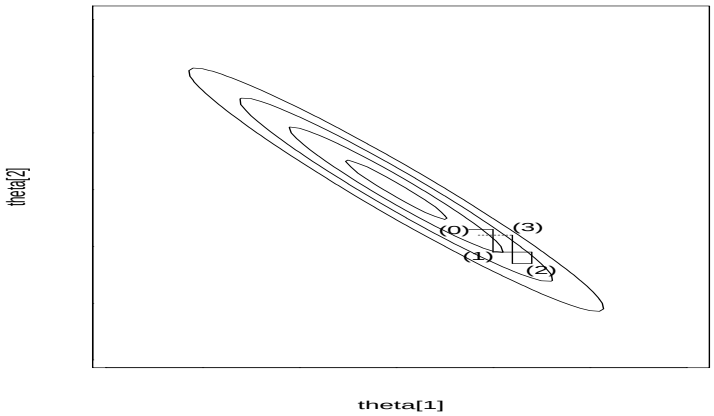
Os critérios acima não asseguram convergência, mas fornecem indícios bastante confiáveis.

b. Reparametrização

Situações típicas em MCMC: espaços com dimensões altas $\Rightarrow$
Divide-se em dimensões menores: amostrar por componentes do espaço de estados.

Problema: Deslocamentos ao longo das componentes depende da correlação em π

Posteriori com correlações altas entre componentes de X $\Rightarrow$ Deslocamentos muito pequenos: convergência lenta



Se não houver forte dependência $\Rightarrow$
movimentos amplos e convergência rápida

Caso extremo: independência a posteriori entre as componentes $\rightarrow$ condicionais coincidem com as marginais e convergência é imediata

Exemplo: No modelo de regressão, $\beta|\phi \sim N(\mathbf{b}_\phi, \mathbf{B}_\phi)$ pode-se obter uma matriz $\mathbf{A}$ tal que $\mathbf{B}_\phi^{-1} = \mathbf{A}\mathbf{A}'$.
Assim, $\alpha = \mathbf{A}\beta \sim N(\mathbf{A}b_\phi, I_p)$ dado ϕ.

$\Rightarrow$ Amostragem das componentes de α mais simples e a cadeia converge mais rápido

Desvantagem: cálculo extra da matriz **A**

De posse de um valor amostrado de α: obtém-se $\beta = \mathbf{A}^{-1}\alpha$

Estratégia geral baseada na normalidade aproximada a posteriori:

Calcular variância a posteriori aproximada

Extrair sua raiz quadrada $\mathbf{A}$ e fazer $\alpha = \mathbf{A}X$, de forma a obter independência aproximada

Dificuldade adicional pode ser causada pela amostragem de α, ao invés de X

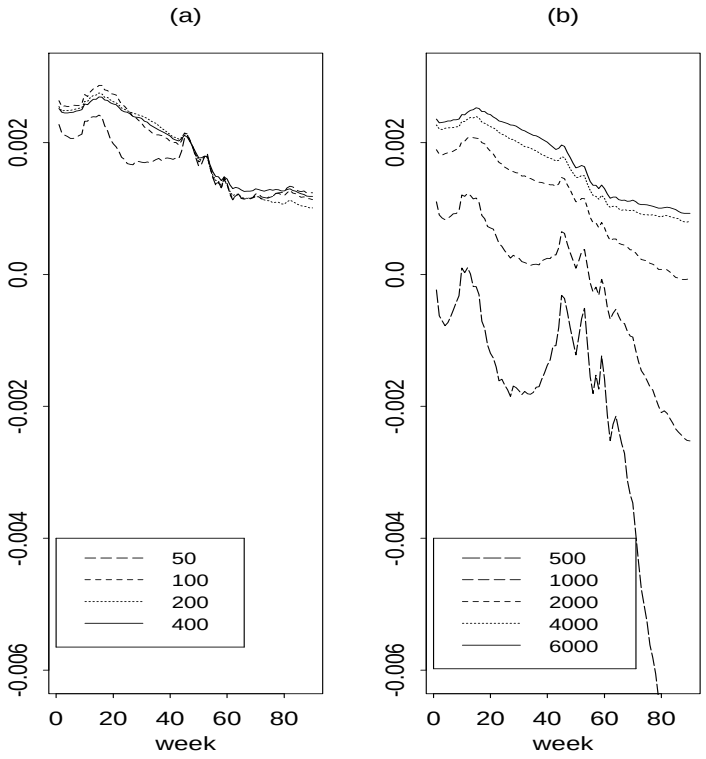
Outro ponto importante: estudar estrutura do modelo e dos parâmetros

Cada parâmetro transformado deve se comportar como normal

Parâmetros de variância tipicamente não terão este comportamento. Como são positivos, recomenda-se tranformação logarítmica

Exemplo onde reparametrização é importante

Modelos dinâmicos não-normais



(a) - modelo com reparametrização

(b) - modelo sem reparametrização

c. Escolha de Blocos

Como dividir X em componentes $X_1, ..., X_r$?

Gerações para blocos de parâmetros: cadeia caminha por direções diferentes das determinadas pelos eixos das componentes univariadas
▷ Melhoramento na velocidade de convergência

Exemplo: agrupando parâmetros altamente correlacionados em um bloco

Estimativas obtidas para componentes em blocos são sempre mais precisas

Vetores paramétricos especificados em bloco formam blocos naturais

Regra básica: agrupar o máximo possível

No limite: um único bloco e condicional completa é a conjunta

Convergência é imediata e o amostrador de Gibbs é desnecessário!

7. Modelos com Estrutura Estocástica Complexa

Extensões do modelo de regressão linear

Distribuições alvo: posterioris de todos os parâmetros do modelo

$y_1, ..., y_n \rightarrow$ observações independentes e $y_i \sim N(\mu_i, \phi)$, para $i = 1, ..., n$

Regressão padrão $\rightarrow \mu_i = x_i'\beta$ onde
x_i é vetor r-dimensional de variéveis explicativas para a i-ésima observação $(i = 1, ..., n)$

Extensão $\rightarrow \mu_i = x_i'\beta_i$, para $i = 1, ..., n$

Quantidades desconhecidas (β, ϕ), onde $\beta = (\beta_1, ..., \beta_n)$ e verossimilhança

$$l(\beta, \phi) = (2\pi)^{-n/2}\phi^{-n/2} \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i - x_i'\beta_i)^2\right\}$$

Priori para (β, ϕ): depende de quantidades desconhecidas Ψ, em geral parâmetros de variância $\rightarrow$ prioris vagas para estes hiperparâmetros

Posteriori dada por

$$\pi(\beta, \phi, \Psi|y) \propto l(\beta, \phi) f(\beta|\Psi) p(\phi) p(\Psi)$$

Possíveis hipóteses:

- Densidade conjunta f não depende de ϕ
- ϕ e Ψ independentes a priori

Modelos hierárquicos

Especificação da priori de β: β_i's são permutáveis $\Rightarrow$ heterogeneidade não estruturada e

$$f(\beta|\Psi) = \prod_i p(\beta_i|\Psi)$$

Variância observacional $\phi \rightarrow$ priori Gama invertida

$$p(\phi) \propto \phi^{-\nu_\phi/2} \exp\{-S_\phi \phi^{-1}/2\}$$

Ψ matriz de variâncias $\rightarrow$ priori Wishart invertida

$$p(\Psi) \propto |\Psi|^{-\nu_\psi/2} \exp\{-tr(S_\psi \Psi^{-1})/2\}$$

Independente da forma de f, posteriori condicional completa para ϕ dada por

$$\pi(\phi|\beta, \Psi, y) \propto l(\beta, \phi) p(\phi)$$

$$\propto \phi^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\phi} \sum_{i=1}^{n} (y_i - x_i'\beta_i)^2\right\} \phi^{-\frac{\nu_\phi}{2}} \exp\left\{-\frac{1}{2\phi} S_\phi\right\}$$

$$\propto \phi^{-(\nu_\phi+n)/2} \exp\left\{-\frac{1}{2}\phi^{-1}\left[S_\phi + \sum_{i=1}^{n} (y_i - x_i'\beta_i)^2\right]\right\}$$

que é Gamma invertida $\Rightarrow$ fácil de gerar

Regressão variando no tempo

Modelo tem dependência temporal entre os coeficientes da regressão

▷ Modelos Dinâmicos Lineares (West e Harrison, 1997) ou
Modelos de Espaço de Estados (Harvey, 1989)

Estrutura temporal dos β_i's dada por cadeia de Markov

Forma usual:
$(\beta_i|\beta_{i-1},...,\beta_1,\Psi) \sim (\beta_i|\beta_{i-1},\Psi) \sim N(G_i\beta_{i-1}, W_i)$

Priori para $\beta|\Psi$ é completada com $\beta_1 \sim N(a_1, R_1)$

Priori conjunta $f(\beta|\Psi)$ dada por

$$\exp\left\{-\frac{1}{2}\left[\sum_{i=2}^{n}(\beta_i - G_i\beta_{i-1})'W_i^{-1}(\beta_i - G_i\beta_{i-1})\right.\right.$$
$$\left.\left. + (\beta_1 - a_1)'R_1^{-1}(\beta_1 - a_1)\right]\right\} \prod_{i=2}^{n}|W_i|^{-\frac{r}{2}}$$

Caso particular: $G_i = G$ e $W_i = W$, para todo i, ambos G e W podem depender de Ψ

Se $G = I_r \Rightarrow$ relação temporal entre os coeficientes da regressão é passeio aleatório e

$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \Rightarrow$ tendência linear local

Componentes sazonais (e modelos com fatores de classificação) $\rightarrow$ pode-se usar matriz de permutação

$$G = \begin{pmatrix} 0_{r-1} & I_{r-1} \\ 1 & 0'_{r-1} \end{pmatrix}$$

r: período da sazonalidade (ou número de níveis de fatores)
0_p: vetor p-dimensional de 0's.

Quando aplicada sobre qualquer vetor $(z_1, ..., z_r)$, a matriz de transição retorna $(z_2, ..., z_r, z_1)$.

Observação: Os modelos podem ser aplicados on-line a qualquer instante $i \rightarrow$ processo de inferência dado por seqüência de posterioris correntes

▷ Aqui: inferência da posteriori completa dada a série temporal completa

Priori dinâmica implica $\beta \sim N(a, R)$ onde $a = (a_1, ..., a_n)$, $R = (R_{ij})$

Blocos a_i e R_{ij} definidos recursivamente por:

$a_i = G_i a_{i-1}$
$R_{ii} = G_i R_{i-1,i-1} G_i' + W_i$
$R_{ij} = \prod_{l=i}^{i-j+1} G_l R_{jj}$

para $i > j = 1, ..., n$

Amostragem dos β's

$1^{\underline{a}}$ **opção:** Amostrar β_i's

Posteriori condicional completa de β_i:

$$\pi(\beta_i|\beta_{-i},\phi,\Psi,y) \propto p(y_i|\beta_i,\phi) f(\beta_i|\beta_{i-1},\Psi) f(\beta_{i+1}|\beta_i,\Psi),$$

para $i = 2,...,n-1$, com ajustes para $i = 1$ e $i = n$

Termos quadráticos em β_i na exponencial $\Rightarrow$ $(\beta_i|\beta_{-i},\phi,\Psi,y) \sim$ Normal: simulação simples e, no máximo, inversão de matrizes $r \times r$

alta correlação a priori entre os β_i's $\Rightarrow$ convergência pode ser muito lenta

$2^{\underline{a}}$ **opção:** Obter posteriori condicional completa de β

Teorema de Bayes: combinação da verossimilhança $l(\beta, \phi)$ com priori $f(\beta|\Psi)$

$\Rightarrow \beta|\phi, \Psi, y \sim N(m_t, C_t)$
onde
$m_t = C_t[R^{-1}a + (x_1'y_1, ..., x_n'y_n)']$
$C_t = (\phi^{-1}diag(x_1x_1', ..., x_nx_n') + R)^{-1}$

$\Rightarrow$ Simulação fácil, mas necessidade de inverter matrizes $nr \times nr$

Tipicamente, r pequeno e n grande $\Rightarrow$ o ganho na velocidade de convergência se perde no custo de cada iteração

Solução$\rightarrow$ Algoritmo que fornece amostras de β com n inversões de matrizes $r \times r$ (Fruhwirth-Schnatter (1994) e Carter e Kohn (1994)):

Estrutura markoviana da priori implica

$$\pi(\beta|\phi,\Psi,y) = \pi(\beta_n|\phi,\Psi,y^n) \prod_{i=1}^{n-1} \pi(\beta_i|\beta_{i+1},...,\beta_n,\phi,\Psi,y^n)$$
$$= \pi(\beta_n|\phi,\Psi,y^n) \prod_{i=1}^{n-1} \pi(\beta_i|\beta_{i+1},\phi,\Psi,y^i)$$

onde $y^j = (y_1,...,y_j)$, para $j = 1,...,n$

$\Rightarrow$ amostragem conjunta de β equivale a amostrar β_n de $(\beta_n|\phi,\Psi,y^n)$ e β_i's de $(\beta_i|\beta_{i+1},\phi,\Psi,y^i)$, sucessivamente para $i = n-1,...,1$.

Amostragem do bloco completo β eleva substancialmente a taxa de convergência do algoritmo MCMC (quando comparado à amostragem dos β_i's separadamente)

Obtenção das distribuições relevantes na expressão de $\pi(\beta|\phi, \Psi, y)$: Rodar o filtro de Kalman para i variando de 1 a n e suavizar de n para 1

Filtro de Kalman: Baseia-se no teorema de Bayes - fornece meios para atualização recursiva de $(\beta_i|\phi, \Psi, y^i)$, $i, ..., n$

Amostragem direta de $\pi(\Psi|\beta, \phi, y)$ raramente é possível

Exceção: G_i's conhecidas e $W_i = W = \Psi$, para todo i, tem priori Wishart invertida

$3^{\underline{a}}$ **opção:** Integrar β fora

$$\pi(\phi, \Psi|y) = \pi(\beta, \phi, \Psi|y)/\pi(\beta|\phi, \Psi, y)$$

expressão da posteriori condicional de $(\beta|\phi, \Psi)$ conhecida $\Rightarrow$ expressão da posteriori marginal de (ϕ, Ψ) conhecida $\Rightarrow$ pode-se usar MH para gerar amostra de (ϕ, Ψ)

Inferência para β é baseada em $\pi(\beta|y)$ que pode ser aproximada por MC simples como

$$\widehat{\pi}(\beta|y) = \frac{1}{n}\sum_{i=1}^{n} \pi(\beta|\phi^{(i)}, \psi^{(i)}, y)$$

com $(\phi, \psi)^{(i)} \sim \pi(\phi, \psi|y)$

Regressão variando no espaço

Uma coleção de quantidades aleatórias $X = (X_1, ..., X_n)$ forma um campo aleatório markoviano (CAM) se a distribuição conjunta de X satisfaz

$$(X_i|X_{-i}) \sim (X_i|X_{\partial i}) \quad , \quad i = 1, ..., n$$

onde $\partial i = \{j : j \text{ é um vizinho de } i\}$

Quando as regiões são ordenadas

$\partial i = \{i-1\,,\, i+1\}$ e $(X_i|X_{-i}) \sim (X_i|X_{i-1}, X_{i+1})$

para todo i fora dos extremos

$\Rightarrow$ CAM fornece generalização das cadeias de Markov

Exemplo de priori para β por CAM:

$$f(\beta|\Psi) \propto |\Psi|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(\beta_i-\beta_j)'\Psi^{-1}(\beta_i-\beta_j)\right\}$$

onde w_{ij} são pesos associados à estrutura espacial

Por exemplo: $w_{ij} = \begin{cases} 1 & , \text{ if } i \sim j \\ 0 & , \text{ otherwise} \end{cases}$,
onde $i \sim j$ significa que i e j são vizinhos

Priori atribui probabilidades maiores a regiões que tenham valores similares para β's vizinhos

Priori imprópria, pois $Var(\beta) = K^{-1} \otimes \Psi$ onde

$$K = (k_{ij}) \quad \text{e} \quad k_{ij} = \begin{cases} n_i & , \text{ if } i = j \\ -1 & , \text{ if } i \sim j \\ 0 & , \text{ otherwise} \end{cases}$$

$\Rightarrow$ linhas da matriz K $(n \times n)$ somam zero $\Rightarrow$ K^{-1} não existe

$1^{\underline{a}}$ **opção:** Amostrar β_i's

Pode-se mostrar que

$$\beta_i|\beta_{-i} \sim \beta_i|\beta_{\partial i} \sim N\left(\bar{\beta}_{\partial i}, \frac{1}{n_i}\Psi\right)$$

$\bar{\beta}_{\partial i}$ é a média dos n_i β's vizinhos
n_i é o número de vizinhos de i, for $i = 1, ..., n$.

Posteriori condicional completa de β_i dada por

$$(\beta_i|\beta_{-i}, \phi, \Psi, y) \sim (\beta_i|\beta_{\partial i}, \phi, \Psi, y) \sim N(a_i, R_i)$$

$a_i = R_i(\phi^{-1}x_i y_i + n_i \Psi^{-1}\bar{\beta}_{\partial i})$
$R_i = (\phi^{-1}x_i x_i' + n_i\Psi^{-1})^{-1}$

para $i = 1, ..., n$

$\Rightarrow$ simulação fácil. Esforço computacional reside na inversão de matrizes de variância $r \times r$

2$^{\underline{a}}$ **opção:** Obter posteriori condicional completa de β

$\beta|\phi, \Psi, y \sim N(m_s, C_s)$
$m_s = C_s(x_1' y_1, ..., x_n' y_n)'$
$C_s = (\phi^{-1} diag(x_1 x_1', ..., x_n x_n') + K \otimes \Psi^{-1})^{-1}$

$\Rightarrow$ simulação fácil, mas esforço computacional elevado, pois as matrizes a inverter são $nr \times nr$

3$^{\underline{a}}$ **opção:** Integrar β fora

$$\pi(\phi, \Psi|y) = \pi(\beta, \phi, \Psi|y)/\pi(\beta|\phi, \Psi, y)$$

expressão da posteriori condicional de $(\beta|\phi, \Psi)$ conhecida $\Rightarrow$ expressão da posteriori marginal de (ϕ, Ψ) conhecida $\Rightarrow$ pode-se usar MH para gerar amostra de (ϕ, Ψ)

Finalmente, obtém-se também a posteriori condicional completa de Ψ

$$\begin{aligned}
\pi(\Psi|\beta,\phi,y) &\propto f(\beta|\Psi)p(\Psi) \\
&\propto |\Psi|^{-n/2}\exp\left\{-\frac{1}{2}tr\left(Q_s(\beta)\Psi^{-1}\right)\right\} \\
&\times \quad |\Psi|^{-\nu_\psi/2}\exp\left\{-\frac{1}{2}tr\left(S_\psi\Psi^{-1}\right)\right\} \\
&\propto |\Psi|^{-\frac{\nu_\psi+n}{2}}\exp\left\{-\frac{1}{2}tr\left[\left(S_\psi+Q_s(\beta)\right)\Psi^{-1}\right]\right\}
\end{aligned}$$

onde $Q_s(\beta) = \sum_{i,j=1}^{n} w_{ij}(\beta_i - \beta_j)(\beta_i - \beta_j)'$

Tem-se novamente densidade Wishart invertida $\Rightarrow$ pode-se usar qualquer esquema MCMC paraa amostrar esta componente

Exemplos de especificação dos pesos:

Dependência do tamanho das regiões envolvidas:

$$W = (w_{ij}) \qquad \text{onde } w_{ij} = 1/A_i A_j$$

onde A_k é a área da região k

Dependência da distância entre as regiões envolvidas:

$$w_{ij} = \exp(-b d_{ij})$$

onde d_{ij} é a distância entre as regiões i and j

Parâmetro b desconhecido $\Rightarrow$ priori para b incluída no esquema MCMC

8. Aplicações

a. Índice de produção industrial

Dados coletados pelo IBGE

Produção de um índice mensal que serve como referência econômica

Índice afetado por sazonalidade e número de dias úteis no mês.

Investigação sobre a existência de ciclos de negócios e adição de tendência variando com o tempo

$\Rightarrow$ Modelo construído com $\mu_i = t_i + s_i + c_i + \eta_i d_i$

onde
t_i: tendência
s_i: efeito sazonal
c_i: efeito cíclico
d_i: número de dias úteis no mês i
para $i = 1, ..., n$

Tendência modelada por crescimento linear:

$$\begin{aligned} t_i &= t_{i-1} + g_i + w_{1i}\,,\ w_{ti} \sim N(0, \psi_t) \\ g_i &= g_{i-1} + w_{gi}\,,\ w_{gi} \sim N(0, \psi_g) \end{aligned}$$

onde g_i é o crescimento local da tendência para o mês i

$\psi_g = \psi_t = 0 \rightarrow$ tendência descrita por uma reta fixa com inclinação $g = g_i$, para todo i

Efeito sazonal:
$s_i = s_{i-12} + w_{si}$, com $w_{si} \sim N(0, \psi_s) \rightarrow$
Soma dos efeitos anuais tem média nula

$\psi_s = 0 \Rightarrow$ padrão sazonal fixo

Modelagem dos ciclos:

$$c_i = \psi_1 c_{i-1} + \psi_2 c_{i-2} + w_{ci} \text{ com } w_{ci} \sim N(0, \psi_c)$$

Comprimento do ciclo: $\lambda = 2\pi / \cos^{-1}(\psi_1 / 2\rho)$
Decaimento do ciclo: $\rho = \sqrt{-\psi_2}$

Inferência amostral de λ e ρ: tomar amostras de (ψ_1, ψ_2) e aplicar as transformações acima.

Resultado: amostra da posteriori marginal de (λ, ρ)

O coeficiente dos dias trabalhados η_i é descrito por um passeio aleatório

$$\eta_i = \eta_{i-1} + w_{di} \text{ com } w_{di} \sim N(0, \psi_d)$$

O modelo formado tem:

regressores $x_i' = (\iota_2', \iota_{12}', \iota_2', d_i)$

coeficientes $\beta_i = (t_i, g_i, s_i, s_{i1}..., s_{i,11}, c_i, c_{i-1}, \eta_i)$,

onde $s_i = -\sum_{j=1}^{11} s_{i,j}$

matriz de transição $G_i = G = diag(G_t, G_s, G_c, 1)$,

variância da evolução $W_i = W = diag(W_t, W_s, W_c, \psi_d)$.

ι_r é o vetor r-dimensional com primeiro elemento 1 e demais elementos nulos

$G_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $G_s = \begin{pmatrix} 0_{11} & I_{11} \\ 1 & 0_{11}' \end{pmatrix}$ e $G_c = \begin{pmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{pmatrix}$

$W_t = \begin{pmatrix} \psi_t + \psi_g & \psi_g \\ \psi_g & \psi_g \end{pmatrix}$, $W_s = \psi_s \; diag(\iota_{12})$ e $W_c = \psi_c \; diag(\iota_2)$

Hiperparâmetro: $\Psi = (\psi_1, \psi_2, \psi_t, \psi_g, \psi_s, \psi_c, \psi_d)$

Gamerman and Moreira (1999): série de Jan/81 a Feb/97

Estratégia amostral (3a. opção): algoritmo MH ao longo das componentes

Cada iteração da cadeia consiste de 7 movimentos propostos para os componentes univariados de Ψ

Cada movimento proposto é dado por $q(x, \cdot) = N(x, c)$

▷ Taxas de aceitação altas podem significar movimentos pequenos e convergência lenta

▷ Taxas de aceitação baixas evidentemente não conduzem a convergência rápida

$\Rightarrow$ Variâncias c dos passeios aleatórios propostos ajustadas para provocar taxas de aceitação em torno de 50%

Quatro cadeias paralelas geradas, partindo de pontos dispersos

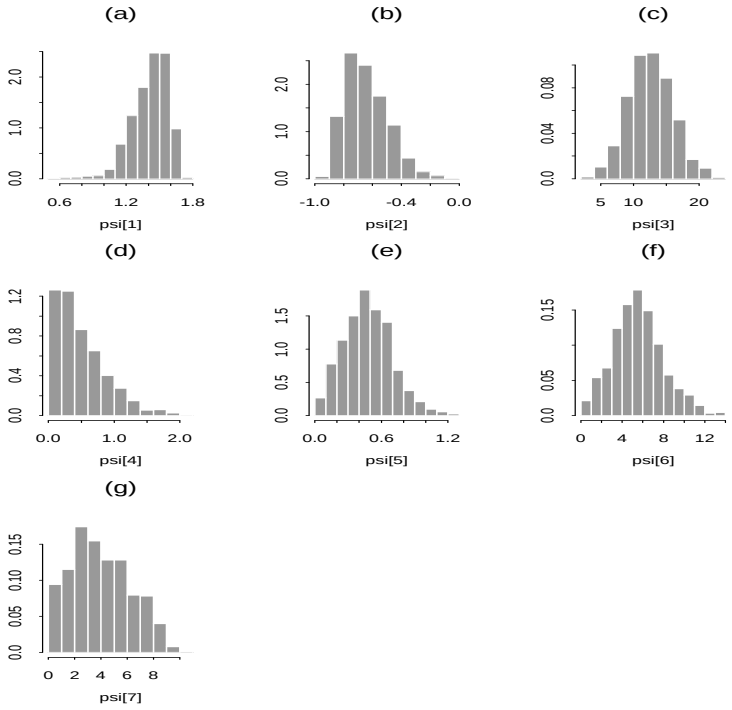
Checagem da convergência:

- Inspeção visual

- Testes de Geweke e de Gelman e Rubin para alguns componentes de Ψ e para $\log \pi(\Psi|y)$

Após equilíbrio: amostras de tamanho 500 de cada cadeia $\rightarrow$ tamanho amostral total $= 2000$

Histogramas dos hiperparâmetros Ψ

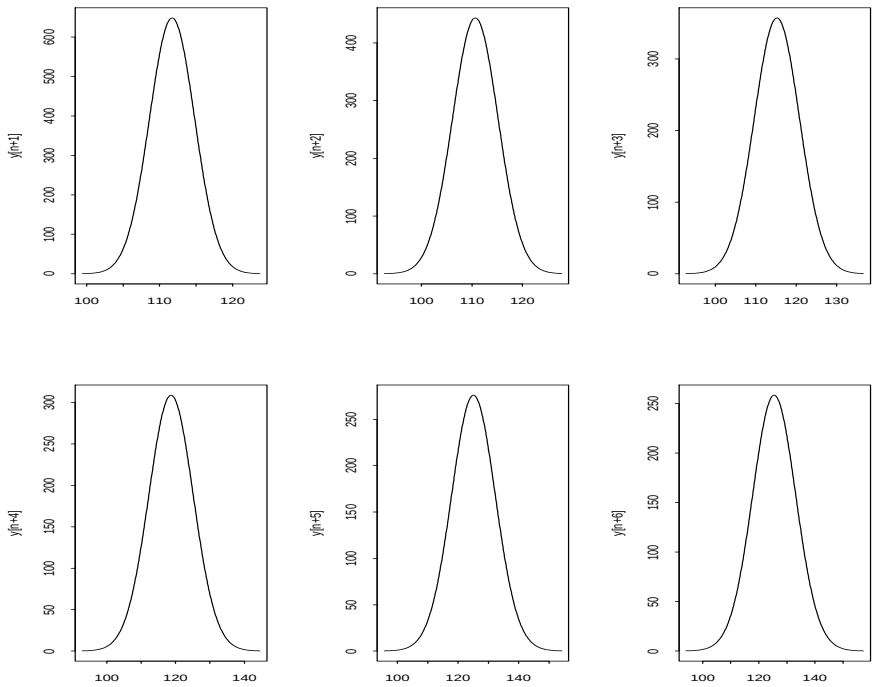


Decaimento dos ciclos concentrado em torno de 1

ψ_s vs ψ_c ou ψ_d: Variabilidade do padrão sazonal menor que a do ciclo ou do regressor

Distribuições preditivas de $y_{n+1}, ..., y_{n+6}$

Padrão Gaussiano, com incerteza crescente para horizontes de previsão crescentes



b. Índice de Produção Agrícola

Estimação de uma função de produção para descrever a produção agrícola em diferentes regiões do Brasil

Assunção, Assunção e Lemos (1998):
estudo do impacto de cada componente de produção sobre a produtividade de uma região

Classificação dos componentes, de acordo com análise de componentes principais: força de trabalho, área trabalhada, insumos mecânicos e insumos bioquímicos

Dados agrupados segundo 558 micro regiões do Brasil

Variável dependente: produção agrícola total da região

Todas as variáveis entram no modelo em logs

Vetor de coeficientes da regressão para região i: contante + 4 coeficientes associados aos 4 fatores de produção

Assunção, Gamerman e Assunção (1999):
Análise com prioris de diferenças de pares e matriz de vizinhança W dada por

$$w_{ij} = \begin{cases} 1/A_i A_j & , \text{ if } i \sim j \\ 0 & , \text{ otherwise} \end{cases},$$

onde A_i é a área da região i, $i = 1, ..., n$.

Outras análises sem priori:
resultados muito pouco suaves espacialmente

Uma única cadeia longa, com GS applicado aos blocos $\beta_1, ..., \beta_{558}$, ϕ e Ψ

5.000 iterações iniciais descartadas e 5.000 valores formam amostra da posteriori dos parâmetros

Padrão consistente com a divisão regional do Brasil:

▷ Áreas mais desenvolvidas do Sudeste mostram maior elasticidade da força de trabalho

▷ Regiões mais pobres do Nordeste exibem menor elasticidade da força de trabalho

▷ Elasticidade da área trabalhada tem comportamento oposto

▷ Saturação dos insumos bioquímicos e mecânicos nas regiões Sul e Sudeste