Econometria Bayesiana

Notas de Aula - 2018-1

Aishameriane Schmidt e Guilherme Moura

Última revisão: 17 de junho de 2018

As notas de aula estão em constante revisão e podem conter erros.
Por favor, envie observações, comentários e correções para
aishameriane arroba gmail ponto com
ou
guilherme ponto moura arroba ufsc ponto br.

Sumário

1 Parte 1 - Motivação e Revisão de Probabilidade

1.1 Introdução

Na[1] estatística bayesiana, a probabilidade é interpretada como o grau de crença na ocorrência de determinado evento. Isso possibilita pensar não apenas nos eventos usualmente trabalhados na estatística frequentista mas também em eventos raros (de ocorrência única, por exemplo) ou ainda contrafactuais (que nem ocorrem). O mesmo não seria possível utilizado a abordagem frequentista, por exemplo, o estimador de máxima verossimilhança (EMV) sofre com dados esparsos.

Exemplo 1.1.1. Considere três lançamentos de uma moeda. Se forem observadas três faces cara, então o estimador de máxima verossimilhança para sair coroa será zero.

Exemplo 1.1.2. Na estatística clássica não faz sentido nos perguntarmos qual a probabilidade de uma pessoa ser "alta" ou "baixa", pois ou a pessoa é alta ou não é (é uma característica intrínseca da pessoa). Essa é a dificuldade da abordagem frequentista, ela não permite modelar eventos únicos. Na bayesiana, como estamos falando de um grau de crença, essa probabilidade será um valor entre $[0 - 1]$.

Na abordagem clássica à inferência, parâmetros são quantidades determinísticas fixas, porém desconhecidas, logo não faz sentido fazer afirmações probabilísticas acerca dos parâmetros. Essa abordagem foi desenvolvida em grande parte no início do século XIX e teve a formalização das ideias a partir dos trabalhos de Karl Pearson, Ronald Fisher e Jerzy Neyman (Paulino et al., 2003). Métodos clássicos de inferência se baseiam na ideia de que existem parâmetros populacionais desconhecidos, mas fixos, que influenciam no processo gerador dos dados. Os dados, por sua vez, serão modelados como variáveis aleatórias que seguem, por hipótese, a forma funcional de uma distribuição de probabilidade conhecida[2] cujos parâmetros (desconhecidos mas fixos!) serão estimados. Sob este prisma, não é possível calcular a probabilidade de um parâmetro θ ser maior do que um determinado valor a, uma vez que θ é constante. Também não é possível fazer inferências ou atribuir probabilidades a eventos que nunca foram observados.

Já na estatística bayesiana, tudo que é desconhecido é tratado como aleatório e, portanto, parâmetros são também considerados variáveis aleatórias. Com isso, todas as regras da inferência bayesiana se baseiam na teoria das probabilidades (Greenberg, 2008). De fato, os métodos bayesianos se ocupam de encontrar as distribuições de probabilidade para os parâmetros e, a partir delas, são realizados os cálculos de interesse. Além disso, comparação de modelos e previsão também são baseadas nas mesmas regras de probabilidade, o que constitui uma vantagem adicional da abordagem bayesiana.

Em economia existem crenças fortes sobre alguns assuntos, como por exemplo, elasticidades serem maiores que 1 em módulo. Essas crenças a priori são herdadas da teoria econômica e a abordagem bayesiana permite incorporá-las de maneira explícita aos modelos.

O método bayesiano consegue incorporar todo o suporte da função de verossimilhança, não apenas seu máximo (como no EMV). Isso implica que as previsões agora podem ser a respeito de uma densidade completa de um parâmetro e não apenas um valor pontual.

Sabemos, da teoria de probabilidade, que a probabilidade de um evento A dado que um evento B ocorreu, é dada por:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{1}$$

[1] Esta aula é baseada no capítulo 1 de Koop (2003).

[2] Os métodos que não assumem nenhuma distribuição subjacente são conhecidos como não paramétricos e não serão vistos na disciplina.

Analogamente, podemos escrever a probabilidade de B ocorrer dado que A ocorreu:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \tag{2}$$

Essas duas expressões combinadas dão origem à forma básica do teorema de Bayes:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3}$$

Note como a expressão acima mostra como usar a informação A para atualizar nosso conhecimento a respeito do evento B. Em econometria estamos interessados em usar dados para aprender algo de interesse do pesquisador. Como em economia trabalha-se muito com modelos, muitas vezes o queremos aprender a respeito de valores de parâmetros. Explorando essa ideia, podemos reescrever (3) em termos do parâmetro θ e dos dados y para obter:

$$\underbrace{\mathbb{P}(\theta|y)}_{\text{densidade a posteriori de } \theta} = \frac{\overbrace{\mathbb{P}(y|\theta)}^{\text{verossimilhança}} \overbrace{\mathbb{P}(\theta)}^{\text{densidade a priori de } \theta}}{\underbrace{\mathbb{P}(y)}_{\text{densid. marginal dos dados}}} \tag{4}$$

Perceba como (3) e (4) são idênticas, mas a formulação (4) explicita o fato de que estamos usando os dados y para aprender a respeito dos parâmetros θ. É importante perceber também que a expressão $\mathbb{P}(\theta|y)$ não faz sentido na abordagem clássica, pois θ é, na abordagem clássica, tratado como uma constante!

Na econometria bayesiana, a densidade a posteriori $\mathbb{P}(\theta|y)$ é de fundamental importância. Ela resume tudo que sabemos a respeito da *variável aleatória* θ após termos observado os dados y. Diferentemente da abordagem clássica, a abordagem bayesiana trata os parâmetros como variáveis aleatórias. Isso é possível, pois ela se baseia na interpretação subjetiva de probabilidades, sob a qual podemos usar as regras de probabilidade para expressar o nosso grau de crença a respeito de qualquer coisa desconhecida.

Como o termo do denominador da equação (4) não depende de θ, podemos simplesmente escrever a posteriori como sendo proporcional ao produto da verossimilhança com a priori:

$$\mathbb{P}(\theta|y) \propto \mathbb{P}(y|\theta)\mathbb{P}(\theta) \tag{5}$$

A distribuição a priori deve resumir todas as informações que o pesquisador possui a respeito de θ antes de observar os dados, como por exemplo, informações de θ obtidas em outros estudos, limites para valores de θ prescritos por alguma teoria, etc.. Por exemplo, suponha que queiramos estimar a participação do capital no produto da economia brasileira, ou seja, o α em uma função Cobb-Douglas. Sem ver os dados brasileiros, sei pelo simples fato de ser a *participação* do capital no produto, que este valor precisa estar entre 0 e 1. Além disso, sei também de estudos para outros países que esse valor costuma ficar entre 0,3 e 0,5. Portanto, a minha distribuição a priori para α deve descrever esse meu conhecimento.

Uma vez bem estabelecida a densidade à posteriori, podemos usá-la para calcular o valor esperado, variância, intervalos, etc. Essa situação é o problema de *estimação*.

Outro problema de interesse para economistas é a *comparação de modelos*. Suponha que tenhamos m modelos disponíveis e gostaríamos de compará-los. Um modelo bayesiano M_i, para $i = 1, \ldots, m$, é formalmente definido como uma função de

verossimilhança vezes uma distribuição a priori. Portanto, a densidade a posteriori para θ calculado com base nos dados y e no modelo M_i é dada por:

$$\mathbb{P}(\theta^i|y, M_i) = \frac{\mathbb{P}(y|\theta^i, M_i)\mathbb{P}(\theta^i|M_i)}{\underbrace{\mathbb{P}(y|M_i)}_{\text{verossimilhança marginal}}} \tag{6}$$

sendo que a expressão para a função de verossimilhança marginal pode ser obtida integrando-se os dois lados de (6) em relação a θ^i, e lembrando que $\int \mathbb{P}(\theta^i|y, M_i)d\theta^i = 1$, o que nos dá:

$$\mathbb{P}(y|M_i) = \int \mathbb{P}(y|\theta^i, M_i)\mathbb{P}(\theta^i|M_i)d\theta^i \tag{7}$$

A lógica da abordagem bayesiana é usar a regra de Bayes para realizar afirmações probabilísticas sobre algo que desconhecemos a partir dos dados que observamos. Portanto, seria interessante derivar a probabilidade do modelo estar corretamente especificado com base nos dados que possuímos. Para isso, usamos a regra de Bayes em (3) com $B = M_i$ e $A = y$ para calcularmos a distribuição a posteriori para cada modelo M_i :

$$\underbrace{\mathbb{P}(M_i|y)}_{\text{Probabilidade do modelo } i|y} = \frac{\mathbb{P}(y|M_i)\mathbb{P}(M_i)}{\mathbb{P}(y)} \tag{8}$$

$\mathbb{P}(M_i)$ é a distribuição a priori do modelo M_i, ou seja, ela descreve toda a informação que o pesquisador tem a respeito da chance do modelo M_i estar corretamente especificado. Se, a priori, não se tem nenhuma indicação a respeito de qual dos m modelos é melhor, faz-se simplesmente $\mathbb{P}(M_i) = 1/m$, dando probabilidade a priori igual para todos os modelos. $\mathbb{P}(y|M_i)$ é a verossimilhança marginal

O cálculo de $\mathbb{P}(y)$ é, na maioria dos casos, difícil, logo, é comum comparar dois modelos através da razão de chances, ou *odds ratio*:

$$\frac{\mathbb{P}(M_1|y)}{\mathbb{P}(M_2|y)} = \frac{\frac{\mathbb{P}(y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y|M_2)\mathbb{P}(M_2)}{\mathbb{P}(y)}} = \frac{\mathbb{P}(y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(y|M_2)\mathbb{P}(M_2)} = PO_{12} \tag{9}$$

Como $\mathbb{P}(y)$ é comum aos dois modelos, esses termos se cancelam e não precisamos calculá-lo!

Como comentado, muitas vezes o pesquisador não tem uma informação a priori a respeito de qual modelo é melhor. Nesse caso, coloca-se a mesma probabilidade a priori para cada um dos dois modelos e, com isso, a razão de a prioris em (9) será igual a 1, dando origem ao que se chama de *fator de bayes* (que é a razão de chances desconsiderando qualquer crença a priori sobre qual o modelo mais plausível):

$$BF_{ij} = \frac{\mathbb{P}(y|M_i)}{\mathbb{P}(y|M_j)} \tag{10}$$

O terceiro problema de interesse está relacionado à *previsão* para novos dados. Seguindo novamente a lógica bayesiana, devemos resumir a nossa incerteza a respeito do que desconhecemos (valores futuros de uma variável) através do uso da regra de Bayes. Ou seja, a previsão deve ser baseada na densidade preditiva $\mathbb{P}(y^*|y)$, sendo y^* os valores futuros de y que se quer prever. Usando regras simples de probabilidade, podemos escrever:

$$\mathbb{P}(y^*|y) = \int \mathbb{P}(y^*, \theta|y)d\theta = \int \mathbb{P}(y^*|y, \theta) \underbrace{\mathbb{P}(\theta|y)}_{posteriori} d\theta \tag{11}$$

A equação (11) é chamada de *densidade preditiva a posteriori* para y^*.

Podemos, então, trabalhar com estimação, comparação de modelos e previsão, incorporando probabilidades, pois a abordagem bayesiana nos permite trabalhar a nossa incerteza em termos aleatórios.

1.2 Revisão de conceitos de probabilidade

1.2.1 Modelo probabilístico

O modelo probabilístico tem dois "ingredientes" básicos: o *espaço amostral* (S) e a *lei de probabilidade*, para falar deles, precisamos entender a noção de *experimento*.

Por experimento podemos entender qualquer atividade cujo resultado final é desconhecido (porém sabemos quais são as possibilidades resultantes dele, apenas não sabemos a priori o que irá sair). Por exemplo, para o lançamento de uma moeda, podemos ter o resultado cara ou coroa e portanto $S = \{(cara), (coroa)\}$. Não sabemos, antes de lançar a moeda, mas sabemos que será[3] *cara* ou *coroa*.

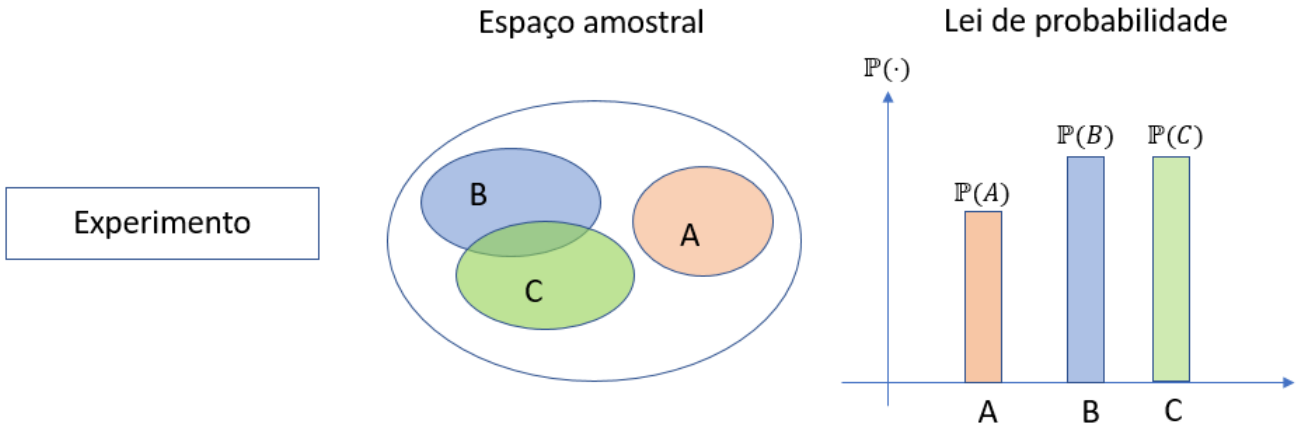


Figura 1: Representação de experimento, espaço amostral e lei de probabilidade.

Definição 1.2.1. Experimento
Qualquer atividade para a qual o resultado final não pode ser especificado de antemão, mas para a qual um conjunto contendo todos os resultados possíveis pode ser indicado.

Definição 1.2.2. Espaço amostral
É o conjunto de todos os resultados possíveis de um dado experimento. Um espaço amostral pode ser discreto ou contínuo, sendo que o primeiro ainda subdivide-se em *finitos* ou *enumeráveis* e os espaços amostrais contínuos serão *não-enumeráveis*.

Definição 1.2.3. Lei de probabilidade
Determina um número não negativo $\mathbb{P}(A)$ a qualquer conjunto de resultados possíveis de A, de forma a resumir o nosso conhecimento ou crença a respeito da chance de ocorrência de A.

Definição 1.2.4. Evento
É um subconjunto do espaço amostral, ou seja, qualquer conjunto de resultados possíveis de S.

Definição 1.2.5. Eventos disjuntos
Dois eventos A e B, $(A, B \subset S)$ são *mutuamente exclusivos* (ou disjuntos) se $A \cap B = \emptyset$. De maneira mais geral, os eventos $A_1, A_2, \ldots$ de S são disjuntos 2 a 2 se $A_i \cap A_J = \emptyset \forall i, j$ onde $i \neq j$.

[3] Estamos abstraindo aqui a possibilidade de outros eventos, como por exemplo, um meteoro cair na terra ou um albatroz pegar a moeda antes de observarmos seu resultado.

Definição 1.2.6. Partição
Se $A_1, A_2 \ldots$, são eventos de S disjuntos 2 a 2 tais que

$$\bigcup_{i=1}^{n} A_i = S$$

Dizemos que a sequência $A_1, A_2, \ldots$ forma uma *partição* de S.

Definição 1.2.7. Espaço de eventos
É o conjunto de todos os eventos de S que inclui $\emptyset$.

Exemplo 1.2.8. Espaço de Eventos
Considere $S = \{1, 2, 3\}$. Então o espaço de eventos de S, denotado por β, será igual a:

$$\beta = \{\emptyset, \underbrace{\{1, 2, 3\}}_{S}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}\}$$

Observe então que os eventos podem ser coleções de resultados possíveis do experimento.

1.2.2 Definição axiomática de probabilidade

Agora que falamos de eventos e do espaço amostral, precisamos definir a lei de probabilidade. Ela atribui uma medida quantitativa para a propensão de ocorrência de cada evento $A \subset S$. Iremos nos apropriar da formulação feita por Kolmogorov.

A medida quantitativa para a propensão de ocorrência de um determinado evento deveria ser um mapa saindo do espaço de eventos indicando um valor entre 0 e 1 para cada um destes eventos de acordo com a chance de ocorrência de cada um deles. Portanto, seria interessante que $\mathbb{P}$ fosse uma função com domínio na família dos subconjuntos de S e imagem no conjunto $[0, 1]$, isto é: gostaríamos que $\mathbb{P}$ fosse tal que $\mathbb{P} : \beta \to [0, 1]$.

Definição 1.2.9. Medida de probabilidade
A função de probabilidade, $\mathbb{P}(\cdot)$, tem domínio em β, espaço de eventos, e domínio em $\mathbb{R}$ e deve atender os seguintes axiomas:

A1. Para qualquer $A \subset S$, $\mathbb{P}(A) \geq 0$ (*axioma da não-negatividade*);

A2. $\mathbb{P}(S) = 1$ (*axioma da normalização*);

- Para o axioma 3, considere as seguintes ideias: se temos dois eventos A e B de S tais que $A \subset B$, seria intuitivo pensar que $\mathbb{P}(A) \leq \mathbb{P}(B)$. Observe ainda que podemos particionar B em A e $B - A$, que são dois conjuntos disjuntos e $B = A \cup (B - A)$. Se $\mathbb{P}(B) > \mathbb{P}(A)$, então o restante da chance de ocorrência de B deve ser dada por $\mathbb{P}(B) - \mathbb{P}(A) \Rightarrow \mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$. Então, a medida de chance de ocorrência precisa ter a propriedade de que para quaisquer eventos A e B, se $A \subset B$, queremos que $\mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(B - A)$. De forma geral, estamos querendo dizer que a medida de probabilidade deve ter a propriedade de que para eventos disjuntos a probabilidade de sua união será a soma das probabilidades.

A3. Seja I um conjunto índice enumerável e seja $\{A_i\}_{i \in I}$ uma coleção de subconjuntos disjuntos de S, então

$$\mathbb{P}(\cup_{i \in I} A_i) = \sum_{i \in I} \mathbb{P}(A_i)$$

(*axioma da aditividade enumerável*)

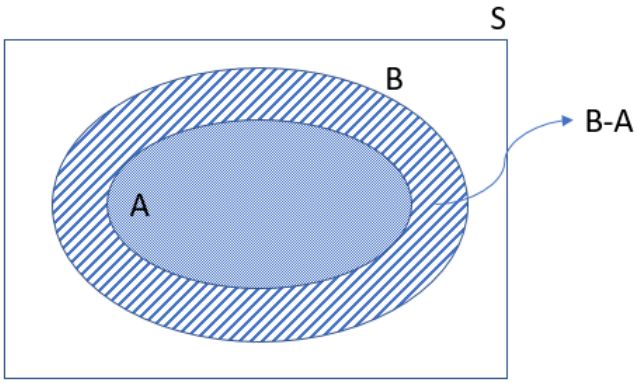


Figura 2: Motivação para o Axioma 3

Qualquer função cujo domínio é o espaço de eventos com contradomínio em $\mathbb{R}$ que atende os axiomas 1, 2 e 3 será chamada de *medida de probabilidade*. Observe que a imagem da função é o intervalo $[0, 1]$ como decorrência dos axiomas. A imagem de um evento $A \subset S$ gerada pela medida de probabilidade $\mathbb{P}(\cdot)$ é chamada de *probabilidade do evento A*.

A tripla $(S, \beta, \mathbb{P}(\cdot))$ é chamada de *espaço de probabilidade* e contém toda a informação necessária para dar probabilidade aos eventos do experimento.

Alguns resultados úteis podem ser demonstrados a partir dos três axiomas.

Lema 1.2.10. ***A probabilidade de nada ocorrer é zero***
$\mathbb{P}(\emptyset) = 0$

Demonstração. Tome $(\{A_n\})_{n\in\mathbb{N}}$ tal que $A_1 = \Omega$ e $A_n = \emptyset \; \forall \; n > 1$. Note que $\cap_{n\in\mathbb{N}} A_n = \emptyset$, portanto:

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

Podemos então escrever:

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \mathbb{P}(\Omega) + \sum_{n>1} \mathbb{P}(\emptyset) \tag{12}$$

Por outro lado, sabemos que $\cup_{n\in\mathbb{N}} A_n = \Omega \cup \emptyset \cup \emptyset \cup \ldots$, de forma que $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \mathbb{P}(\Omega)$. Assim, juntando com 12,

$$\begin{aligned}
\mathbb{P}(\Omega) &= \mathbb{P}(\Omega) + \sum_{n>1} \mathbb{P}(\emptyset) \\
\mathbb{P}(\Omega) - \mathbb{P}(\Omega) &= \sum_{n>1} \mathbb{P}(\emptyset) \\
0 = \sum_{n>1} \mathbb{P}(\emptyset) &= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(\emptyset)
\end{aligned}$$

Note que $\sum_{k=1}^{n} P(\emptyset)$ é não decrescente pois a probabilidade é sempre não negativa. Como o limite é igual a zero e as somas

parciais formam uma sequência não decrescente, as somas finitas são zero. Logo, $\mathbb{P}(\emptyset) = 0$. □

Lema 1.2.11. ***A probabilidade da união é a soma das probabilidades se os eventos forem disjuntos***
Se $A_1, A_2, \ldots, A_n$ são disjuntos, então $\mathbb{P}(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Demonstração. Tome $\{B_n\}_{n\in\mathbb{N}}$ tal que $B_i = A_i \; \forall \; i \in \{1, \ldots, n\}$ e $B_i = \emptyset$ para $i > n$. Por construção, os B_i's são disjuntos e portanto:

$$\begin{aligned}
\mathbb{P}(\cup_{i=1}^{n} A_i) &= \mathbb{P}(\cup_{i\in\mathbb{N}} B_i) \\
&= \sum_{i\in\mathbb{N}} \mathbb{P}(B_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(B_i) + \sum_{k=n+1}^{\infty} \mathbb{P}(B_k) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{k=n+1}^{\infty} \mathbb{P}(\emptyset) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) + 0 \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i)
\end{aligned}$$

Juntando as duas extremidades, segue o resultado desejado. □

Lema 1.2.12. ***Probabilidade do complementar***
Para todo evento A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Demonstração. Sabemos que A e A^c são disjuntos pois $A \cap A^c = \emptyset$. Por outro lado, temos que $A \cup A^c = \Omega$. Então:

$$\mathbb{P}(\Omega) = 1 = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

Segue que $1 = \mathbb{P}(A) + \mathbb{P}(A^c) \Rightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. □

Lema 1.2.13. ***Probabilidade da união***
Para todos os eventos A e B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Demonstração. Note que podemos escrever o conjunto A como a união de dois conjuntos disjuntos:

$$A = (B^c \cap A) \cup (B \cap A)$$

Então:

$$\mathbb{P}(A) = \mathbb{P}[(B^c \cap A) \cup (B \cap A)] = \mathbb{P}(B^c \cap A) + \mathbb{P}(B \cap A)$$

De forma que:

$$\mathbb{P}(B^c \cap A) = \mathbb{P}(A) - \mathbb{P}(B \cap A) \tag{13}$$

Note que B é disjunto de $B^c \cap A$ e que $B \cup (B^c \cap A) = (B \cup B^c) \cap (B \cup A) = \Omega \cap (B \cup A) = (B \cup A)$, portanto,

$$\mathbb{P}(B \cup A) = \mathbb{P}(B) + \mathbb{P}(B^c \cap A) \tag{14}$$

Juntando 13 e 14, temos:

$$\mathbb{P}(B \cup A) = \mathbb{P}(B) + \mathbb{P}(B^c \cap A) = \mathbb{P}(B) + \mathbb{P}(A) - \mathbb{P}(A \cap B)$$

Um resultado que decorre do lema é que $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (pois $\mathbb{P}(A \cap B)$ é sempre não negativa). Observe que a igualdade só ocorre quando os eventos forem disjuntos. □

Lema 1.2.14. ***Probabilidade de subconjuntos***
Se $A \subset B$, então $\mathbb{P}(B) \geq \mathbb{P}(A)$.

Demonstração. Já vimos que $B = (A^c \cap B) \cup (A \cap B)$, que são disjuntos. Então:

$$\mathbb{P}(B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) \tag{15}$$

Mas $A = (A \cap B^c) \cup (A \cap B) \Rightarrow \mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c)$.

Juntando com 15, temos que $\mathbb{P}(B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B^c)$. Mas $A \subset B$, logo, $(A \cap B^c) = \emptyset$, de forma que $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$.

Como $\mathbb{P}(A^c \cap B) \geq 0$, temos $\mathbb{P}(B) \geq \mathbb{P}(A)$. □

Lema 1.2.15. ***Desigualdade de Bonferroni para 2 eventos***

$$\mathbb{P}(A \cap B) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$$

Demonstração. Do lema 1.2.12, sabemos que:

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}[(A \cap B)^c]$$

Pelas leis de DeMorgan, temos $\mathbb{P}[(A \cap B)^c] = \mathbb{P}(A^c \cup B^c)$ e portanto $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c)$.

Do lema 1.2.13, temos $\mathbb{P}(A^c \cup B^c) = \mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c)$, de forma que $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) + \mathbb{P}(A^c \cap B^c)$. Uma vez que as probabilidades são sempre não nulas, segue o resultado desejado. □

Lema 1.2.16. ***Desigualdade de Bonferroni - Caso geral***

$$\mathbb{P}\left(\cap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} \mathbb{P}(A_i^c)$$

Demonstração. A demonstração utiliza a desigualdade do caso bivariado e faz a prova por indução: assuma que é válido para k e vamos provar que vale para $k+1$. Isto é, assuma como verdadeiro

$$\mathbb{P}\left(\cap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} \mathbb{P}(A_i^c)$$

Do lema anterior, vale que:

$$\mathbb{P}\left(\cap_{i=1}^{k+1} A_i\right) = \mathbb{P}\left(\left(\cap_{i=1}^{k} A_i\right) \cap A_{k+1}\right) \geq 1 - \mathbb{P}\left[\left(\cap_{i=1}^{k} A_i\right)^c\right] - \mathbb{P}(A_{k+1}^c)$$

Utilizando a probabilidade do complementar (1.2.12), teremos:

$$\mathbb{P}\left(\cap_{i=1}^{k+1} A_i\right) \geq \mathbb{P}(\cap_{i=1}^{k} A_i) - \mathbb{P}(A_{k+1}^c)$$

E usando a hipótese de indução concluímos que:

$$\mathbb{P}\left(\cap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} \mathbb{P}(A_i^c) - \mathbb{P}(A_{k+1}^c)$$

□

1.2.3 Probabilidade condicional

O único evento *certo* é o próprio S, pois ele incorpora a ideia de que *alguma coisa ocorreu*. Isso está representado no axioma $A2$, da normalização: a probabilidade de algo ocorrer é sempre 1. Agora, considere a seguinte situação: você sabe que um determinado evento $B \subset S$ ocorreu. Isso de certa forma nos restringe em qual região de S estamos interessados, pois tudo aquilo que não tem interseção com B poderá ser descartado. Em outras palavras, se sabemos que o resultado de um experimento é um elemento $B \subset S$, podemos utilizar essa informação para definir o resultado de $S - B$ como irrelevante.

Agora o evento certo passa a ser B e não mais S, de forma que o evento B passa a ser nosso novo *espaço amostral condicional*.

Denotando a probabilidade condicional de A dado que B ocorreu como $\mathbb{P}(A|B)$, teremos que $\mathbb{P}(B|B) = 1$, de forma que A só pode ocorrer se ele ocorre conjuntamente com uma parte de B, sugerindo que $\mathbb{P}(A|B) = \mathbb{P}(A \cap B|B)\ \forall A \subset S$.

Podemos particionar B em dois eventos disjuntos de forma que $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)$. Se $\mathbb{P}(B) \neq 0$, então podemos usá-la para dividir ambos lados da igualdade:

$$\underbrace{\frac{\mathbb{P}(B)}{\mathbb{P}(B)}}_{1} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} \tag{16}$$

Observe que 16 é uma decomposição de chance de ocorrência de B. $\mathbb{P}(A|B)$ será a parte da decomposição relativa à chance do evento $A \cap B$. A probabilidade condicional então será definida como em 1:

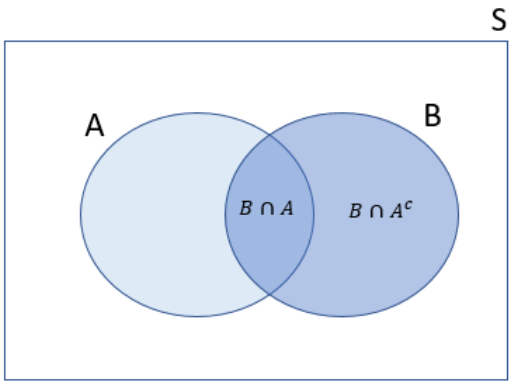


Figura 3: Probabilidades condicionais

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{17}$$

Exemplo 1.2.17. *(Exemplo 1.9 de Bertsekas and Tsitsiklis (2008))*

Se uma aeronave está presente em determinada área, um radar a detecta e aciona um alarme com probabilidade 0.99. Se a aeronave não está na área, o radar tem probabilidade 0.01 de acionar um alarme falso. Assumimos que a probabilidade de uma aeronave estar na região de cobertura do radar é de 0.05. Qual é a probabilidade de que tenhamos um alarme sem que uma aeronave esteja na área? Qual a probabilidade de que a aeronave esteja na região e o alarme não soe?

Podemos definir os seguintes eventos:

A = {(aeronave presente)} com $\mathbb{P}(A) = 0.05$

B = {o radar aciona o alarme}

E os respectivos complementos:

A^c = {(aeronave não está presente)} com $\mathbb{P}(A^c) = 0.95$

B^c = {o radar não aciona o alarme}

Note ainda que: $\mathbb{P}(B|A^c) = 0.1 \Rightarrow \mathbb{P}(B^c|A^c) = 0.9$ e $\mathbb{P}(B|A) = 0.99 \Rightarrow \mathbb{P}(B^c|A) = 0.01$

Podemos agora nos perguntar qual a probabilidade de ter um alarme e a aeronave ao mesmo tempo, sem as condicionais, isto é:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

E também:

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

Essas duas probabilidades são diferentes das que já tínhamos conhecimento: nas probabilidades condicionais nós temos uma informação sobre a aeronave. Agora, nós estamos interessados em uma situação onde o alarme está soando e queremos saber a probabilidade de, ao mesmo tempo, a aeronave estar passando (e por isso a interseção).

Qual é a probabilidade de um alarme falso? Isto é, qual a probabilidade do alarme soar sem que haja uma aeronave passando pelo radar ao mesmo tempo?

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B|A^c)\mathbb{P}(A^c) = 0.1 \cdot 0.95 = 0.095$$

E qual é a probabilidade de não soar alarme e a aeronave estar presente?

$$\mathbb{P}(B^c \cap A) = \mathbb{P}(B^c|A)\mathbb{P}(A) = 0.01 \cdot 0.05 = 0.0005$$

Essas últimas duas probabilidades dizem que o alarme praticamente não deixa de funcionar na presença de uma aeronave ($\mathbb{P}(B^c \cap A)$), pois essa probabilidade é bastante pequena, da ordem de 0.0005. Ao mesmo tempo, ele tem muitas ocorrências de alarmes falsos, representado por $\mathbb{P}(A^c \cap B)$, aproximadamente 10%.

Definição 1.2.18. Regra da multiplicação

$$\mathbb{P}(\cap_{i=1}^{n} A_i) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdot \mathbb{P}(A_4|A_1 \cap A_2 \cap A_3) \cdot \ldots \cdot \mathbb{P}(A_n| \cap_{i=1}^{n-1} A_i)$$

Exemplo 1.2.19. Qual a probabilidade de conseguir 4 ases consecutivos em um baralho de 52 cartas?

Temos 4 naipes e estamos interessados na probabilidade dos 4 ocorrerem simultaneamente (interseção). Logo, a probabilidade desejada é:

$$\mathbb{P}(\cap_{i=1}^{4} A_i) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdot \mathbb{P}(A_4|A_1 \cap A_2 \cap A_3) = \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50}\frac{1}{49} \approx 0.3693 \times 10^{-5}$$

Podemoschecar que de fato $\mathbb{P}(\cdot|B)$ é uma medida de probabilidade. Para isso, verificamos se atende os axiomas $A1$ a $A3$:

(i) $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}(B)} \geq 0 \quad \forall\, A \in \beta$;

(ii) $\mathbb{P}[S|B] = \frac{\mathbb{P}[S \cap B]}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$;

(iii) Se $A_1, A_2, \cdots$ é uma sequência de eventos mutuamente exclusivos em β e $\bigcup_{i=1}^{\infty} A_i \in \beta$, então

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i|B\right] = \frac{\mathbb{P}\left[\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right]}{\mathbb{P}(B)} = \frac{\mathbb{P}\left[\bigcup_{i=1}^{\infty} (A_i \cap B)\right]}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}[A_i|B]$$

Então, $P[\,\cdot\,|B]$ para um dado B que satisfaz $\mathbb{P}(B) > 0$ é uma função de probabilidade, o que justifica chamá-la de probabilidade condicional. $\mathbb{P}[\,\cdot\,|B]$ também apresenta as mesmas propriedades que uma probabilidade não condicionada. Logo, podemos enunciar os seguintes resultados que são similares aos já obtidos para probabilidades não condicionais:

Teorema 1.2.20. *A probabilidade condicional do vazio é zero, isto é,* $\mathbb{P}[\,\emptyset\,|B] = 0$.

Demonstração.

$$\mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = \frac{0}{\mathbb{P}(B)} = 0$$

□

Teorema 1.2.21. *Se $A_1, A_2, \cdots, A_n$ são eventos mutuamente exclusivos em β, então*

$$\mathbb{P}[A_1 \cup \cdots \cup A_n|B] = \sum_{i=1}^{n} \mathbb{P}[A_i|B]$$

Demonstração.

$$\begin{aligned}
\mathbb{P}[A_1 \cup \cdots \cup A_n|B] &= && \text{(definição prop. condicional)} \\
&= \frac{\mathbb{P}[(A_1 \cup \cdots \cup A_n) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)} \\
&= \frac{\mathbb{P}[(A_1 \cap B) \cup \cdots \cup (A_n \cap B)]}{\mathbb{P}(B)} && \text{(independência dos } A_i\text{'s)} \\
&= \frac{\mathbb{P}\left[\sum_{i=1}^{n}(A_i \cap B)\right]}{\mathbb{P}(B)} && \text{(definição prob. condicional)} \\
&= \sum_{i=1}^{n} \mathbb{P}[A_i|B]
\end{aligned}$$

□

Teorema 1.2.22. *Se A é um evento em β, então*

$$\mathbb{P}[A^c|B] = 1 - \mathbb{P}[A|B]$$

onde A^c é o evento complementar de A.

Demonstração.

$$\begin{aligned}
\mathbb{P}[A^c|B] &= && \text{(def. prob. condicional)} \\
&= \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} && \text{(forma alternativa de } \mathbb{P}(B)\text{)} \\
&= \frac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B)}{\mathbb{P}(B)} - \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= 1 - \mathbb{P}[A|B]
\end{aligned}$$

□

Teorema 1.2.23. *Se A_1 e A_2 pertencem a β, então*

$$\mathbb{P}[A_1|B] = \mathbb{P}[A_1 \cap A_2|B] + \mathbb{P}[A_1 \cap A_2^c|B]$$

Demonstração.

$$
\begin{aligned}
\mathbb{P}[A_1|B] = & && \text{(def. prob. cond.)}\\
&= \frac{\mathbb{P}[A_1 \cap B]}{\mathbb{P}(B)} && \text{(reescrevendo } A_1\text{)}\\
&= \frac{\mathbb{P}[((A_1 \cap A_2) \cup (A_1 \cap A_2^c)) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B \cup (A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(independência)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B] + \mathbb{P}[(A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(reorganizando)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B]}{\mathbb{P}(B)} + \frac{\mathbb{P}[(A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(def. prob. cond.)}\\
&= \mathbb{P}[A_1 \cap A_2|B] + \mathbb{P}[A_1 \cap A_2^c|B]
\end{aligned}
$$

□

Teorema 1.2.24. *Para quaisquer dois eventos A_1 e $A_2 \in \beta$,*

$$\mathbb{P}[A_1 \cup A_2|B] = \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] - \mathbb{P}[A_1 \cap A_2|B]$$

Demonstração.

$$
\begin{aligned}
\mathbb{P}[A_1 \cup A_2|B] = & && \text{(def. prob. cond.)}\\
&= \frac{\mathbb{P}[(A_1 \cup A_2) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)}\\
&= \frac{\mathbb{P}[(A_1 \cap B) \cup (A_2 \cap B)]}{\mathbb{P}(B)} && \text{(def. prob. união.)}\\
&= \frac{\mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) - \mathbb{P}(A_1 \cap B \cap A_2 \cap B)}{\mathbb{P}(B)} && \text{(rearranjando)}\\
&= \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} - \frac{\mathbb{P}(A_1 \cap A_2 \cap B)}{\mathbb{P}(B)} && \text{(def. prob. cond.)}\\
&= \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] - \mathbb{P}[A_1 \cap A_2|B]
\end{aligned}
$$

□

Teorema 1.2.25. *Se A_1 e $A_2 \in \beta$ com $A_1 \subset A_2$, então*

$$\mathbb{P}[A_1|B] \leq \mathbb{P}[A_2|B]$$

Demonstração. Esse resultado decorre do lema 1.2.14. □

1.2.4 Teorema da probabilidade total e teorema de Bayes

Da aula passada, discutimos as probabilidades de interseção de dois eventos em termos da probabilidade condicional:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{18}$$

Ao mesmo tempo, podemos escrever a equação em termos da probabilidade inversa:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{19}$$

Combinando 18 com 20, temos a regra de Bayes para o caso simples:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{20}$$

Teorema 1.2.26. ***Teorema da probabilidade Total:***

Sejam $A_1, A_2, \ldots, A_n$ *eventos disjuntos que formam uma partição do espaço amostral e suponha que* $\mathbb{P}(A_i) > 0 \ \forall \ i \in \{1, \ldots, n\}$. *Então, para qualquer evento B, temos:*

$$\mathbb{P}(B) = \mathbb{P}(A_1 \cap B) + \ldots + \mathbb{P}(A_n \cap B) \tag{21}$$

Utilizando a regra da multiplicação, podemos reescrever cada uma das interseções como $P(B|A_i)\mathbb{P}(A_i)$*:*

$$\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \ldots + \mathbb{P}(B|A_n)\mathbb{P}(A_n) \tag{22}$$

A figura 4 ilustra a ideia do teorema:

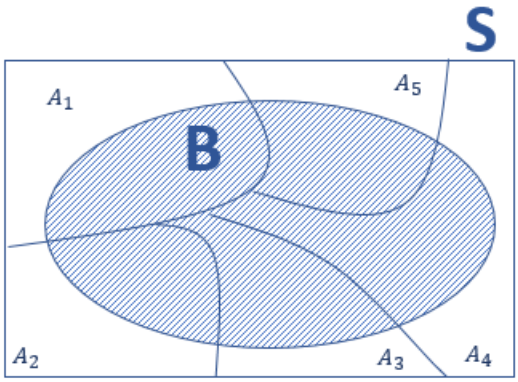


Figura 4: O espaço amostral S particionado pelos eventos $A_1, \ldots, A_5$ e o evento $B \subset S$

Teorema 1.2.27. ***Regra de Bayes***

Sejam $A_1, \ldots, A_n$ *uma partição do espaço amostral e B um evento qualquer. Então, para cada* $i \in \{1, \ldots, n\}$, *temos:*

$$\mathbb{P}(A_i|B) = \frac{(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \tag{23}$$

Exemplo 1.2.28. Exemplo 1.18 de Mittelhammer (2013)

Um teste sanguineo desenvolvido por uma indústria farmacêutica para detectar determinada doença consegue detectar a presença da doença, com 98% de acurácia, dado que o indivíduo esteja de fato infectado. O teste leva a um resultado "falso positivo[4]" para apenas 1% das pessoas não doentes que são testadas.

[4]Falso positivo é quando o teste diz que um indivíduo sadio tem a doença.

Se uma pessoa escolhida aleatoriamente da população é testada para a doença, dado que existe .1% de probabilidade dela de fato estar doente, qual a probabilidade de que a pessoa esteja doente se o resultado do teste foi positivo para a doença?

Nste caso, seja A o evento que indica que o teste deu positivo e B o evento que indica que a pessoa tem a doença. Então, do enunciado, temos:

1. $\mathbb{P}(A|B) = .98$
2. $\mathbb{P}(B) = .001$
3. $\mathbb{P}(A|B^c) = .01$

Utilizando a fórmula de bayes, temos:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{.98 \cdot .001}{.98 \cdot .001 + .01 \cdot .999} - .89$$

Neste caso, o teste é pouco confiável para verificar se a pessoa está de fato doente.

A ideia do teorema é que após sabermos a ocorrência de B podemos "revisar" a chance de ocorrência de todos os pedaços do espaço amostral. Quanto maior a interseção de B com um determinado A_i, maior será a probabilidade condicional desse A_i ocorrer. Na figura 5 temos que a partição funciona como se fosse a nossa probabilidade a priori. Após observar B, temos uma atualização do nosso espaço de maneira que as probabilidades de ocorrência de cada partição se alteram, como na figura 6.

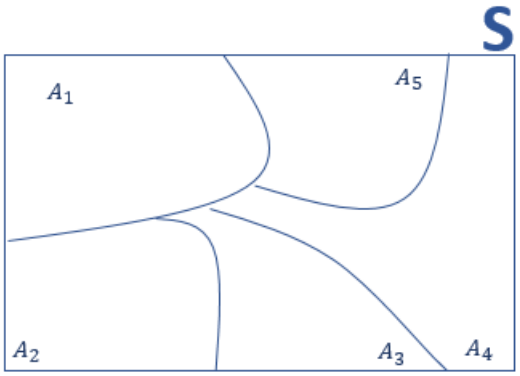


Figura 5: Espaço amostral original

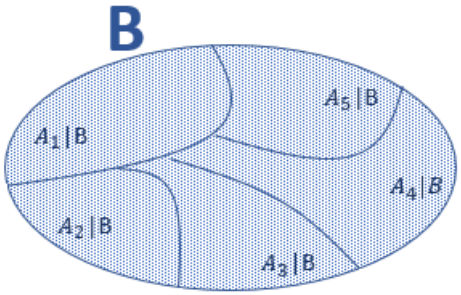


Figura 6: Novo espaço amostral após observar B

Posto de outra maneira, a ideia de que o espaço amostral é particionado em conjuntos disjuntos A_i, e que a ocorrência de um outro evento B com probabilidade conhecida nos permite "revisar", ou atualizar as probabilidades de ocorrência de cada uma das partes A_i, foi o que levou à interpretação de que a informação a respeito de A_i é atualizada à luz da nova informação a respeito da ocorrência de B. Ou seja, $\mathbb{P}(A_i)$ é a probabilidade do evento A_i *antes* de observarmos o evento B e por isso é chamada de *probabilidade à priori* e $\mathbb{P}(A_i|B)$ é a probabilidade de A_i posterior à ocorrência de B.

Independência

A ideia que a probabilidade condicional traz é que ao observar B conseguimos melhorar a nossa informação a respeito de cada um dos A_i. Um caso importante é quando o evento B não traz nenhuma informação a respeito dos eventos A_i. É um caso especial no qual B não traz informação relevante nenhuma a respeito de A. Formalmente, temos que $\mathbb{P}(A|B) = \mathbb{P}(A)$. Podemos desenvolver melhor a conta de forma que:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A) \tag{24}$$

Usualmente utilizamos como definição de independência $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Esse fato vem da ideia de que a probabilidade condicional deve ser a própria probabilidade do evento original. Essa é a definição mais utilizada pois está bem definida mesmo quando $\mathbb{P}(B) = 0$.

Note que dois eventos disjuntos jamais serão independentes: quando dois eventos A e B são disjuntos e sabemos que A ocorreu, temos toda a informação necessária sobre B, neste caso que ele não ocorre.

Definição 1.2.29. Independência Condicional

Seja C um evento qualquer, então dizemos que A e B são condicionalmente independentes se:

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B \cap C) \tag{25}$$

Podemos utilizar a definição de independência não condicional com a regra da multiplicação para obter:

$$\begin{aligned}
\mathbb{P}(A \cap B|C) &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} \quad \text{(prob. cond.)} \\
&= \frac{\mathbb{P}(C)\mathbb{P}(B \cap C)\mathbb{P}(A|B \cap C)}{\mathbb{P}(C)} \quad \text{(regra mult.)} \\
&= \mathbb{P}(B|C)\mathbb{P}(A|B \cap C)
\end{aligned}$$

Igualando as duas pontas da equação acima, temos:

$$\begin{aligned}
\mathbb{P}(A|C)\mathbb{P}(B|C) &= \mathbb{P}(B|C)\mathbb{P}(A|B \cap C) \\
\mathbb{P}(A|C) = \frac{P(B|C)\mathbb{P}(A|B \cap B)}{\mathbb{P}(B|C)} &= \mathbb{P}(A|B \cap C)
\end{aligned}$$

De maneira que $\mathbb{P}(A|C) = \mathbb{P}(A|B \cap C)$.

Note que isso não significa que A e B são independentes. Caso o C seja desconhecido, a nossa probabilidade é alterada. É possível que a informação que C tenha para A seja a mesma que B agregaria e por isso que, condicionado a C, a informação de B para A não seja relevante. Assim, *independência condicional não implica independência incondicional.*

1.2.5 Variáveis aleatórias

Em muitas aplicações, os eventos do espaço amostral não são números reais e a notação começa a ficar pesada e pouco prática. É interessante então termos uma maneira de escrever esses eventos como subconjuntos do $\mathbb{R}^n$. Este processo de relacionar valores reais com resultados de experimentos é feito através da noção de variável aleatória. Dado um experimento e seu respectivo espaço amostral S, uma variável aleatória associa um número particular com cada um dos resultados de S, isto é, $X : S \to \mathbb{R}$, onde X é uma *variável aleatória univariada*.

Definição 1.2.30. Para um dado espaço de probabilidade $(S, \beta, P(\cdot))$ uma variável aleatória, denotada por X ou $X(\cdot)$, é uma função com domínio em S e contradomínio em $\mathbb{R}$, como exemplificado na figura 7.

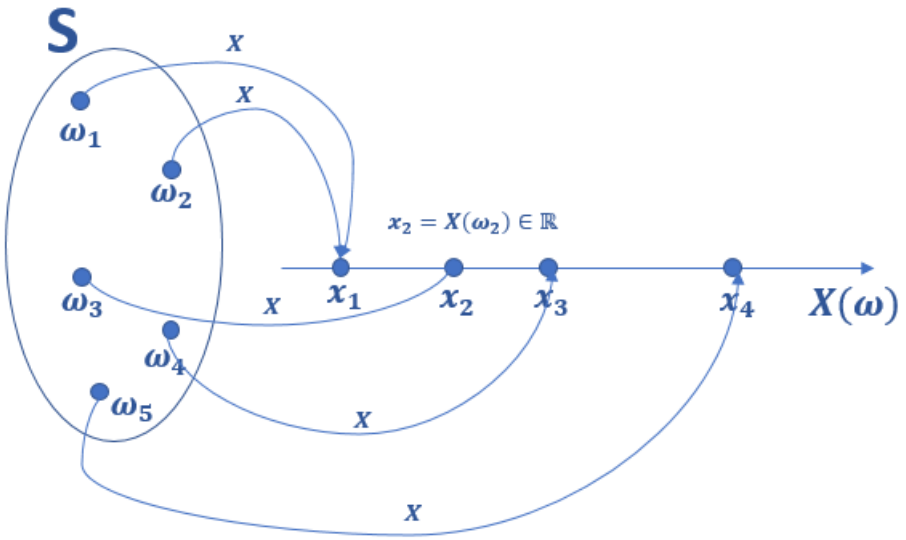


Figura 7: Representação da variável aleatória X levando elementos de S para a reta real

O resultado de uma variável aleatória é o elemento da imagem de X gerado pela observação do resultado do experimento. Ou seja, se o resultado do experimento foi $w_i \in S$, então o resultado da v.a. é $x_i = X(w_i)$.

Definição 1.2.31. Imagem de uma v.a. A imagem de uma v.a. representa a transformação do espaço amostral original para um espaço amostral com valores reais. Formalmente, a imagem de uma v.a. é definida por:

$$R(X) = \{x \in \mathbb{R} : x = X(w), w \in S\}$$

1.2.6 O espaço de probabilidade definido por uma v.a.

Um novo espaço de probabilidade será necessário para poder estabelecer probabilidades a subconjuntos do novo espaço amostral real definido pela imagem da v.a.. O espaço de probabilidade $\{S, \beta, \mathbb{P}(\cdot)\}$ nos permite determinar a probabilidade para eventos em S, porém qual é a probabilidade de que um resultado da v.a. X fique no subconjunto $A \subset R(X)$?

Como X é um mapa de S para a reta real, é possível definir o evento B em S tal que o evento B ocorre se e somente se $A \subset R(X)$ ocorre. Uma vez que os eventos A e B ocorrem simultaneamente, a probabilidade deles deve ser a mesma, ou seja, $\mathbb{P}_X(A) \equiv \mathbb{P}(B)$, onde $\mathbb{P}_X(\cdot)$ denota a medida de probabilidade que atribui a resultados de X suas probabilidades. Se dois eventos ocorrem sempre simultaneamente, eles são ditos equivalentes e ocorrem em espaços de probabilidades distintos, pois se ocorressem no mesmo espaço, eles seriam o mesmo evento. Logo, $\mathbb{P}_X(A) \equiv \mathbb{P}(B)$ para $B = \{w : X(w) \in A, w \in S\}$. A Figura (8) ilustra esta relação.

Probabilidades definidas para eventos em S são transferidas para eventos em $R(X)$ através da relação funcional que define uma v.a., $x_i = X(w_i)$. Então, sabendo que o domínio de $\mathbb{P}(\cdot)$ é β (espaço de eventos), qual é o domínio de $\mathbb{P}_X(\cdot)$? Podemos dizer informalmente que β_X é o espaço de eventos do espaço de probabilidade associado à variável aleatória X e é dado por todos os subconjuntos da imagem de X, $R(X)$.

Então, o espaço de probabilidade definido por X é $(R(X), \mathcal{B}_X, \mathbb{P}_X(\cdot)$. Lidar com espaços amostrais reais é muito mais conveniente pois nos permite utilizar os conhecimentos matemáticos.

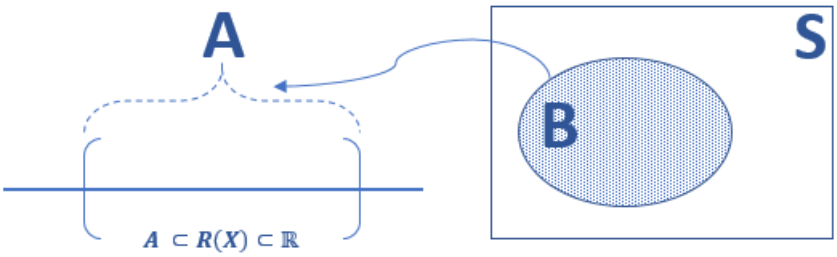


Figura 8: Correspondência entre $B \subset S$ e $A \subset R(X)$.

Exemplo 1.2.32. Exemplo 2.1 de Mittelhammer (2013)
Seja $S = \{1, 2, 3, \ldots, 10\}$ o número de carros potencialmente vendidos por um vendedor em determinada semana e seja $\mathcal{B}$ o conjunto de todos subconjuntos de S. Defina a função de probabilidade $\mathbb{P}(\gamma) = (1/55)\sum_{\omega\in\gamma}\omega$ para $\gamma \in \mathcal{B}$. Suponha que o vendedor recebe semanalmente um salário base de \$100 mais uma comissão de \$100 para cada carro vendido. O salário semanal do vendedor pode ser representado por uma variável aleatória $X(\omega) = 100 + 100\omega$, para $\omega \in S$. O espaço de probabilidade induzido $(R(X), \mathcal{B}_X, \mathbb{P}_X)$ é então caracterizado por $R(X) = \{200, 300, \ldots, 1100\}$, $\mathcal{B}_X = \{A : A \subset R(X)\}$, e $\mathbb{P}_X(A) = (1/55)\sum_{\omega\in\gamma}\omega$ para $\gamma = \{\omega : (100 + 100\omega) \subset A, \omega \in S\}$. Então, por exemplo, o evento onde o vendedor consegue um salário menor que \$300 em uma semana, $A = \{200, 300\}$, tem probabilidade $\mathbb{P}_X(A) = (1/55)\sum_{\omega\in\{1,2\}}\omega = 3/55$.

Função densidade de probabilidade

Quando o espaço amostral real $R(X)$ é enumerável, então qualquer subconjunto de $R(X)$ pode ser representado como a união de elementos específicos de $R(X)$. Ou seja, se $A \subset R(X)$, então A pode ser escrito como:

$$A = \bigcup_{x\in A}\{x\}$$

Como os elementos de $R(X)$ são disjuntos, o axioma 3 nos dá:

$$\mathbb{P}_X(A) = \sum_{x\in A}\mathbb{P}(\{x\}) = \sum_{x\in A}\sum_{\omega\in S}\mathbb{P}(\{\omega \in S : X(\omega) = x\}) \tag{26}$$

Isso sugere uma função $f : R(X) \to \mathbb{R}$ que nos dá a probabilidade de cada elemento de $R(X)$. Uma vez que $f(x)$ foi definida, $\mathbb{P}_X$ pode ser redefinida como $\mathbb{P}_X(A) = \sum_{x\in A} f(x)$, agora sem fazer referência à elementos w do espaço amostral original S. Observe que mesmo que o contradomínio de $f(\cdot)$ seja a reta real, os axiomas definidos para $\mathbb{P}(\cdot)$ implicarão que necessariamente sua imagem será sempre o intervalo $[0, 1]$.

Definição 1.2.33. Variável aleatória discreta
Uma variável aleatória é dita *variável aleatória discreta* se sua imagem é um conjunto enumerável.

Definição 1.2.34. Função de probabilidade discreta
A função densidade de probabilidade discreta (ou função de probabilidade), denotada por $f(\cdot)$, é definida por:

$$f(x) \equiv \begin{cases} \text{probabilidade de } x \; \forall \; x \in R(X) \\ 0 \; \forall x \notin R(X) \end{cases}$$

Se tivermos uma v.a. definida em um subconjunto não-contável, ou não enumerável de $\mathbb{R}$, então 26 não será definida.

Porém, integração em relação a conjuntos não enumeráveis é possível, sugerindo a seguinte definição:

$$\mathbb{P}_X(A) = \int_{x \in A} f(x)dx \tag{27}$$

Exemplo 1.2.35. Exemplo 2.2 de Mittelhammer (2013)

Seja o experimento que consiste em jogar dois dados equilibrados e observar a face virada para cima de cada um deles. Sejam i e j a face de cada um dos dados, respectivamente. O espaço amostral será dado por $S = \{(i, j) : i \text{ e } j \in \{1, 2, 3, 4, 5, 6\}\}$. Agora defina a variável aleatória $x = X((i, j)) = i + j$ para $i, j \in S$. Então a seguinte correspondência pode ser feita entre os resultados de X, eventos de S e as respectivas probabilidades:

	$X(w) = x$	$B_x = \{w : X(w) = x, w \in S\}$	$f(x) = P(B_x)$
$R(X)$	2	{(1,1)}	1/36
	3	{(1,2), (2,1)}	2/36
	4	{(1,3), (2,2), (3,1)}	3/36
	5	{(1,4), (2,3), (3,2), (4,1)}	4/36
	6	{(1,5), (2,4), (3,3), (4,2), (5,1)}	5/36
	7	{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}	6/36
	8	{(2,6), (3,5), (4,4), (5,3), (6,2)}	5/36
	9	{(3,6), (4,5), (5,4), (6,3)}	4/36
	10	{(4,6), (5,5), (6,4)}	3/36
	11	{(5,6), (6,5)}	2/36
	12	{(6,6)}	1/36

Figura 9: Representação do Exemplo (1.2.35) - Retirado de Mittelhammer (2013)

A imagem da v.a. é $R(X) = \{2, 3, \ldots, 12\}$, que representa a coleção de imagens dos pontos $(i, j) \in S$ pela função $x = X((i, j)) = i + j$. Probabilidades dos possíveis resultados de X são dados por $f(x) = \mathbb{P}(\mathcal{B}_X)$, onde $\mathcal{B}_X$ é a coleção de imagens inversas de x.

Se desejamos obter a probabilidade de um evento $x \in A = \{7, 11\}$, então $\mathbb{P}_X(A) = \sum_{x \in A} f(x) = f(7) + f(11) = \frac{8}{36}$. Se $A = \{2\}$, então teremos $\mathbb{P}_X(A) = \sum_{x \in A} f(x) = f(2) = \frac{1}{36}$.

Definição 1.2.36. Variável aleatória contínua

Uma variável aleatória é dita *contínua* se o seu domínio é (1) não-enumerável e (2) se existe uma função não-negativa $f(x)$ definida para todo $x \in (-\infty, +\infty)$ de forma que para qualquer evento $A \subset R(X)$ temos:

$$\mathbb{P}_X(A) = \int_{x \in A} f(x)dx$$
$$f(x) \equiv 0 \; \forall \; x \notin R(X)$$

A função $f(x)$ é chamada de *função densidade de probabilidade contínua*.

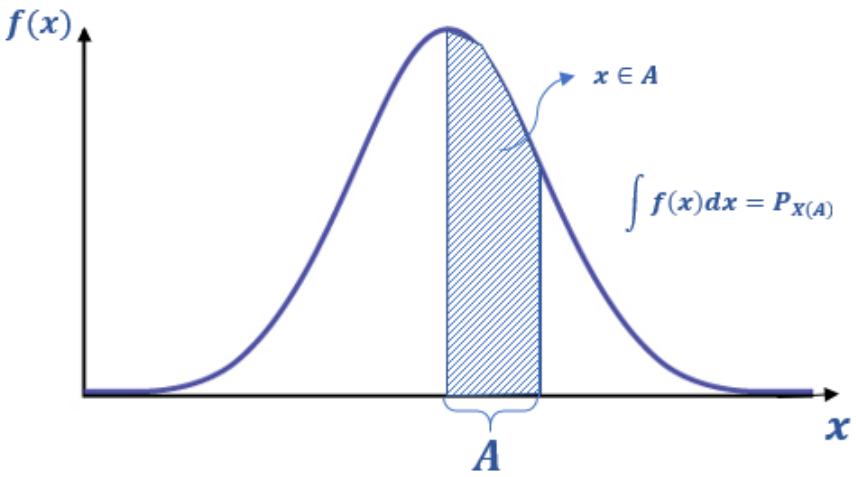


Figura 10: Exemplo de $f(x)$ para uma v.a. contínua X.

Enquanto $f(x)$ do caso discreto nos dá a probabilidade de ocorrência de x, no caso contínuo isso não ocorre, pois demandaria que $f(x) = 0 \forall x \in R(X)$, dado que a probabilidade de eventos elementares é 0 (em decorrência de ser calculado via uma integral) (veja figura 10).

1.2.7 Função distribuição acumulada

A f.d.a. nos dá a probabilidade de que uma variável aleatória assuma um valor menor do que um número real específico, isto é, ela nos dá a probabilidade do evento $\{x : x \leq b, x \in R(X)\}$ para um valor $b \in \mathbb{R}$.

Definição 1.2.37. Função distribuição acumulada
A função de distribuição acumulada de uma v.a. X é definida por $F(b) \equiv \mathbb{P}(X \leq b)\ \forall b \in \mathbb{R}$.

- **Caso discreto:** $F(b) = \sum_{x \leq b} f(x),\ b \in \mathbb{R}$;
- **Caso contínuo:** $F(b) = \int_{-\infty}^{b} f(x)dx,\ b \in \mathbb{R}$.

Teorema 1.2.38. *Sejam $f(x)$ e $F(x)$ a f.d.p. e a f.d.a. de uma v.a. contínua X. A f.d.p. de X pode ser definida como:*

$$f(x) = \frac{d}{dx}[F(x)]$$

Sempre que $f(x)$ for contínua e $f(x) = 0$ nas outras partes.

Demonstração. A prova envolve o teorema fundamental do cálculo. □

1.2.8 Variáveis aleatórias multivariadas

Considere duas variáveis aleatórias discretas, X e Y, associadas ao mesmo experimento. As probabilidades que X e Y podem assumir são descritas pela função densidade de probabilidade conjunta de X e Y:

$$f(X, Y) = \mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x, Y = y) \tag{28}$$

A f.d.p. conjunta de X e Y determina a probabilidade de qualquer evento que possa ser especificado em termos de X e Y.

$$\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} f(X,Y) \tag{29}$$

Pode-se calcular a f.d.p. somente de X ou de Y através da f.d.p. conjunta de X e Y:

$$f(x) = \sum_{y} f(x,y) \quad f(y) = \sum_{x} f(x,y)$$

As funções acima são conhecidas como *funções de probabilidade marginais* de X e Y, respectivamente.

Exemplo 1.2.39. Seja $f(x_1, x_2)$ a f.d.p. conjunta e $R(X)$ a imagem da v.a. bivariada discreta $X = (X_1, X_2)$. Suponha agora que queiramos designar probabilidade ao evento $X_1 \in B$. Qual evento para a v.a. bivariada é equivalente a B para a v.a. univariada X_1?

Ovento de interesse, equivalente ao B, é dado por $A = \{(x_1, x_2) : x_1 \in B \; e \; (x_1, x_2) \in R(X)\}$. Logo, A e B são equivalentes pois não há como A ocorrer sem que $x_1 \in B$ e não há como B ocorrer sem que A ocorra, de maneira que suas probabilidades são as mesmas:

$$\mathbb{P}_{X_1}(B) = \mathbb{P}_{X_1,X_2}(\{X_1 \in B\}) = \mathbb{P}_{X_1,X_2}(A) = \sum_{(X_1,X_2)\in A} f(x_1, x_2)$$

Como $f(x_1, x_2) = 0 \; \forall (x_1, x_2) \notin R(X)$, temos:

$$\mathbb{P}_{X_1}(B) = \sum_{X_1 \in B} \sum_{X_2 \in R(X_2)} f(x_1, x_2)$$

A função densidade de probabilidade marginal de X_1 é dada por:

$$f_{X_1}(x_1) = \sum_{X_2 \in R(X_2)} f(x_1, x_2)$$

Caso as variáveis fossem contínuas, a função densidade de probabilidade marginal contínua:

$$\mathbb{P}_{X_1}(B) = \mathbb{P}_{X_1,X_2}(\{X_1 \in B\}) = \mathbb{P}_{X_1,X_2}(A) = \int\limits_{(X_1,X_2)\in A} f(x_1, x_2)dX$$

Como $f(x_1, x_2) = 0 \; \forall \; (x_1, x_2) \notin R(X)$, podemos reescrever como uma integral dupla:

$$\mathbb{P}_{X_1}(B) = \int\limits_{X_1 \in B} \int\limits_{-\infty}^{+\infty} f(x_1, x_2)dx_1dx_2 \Rightarrow f_{x_1}(X_1) = \int\limits_{-\infty}^{+\infty} f(x_1, x_2)dx_2$$

1.2.9 Função densidade de probabilidade condicional

Suponha que conheçamos o espaço de probabilidade correspondente a um experimento envolvendo o resultado da variável aleatória n-dimensional $X_{(n)} = (X_1, \ldots, X_m, X_{m+1}, \ldots, X_n)$ mas estamos interessados em dar probabilidades para o evento

$(X_1, \ldots, X_m) \in C$, condicionado ao fato de que $(X_{m+1}, X_{m+2}, \ldots, X_n) \in D$.

Exemplo 1.2.40. Seja $f(x_1, x_2)$ a f.d.p. conjunta e $R(X)$ a imagem da v.a. $X = (X_1, X_2)$. *O evento para a variável aleatória bivariada que é equivalente ao evento $X_1 \in C$ para a v.a. X_1 é dado por*[5] $A = \{(x_1, x_2) : x_1 \in C, (x_1, x_2) \in R(X)\}$. Analogamente, o evento para a variável aleatória bivariada X que é equivalente ao evento D para a variável univariada X_2 pe dado por $B = \{(x_1, x_2) : x_2 \in D, (x_1, x_2) \in R(X)\}$.

Logo, a probabilidade de $X_1 \in C$, dado que $X_2 \in D$ pode ser definido como:

$$\mathbb{P}_{X_1|X_2}(C|D) = \mathbb{P}_{X_1|X_2}(\{x_1 \in C\}|\{x_2 \in D\}) = \mathbb{P}(A|B) = \frac{A \cap B}{\mathbb{P}(B)}, \text{ se } \mathbb{P}(B) \neq 0 \tag{30}$$

Sendo que $A \cap B = \{(x_1, x_2) : x_1 \in C,\ x_2 \in D,\ (x_1, x_2) \in R(X)\}$.

No caso discreto, temos:

$$\mathbb{P}_{X_1|X_2}(C|D) = \mathbb{P}(A|B) = \frac{\overbrace{\sum_{(x_1,x_2)\in A\cap B} f(x_1, x_2)}^{\mathbb{P}(A\cap B)}}{\underbrace{\sum_{(x_1,x_2)\in B} f(x_1, x_2)}_{\mathbb{P}(B)}} = \underbrace{\frac{\sum_{x_1\in C}\sum_{x_2\in D} f(x_1, x_2)}{\sum_{x_2\in D}\sum_{x_1\in R(X)} f(x_1, x_2)}}_{\text{marginal de } x_2} = \sum_{x_1\in C}\left[\underbrace{\frac{\sum_{x_2\in D} f(x_1, x_2)}{\sum_{x_2\in D} f_2(x_2)}}_{\text{f.d.p condicional}}\right] \tag{31}$$

Ou seja, o miolo sempre se mantém, o que vai mudar é onde estamos somando "do lado de fora":

$$\mathbb{P}(W|D) = \sum_{(x_1,x_2)\in W}\left[\frac{\sum_{x_2\in D} f(x_1, x_2)}{\sum_{x_2\in D} f_2(x_2)}\right]$$

No caso de uma v.a. contínua, teremos:

$$\mathbb{P}_{X_1|X_2}(C|D) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \text{ para } \mathbb{P}(B) \neq 0$$

Isto é,

$$\mathbb{P}_{X_1|X_2}(C|D) = \frac{\int_{(x_1,x_2)\in A\cap B} f(x_1, x_2)dX}{\int_{(x_1,x_2)\in B} f(x_1, x_2)dX} = \frac{\int_{x_1\in C}\int_{x_2\in D} f(x_1, x_2)dx_1dx_2}{\int_{x_2\in D}\underbrace{\int_{-\infty}^{\infty} f(x_1, x_2)dx_1}_{\textbf{marginal de } X_2} dx_2} \tag{32}$$

E assim:

$$\mathbb{P}_{X_1|X_2}(C|D) = \int_{x_1\in C}\left[\frac{\int_{x_2\in D} f(x_1, x_2)dx_2}{\int_{x_2\in D} f_2(x_2)dx_2}\right]dx_1 \tag{33}$$

[5] revisar a frase em itálico

1.2.10 Esperança de uma v.a.

O conceito de esperança de uma variável aleatória pode ser motivado tanto através do conceito de média ponderada, quanto pelo conceito físico de centro de gravidade.

Exemplo 1.2.41. Suponha que uma haste seja colocada em um suporte (fulcro). Suponha também que um peso de 10 kg seja colocado a meio metro do ponto de suporte, enquanto um peso de 5 kg seja colocado a um metro do ponto de suporte como na Figura 11 abaixo:

Figura 11: Representação do sistema de pesos do exemplo 1.2.41

Denotando o ponto de suporte por $\delta = 0$, o ponto onde está localizado o peso de 10 kg por $x_1 = 0,5$, o ponto onde colocou-se o peso de 5 kg por $x_2 = -1$ e a massa colocada em um ponto x qualquer como massa(x), podemos calcular os momento físico de qualquer massa colocada em qualquer ponto x da haste. Mais especificamente, o momento físico do ponto x é definido como o produto da massa naquele ponto pela distância até o ponto de suporte:

$$\text{momento} = \text{massa}(x)(x - \delta).$$

Logo, temos:

- $\text{momento}_{10kg} = 10(1/2 - 0) = 5$
- $\text{momento}_{5kg} = 5(-1 - 0) = -5$

Um sistema de alavanca como este estará em equilíbrio se a soma dos momentos for zero, o que é conhecido como o *momento total do sistema*:

$$\sum_{i=1}^{n} \text{massa}(x_i)(x_i - \delta).$$

O Conceito físico de momento deu origem ao conceito estatístico de momento e pode ser usado para identificar o ponto no qual uma função densidade de probabilidade se equilibra.

Exemplo 1.2.42. Suponha que o experimento em questão seja o mesmo do exemplo 1.2.35 (rolar de dois dados). Sejam j e i os valores observado na face de cada um dos dados. Então, o espaço amostral é dado por $S = \{(i, j) : i \text{ e } j \in \{1, 2, 3, 4, 5, 6\}\}$. Definindo uma variável aleatória $x = X(i, j) = i + j \ \forall \ (i, j) \in S$, podemos então determinar as probabilidades de cada resultado possível da v.a. X.

A imagem de X é $R(X) = \{2, 3, 4, \ldots, 12\}$. As probabilidades de cada um dos resultados possíveis de X é dada pela f.d.p. de X:

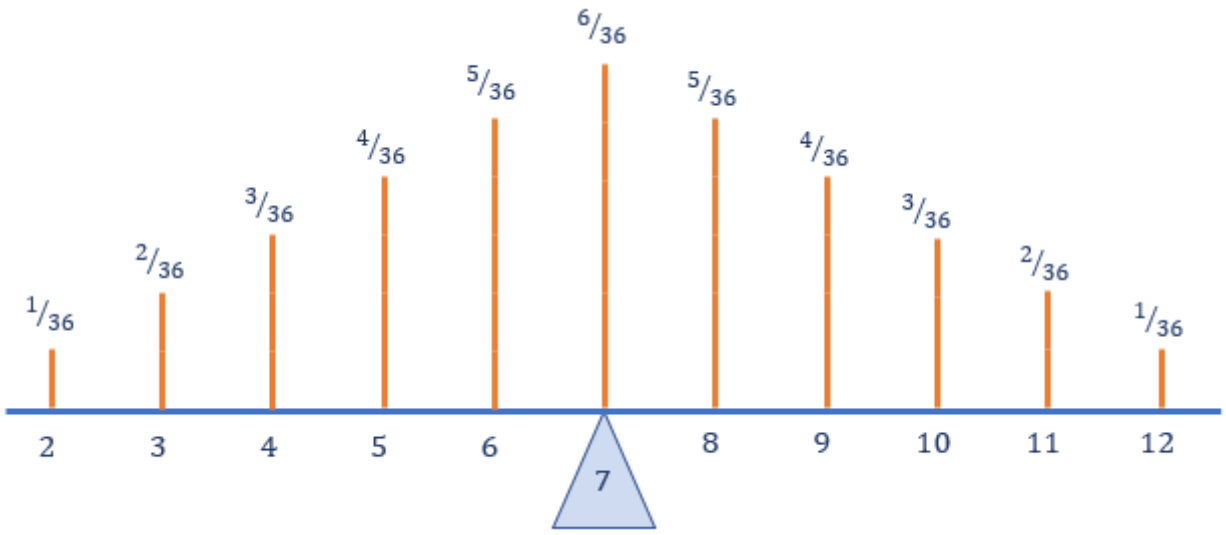


Figura 12: Representação das massas de probabilidade dos resultados do experimento do exemplo 1.2.42

Supondo que as probabilidades funcionassem como os pesos do exemplo anterior, em qual ponto devemos colocar nosso suporte para que a f.d.p. se equilibre? Vamos calcular o momento total do sistema:

$$\sum_{i=1}^{11} \text{massa}(x_i)(x_1 - \delta) = \sum_{i=1}^{11} f(x_i)(x_i - \delta) \tag{34}$$

A equação acima, para que o momento total seja zero, nos dá o valor de δ que equilibra o sistema, de forma que:

$$\sum_{i=1}^{11} f(x_i)(x_i - \delta) = 0 \Rightarrow \sum_{i=1}^{11} f(x_i)(x_i) = \delta \underbrace{\sum_{i=1}^{11} f(x_i)}_{1} \Rightarrow \delta = \sum_{i=1}^{11} f(x_i)(x_i) = 7 \tag{35}$$

Definição 1.2.43. O valor esperado de uma v.a. discreta é definido como:

$$\mathbb{E}(X) = \sum_{x \in R(X)} x f(x) \tag{36}$$

Dado que essa soma exista. Perceba que a equação acima implica que a esperança de uma variável aleatória discreta nada mais é que uma média ponderada dos valores de x por suas probabilidades.

Para o caso contínuo, temos:

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x) dx \tag{37}$$

Dado que a integral acima seja convergente.

2 Parte 2 - Introdução à Estatística Bayesiana

Já vimos que o tratamento bayesiano da informação amostral é caracterizado pela regra de Bayes:

$$\underbrace{\mathbb{P}(\theta|y)}_{\text{a posteriori}} = \frac{\overbrace{\mathbb{P}(y|\theta)\ \mathbb{P}(\theta)}^{\text{a priori}}}{\underbrace{\mathbb{P}(y)}_{\text{constante em relação a } \theta}} \tag{38}$$

com $\mathbb{P}(y) = \int_{\Theta} \mathbb{P}(y|\theta)d\theta$, onde Θ é o espaço de eventos para a variável aleatória θ.

A equação (38) deixa claro que todo o conhecimento a respeito de θ na distribuição a posteriori é determinado pelo produto entre $\mathbb{P}(y|\theta)$ e $\mathbb{P}(\theta)$, pois o denominador de (38) não depende de θ e serve apenas como constante de integração para garantir que a densidade $\mathbb{P}(\theta|y)$ integre em 1. Logo, podemos reescrever a posteriori como proporcional ao produto da verossimilhança e priori:

$$\mathbb{P}(\theta|y) \propto \mathbb{P}(y|\theta) \cdot \mathbb{P}(\theta) \tag{39}$$

Como $\mathbb{P}(\theta)$ caracteriza a informação a priori e é não-amostral, toda a informação amostral relevante para a inferência a posteriori está contida em $\mathbb{P}(y|\theta)$ que, vista como função de *theta*, é conhecida como função de verossimilhança.

Definição 2.0.1. Função de Verossimilhança
Para qualquer amostra aleatória y, a função $\mathbb{P}(y|\theta) = f(\theta|y)$ vista como função de θ é chamada de *função de verossimilhança*.

Definição 2.0.2. Princípio da Verossimilhança
Ao fazermos inferência a respeito de θ após observarmos o vetor de dados y, toda a informação amostral relevante está contida na função de verossimilhança $f(\theta|y)$. Isto significa que a função de verossimilhança tem toda a informação dos dados que é relevante para fazer inferências sobre θ.

O princípio da verossimilhança deixa explícito que apenas os dados observados são relevantes para fazermos inferência sobre θ, e outras realizações de $y^i \in S$ que poderiam ter ocorrido são irrelevantes. Observe que isto não é válido na abordagem clássica, que se interessa, por exemplo, no procedimento de amostragem para o cálculo de p-valores. Mais detalhes podem ser obtidos em http://www2.isye.gatech.edu/~brani/isyebayes/bank/handout2.pdf e no capítulo 2 de Bauwens et al. (2003).

Exemplo 2.0.3. Lançamento de uma moeda (Adaptado de Greenberg (2008))
Considere o experimento de lançar uma moeda. O espaço amostral do lançamento de uma moeda é $S = \{\text{cara, coroa}\}$. O espaço de eventos é dado por $\mathcal{B} = \{\emptyset, S, \text{cara, coroa}\}$. A medida de probabilidade associada ao experimento será:

$$\mathbb{P}(\omega) = \begin{cases} \theta, \text{ se } \omega = \text{cara} \\ 1 - \theta, \text{ se } \omega = \text{coroa} \end{cases}$$

onde ω é um elemento qualquer de S e θ é o parâmetro que determina a chance do resultado do lançamento ser cara ou coroa. O espaço de probabilidades do evento então será dado por $(S, \mathcal{B}, \mathbb{P}(\cdot))$.

Como é pouco prático ficar lidando com o espaço amostral S, que contém elementos não numéricos, vamos definir uma variável aleatória $Y(\omega)$ que mapeie os elementos de S para algum subconjunto de $\mathbb{R}$:

$$Y(\omega) = \begin{cases} 1, \text{ se } \omega = \text{ cara} \\ 0, \text{ se } \omega = \text{ coroa} \end{cases}$$

A variável aleatória $Y(\omega)$ define um novo espaço amostral dado pela imagem da função $Y(\omega)$, ou seja, $R(Y) = \{1, 0\}$. Note que ao utilizar a variável aleatória, passamos de um espaço amostral não numérico para um espaço amostral numérico, o que facilita a análise quantitativa do problema. Além disso, o novo espaço amostral $R(Y)$ define um novo espaço de eventos $\mathcal{B}_Y = \{\emptyset, R(Y), 1, 0\}$. O evento *cara* e o evento 1 para a variável aleatória Y são *eventos equivalentes*, porém em *espaços de probabilidade* distintos. Isso implica que a probabilidade de ocorrência destes dois eventos serão iguais. Portanto, para completarmos o espaço de probabilidade de Y, precisamos definir a medida de probabilidade $\mathbb{P}_Y$. Para isso, exploramos o fato dos eventos equivalentes para fazer com que P_Y seja definida como: $\mathbb{P}_Y(1) = \mathbb{P}(\text{cara}) = \theta$ e $\mathbb{P}_Y(0) = \mathbb{P}(\text{coroa}) = 1 - \theta$.

Entretanto, será interessante definir uma função que leve os elementos da imagem de Y até subconjuntos da reta real sem necessariamente passarmos pelos eventos de S, isto é, queremos definir a função $p : R(Y) \to \mathbb{R}$ que nos dê a probabilidade de ocorrência de cada um dos elementos de $R(Y)$ sem fazer referência ao espaço amostral inicial, S. Essa função, como já vimos, é chamada de *função densidade de probabilidade*, que neste caso específico será dada por:

$$\mathbb{P}(Y = y|\theta) = \begin{cases} \theta, \text{ se } y = 1 \\ 1 - \theta, \text{ se } y = 0 \end{cases}$$

Podemos reescrever a função acima como:

$$\mathbb{P}(Y = y|\theta) = \theta^y(1 - \theta)^{1-y} \tag{40}$$

Para verificar que (40) de fato representa a probabilidade de Y, substitua Y por 1 e 0.

Agora estamos prontos para fazer inferências a respeito de θ.

Se fizermos n lançamentos independentes da moeda, então a função densidade de probabilidade conjunta dos n lançamentos é dada por:

$$\begin{aligned} \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\theta) = & \\ &= \mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = y_2\} \cap \ldots \cap \{Y_n = y_n\}|\theta) \\ &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\theta) = \mathbb{P}(Y_1 = y_1|\theta) \cdot \mathbb{P}(Y_2 = y_2|\theta) \ldots \cdot \mathbb{P}(Y_n = y_n|\theta) \\ &= \theta^{y_1}(1 - \theta)^{1-y_1} \ldots \theta^{y_n}(1 - \theta)^{1-y_n} \\ &= \prod_{i=1}^{n} \theta^{y_i}(1 - \theta)^{1-y_i} \end{aligned}$$

E obtemos:

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\theta) = \theta^{\sum y_i}(1 - \theta)^{n-\sum y_i} \tag{41}$$

Observação: $\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\theta) = \mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = y_2\} \cap \ldots \cap \{Y_n = y_n\}|\theta)$ e como os eventos são independentes, a probabilidade da interseção é o produto das probabilidades, de forma que temos $\mathbb{P}(Y_1 = y_1, Y_2 =$

$y_2, \ldots, Y_n = y_n|\theta) = \mathbb{P}(Y_1 = y_1|\theta) \cdot \mathbb{P}(Y_2 = y_2|\theta) \ldots \cdot \mathbb{P}(Y_n = y_n|\theta)$.

Após observarmos n lançamentos, (41) é uma função apenas do parâmetro desconhecido θ. $\mathbb{P}(Y_1, Y_2, \ldots, Y_n|\theta)$ vista como função de θ é chamada de função de verossimilhança e é fundamental na inferência clássica e também na bayesiana. $\sum y_i$ é chamada de *estatística suficiente*. Toda a informação que precisamos da amostra para fazer inferência para θ está condensada nesta quantidade - pois o *princípio da verossimilhança* estabelece que toda a informação dos dados está contida na função de verossimilhança (Bauwens et al., 2003). Note, porém, que a forma como a informação é utilizada é diferente se formos clássicos ou bayesianos.

Faz diferença obter 6 caras em 10 lançamentos dependendo como o experimento é arranjado, se formos clássicos. Caso o experimento seja "lançar a moeda até obter 6 caras" e interrompemos na décima jogada (onde tivemos a sexta cara) ou se o experimento for "jogar a moeda 10 vezes" e observamos 6 caras. A informação da amostra é a mesma (6 caras em 10 lançamentos), porém para o clássico isso é tratado diferente e irá influenciar em quantidades como p-valores (uma vez que calculamos a probabilidade de observar os dados sob a hipótese nula e para cada experimento a hipótese nula é diferente). Para o Bayesiano, a organização do experimento é irrelevante.

É importante perceber que a função de verossimilhança **não** é uma função densidade de probabilidade para θ, ou seja, sua integral ou somatório em relação a θ é diferente de 1. Entretanto, sua integral em relação a y, ou soma, nesse caso específico de uma variável aleatória discreta, é sim igual a 1.

Após observarmos os n lançamentos, podemos usar $\underline{y} = \{y_i\}_{i=1}^n$ para estimar θ. Na abordagem clássica, a função de verossimilhança $f(\theta|y)$ definida por (41) será maximizada em relação a θ, de forma a encontrarmos $\hat{\theta}_{ML}$ que maximiza a chance de ocorrência de $\underline{y}$. No entanto, o seu logaritmo é uma função mais suave e é uma transformação monotônica e portanto $\max_\theta\{\mathbb{P}(\underline{y}|\theta)\} = \max_\theta\{\log \mathbb{P}(\underline{y}|\theta)\}$. Logo, podemos trabalhar com o log de (41), ou seja:

$$l(\theta|y) = log\left(\mathbb{P}(y|\theta)\right) = \left(\sum_{i=1}^{n} y_i\right) log(\theta) + \left(n - \sum_{i=1}^{n} y_i\right) log(1-\theta) \tag{42}$$

Diferenciando (42) em relação a θ e igualando a 0:

$$\begin{aligned} \frac{dl(\theta)}{d\theta}\Big|_{\hat{\theta}_{ML}} &= \frac{\sum y_i}{\hat{\theta}_{ML}} - \frac{n - \sum y_i}{1 - \hat{\theta}_{ML}} = 0 \\ \Rightarrow \sum y_1(1 - \hat{\theta}_{ML}) &= \left(n - \sum y_i\right)\hat{\theta}_{ML} \end{aligned}$$

De forma que temos:

$$\hat{\theta}_{ML} = \frac{\sum y_i}{n} = \bar{Y} \tag{43}$$

Na abordagem bayesiana podemos incluir informações que não estão contidas na base de dados às nossas estimativas, através da determinação de uma priori. Isto é, a abordagem bayesiana utiliza, além da função de verossimilhança dada em (41), também uma distribuição a priori[6] para θ. Esta distribuição caracteriza o conhecimento prévio respeito desse valor desconhecido.

Para realizar inferência bayesiana a respeito de θ precisamos, além da função de verossimilhança[7], de uma distribuição

[6]Iremos ver mais adiante que não necessariamente precisa ser uma densidade, então estamos fazendo um abuso de linguagem aqui.
[7]Observe que a *função de verossimilhança* (Equação 41) **não** é a mesma coisa que o *estimador de máxima verossimilhança* (Equação 43).

a priori para θ que caracterize nosso conhecimento prévio a respeito dessa variável aleatória. Uma coisa que sabemos a respeito de θ é que $0 \leq \theta \leq 1$, uma vez que se trata de uma probabilidade. Portanto, podemos usar uma f.d.p. *Beta* como distribuição a priori, pois sabemos que o suporte da distribuição *Beta* é tal que $\theta \in [0, 1]$. A distribuição *Beta* tem dois parâmetros, α e β, ambos positivos e é dada por:

$$\mathbb{P}(\theta|\underline{\alpha},\underline{\beta}) = \frac{\theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\beta}-1}}{\int_0^1 \theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\beta}-1}} = \frac{\Gamma(\underline{\alpha}-1+\underline{\beta})}{\Gamma(\underline{\alpha})\Gamma(\underline{\beta})}\theta^{\underline{\alpha}}(1-\theta)^{\underline{\beta}-1} \tag{44}$$

onde $\Gamma(\cdot)$ é a função Gama e $0 \leq \theta \leq 1$. $\underline{\alpha}$ e $\underline{\beta}$ são os parâmetros da distribuição a priori que representam nosso conhecimento prévio, e são muitas vezes chamados de *hiperparâmetros* para se diferenciar dos parâmetros de interesse. A média e a variância de uma variável aleatória $\theta \sim Beta(\underline{\alpha},\underline{\beta})$ são:

$$\mathbb{E}(\theta) = \frac{\underline{\alpha}}{\underline{\alpha}+\underline{\beta}} \quad \text{e} \quad Var(\theta) = \frac{\underline{\alpha}\underline{\beta}}{(\underline{\alpha}+\underline{\beta})^2(\underline{\alpha}+\underline{\beta}+1)} \tag{45}$$

Note que, ao variar $\underline{\alpha}$ e $\underline{\beta}$, podemos caracterizar diferentes crenças a respeito de θ. A figura 13 apresenta quatro representações diferentes para a distribuição $Beta(\alpha,\beta)$, onde é possível ver o efeito da variação dos parâmetros. Sempre que $\alpha = \beta$, teremos uma distribuição com média $0,5$. Valores maiores de α em relação a β irão resultar em uma curva com cauda à esquerda e o contrário (quando $\beta > \alpha$) resulta em uma distribuição com cauda para a esquerda. Um caso particular da distribuição está no terceiro gráfico, onde vemos que $Beta(1,1) = \text{Uniforme}(0,1)$. Apesar dos três primeiros gráficos apresentarem distribuições *Beta* com médias iguais a $0,5$, elas têm implicações muito distintas em relação à possível crença a respeito de θ. Por exemplo, quando $\alpha = \beta = 0,5$, a média da distribuição é $0,5$, mas θ tem mais chances de assumir valores extremos, próximos de 0 ou 1, do que próximos da média de $0,5$. Por outro lado, se $\alpha = \beta = 10$, então a θ tem mais chances de assumir valores próximos de $0,5$, e descreveria a situação na qual tem-se conhecimento a priori de que a moeda tem maior chance de ser "justa". Quando $\alpha = \beta = 1$, todas as regiões do intervalo $[0, 1]$ têm a mesma chance de ocorrência, e esta distribuição pode ser usada para caracterizar ausência de conhecimento a priori. Já o caso $\alpha = 5$ e $\beta = 2$ pode descrever uma situação na qual se tem informação a priori que a moeda é viesada em direção ao lado contendo cara.

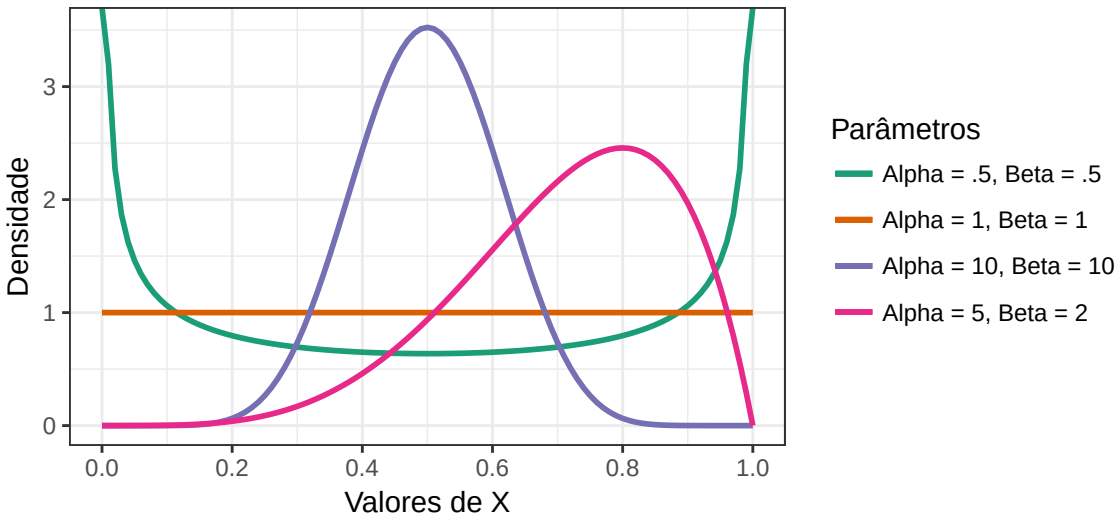


Figura 13: Diferentes formas da distribuição $Beta(\alpha,\beta)$

Outra vantagem da escolha de uma priori Beta, neste exemplo, é que ela é uma priori *conjugada* ao modelo Binomial. A priori conjugada é aquela que combinada com a verossimilhança produz uma posteriori da mesma família que a priori. Logo, quando temos uma verossimilhança Binomial (que é o caso de 41) e a combinamos com uma a priori Beta (equação 44), temos uma distribuição a posteriori que também é Beta, porém com outros parâmetros. Isto é, a posteriori está na mesma família da priori. Esse processo facilita o processo de atualização da posteriori quando mais dados estiverem disponíveis.

Utilizando (38), (41) e (44), temos:

$$\mathbb{P}(\theta|y) = \frac{\overbrace{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}}^{\text{eq. (41)}} \overbrace{\theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\beta}-1}\frac{\Gamma(\underline{\alpha}+\underline{\beta})}{\Gamma(\underline{\alpha})\Gamma(\underline{\beta})}}^{\text{eq. (44)}}}{\int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\theta^{\underline{\alpha}-1}(1-\theta)^{\underline{\beta}-1}\frac{\Gamma(\underline{\alpha}+\underline{\beta})}{\Gamma(\underline{\alpha})\Gamma(\underline{\beta})}d\theta} = \frac{\overbrace{\mathbb{P}(Y|\theta)\mathbb{P}(\theta)}^{\text{eq (38)}}}{\int_0^1 \mathbb{P}(Y|\theta)\mathbb{P}(\theta)d\theta}$$

De forma que:

$$\mathbb{P}(\theta|y) = \frac{\theta^{(\overbrace{\underline{\alpha}+\Sigma y_i}^{\bar{\alpha}})-1}(1-\theta)^{(\overbrace{\underline{\beta}+n-\Sigma y_i}^{\bar{\beta}})-1}}{\int_0^1 \theta^{(\underline{\alpha}+\sum y_i)-1}(1-\theta)^{(\underline{\beta}+n-\sum y_i)-1}\,d\theta} \tag{46}$$

O numerador (46) é o *núcleo* de uma distribuição *Beta* com parâmetros $\bar{\alpha} = (\underline{\alpha}+\sum y_i)$ e $\bar{\beta} = (\underline{\beta}+n-\sum y_i)$. O denominador de (46), nada mais é do que a integral do numerador e é constante em relação a θ, pois todo o efeito de θ foi integrado para fora. Por isso, o inverso do denominador de (46) é conhecido como *constante de integração* da densidade. Portanto, a forma da distribuição é completamente definida pelo *núcleo* expresso no numerador, enquanto o denominador é apenas uma constante que garante com que a integral da densidade seja igual a 1. Para entender esse ponto, perceba que a densidade a posteriori precisa, como toda densidade, integrar 1, ou seja, $\int \mathbb{P}(\theta|y)\,d\theta \equiv 1$. Integrando os dois lados de (46), temos

$$\int \mathbb{P}(\theta|y)\,d\theta = 1 = \int \frac{\theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}}{\int_0^1 \theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}\,d\theta}\,d\theta = \frac{\int \theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}\,d\theta}{\int_0^1 \theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}\,d\theta} = 1 \tag{47}$$

Portanto, (46) é uma densidade $Beta(\bar{\alpha},\bar{\beta})$ e podemos encontrar o valor da integral do denominador sem usar regras de cálculo integral, simplesmente observando que $\int \mathbb{P}(\theta|y)\,d\theta \equiv 1$. Com isso, da fórmula da distribuição *Beta* em (44), sabemos que:

$$\frac{\Gamma(\bar{\alpha}+\bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})}\int \theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}\,d\theta = 1 \tag{48}$$

Multiplicando os dois lados de (48) por $\frac{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})}{\Gamma(\bar{\alpha}+\bar{\beta})}$, temos:

$$\int \theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1}\,d\theta = \frac{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})}{\Gamma(\bar{\alpha}+\bar{\beta})} \tag{49}$$

O que nos dá a forma funcional da distribuição a posteriori $Beta(\bar{\alpha},\bar{\beta})$:

$$\mathbb{P}(\theta|y) = \frac{\Gamma(\bar{\alpha}+\bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})}\theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\beta}-1} \tag{50}$$

Note que os parâmetros da distribuição a posteriori $\bar{\alpha} = \underline{\alpha} + \sum y_i$ e $\bar{\beta} = \underline{\beta} + n - \sum y_i$ são os parâmetros da distribuição a priori, $\underline{\alpha}$ e $\underline{\beta}$, somados ao número de caras e coroas observados na amostra. Isso permite a interpretação de que os parâmetros da priori são o números de caras e coroas que obtivemos no "experimento" que baseia o nosso conhecimento prévio.

A média da posteriori pode ser calculada utilizando:

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \int_0^1 \theta \cdot \mathbb{P}(\theta|y)d\theta \\
&= \frac{\overbrace{\underline{\alpha} + \sum y_i}^{\bar{\alpha}}}{\underbrace{\underline{\alpha} + \sum y_i}_{\bar{\alpha}} + \underbrace{\underline{\beta} + n - \sum y_i}_{\bar{\beta}}} \\
&= \frac{\underline{\alpha} + \sum y_i}{\underline{\alpha} + \underline{\beta} + n} = \frac{\underline{\alpha}}{\underline{\alpha} + \underline{\beta} + n} + \frac{\sum y_i}{\underline{\alpha} + \underline{\beta} + n}
\end{aligned}$$

Que podemos reescrever como:

$$\mathbb{E}[\theta|y] = \underbrace{\frac{\underline{\alpha}}{\underline{\alpha} + \underline{\beta}}}_{\mathbb{E}[\theta]}\left[\frac{\underline{\alpha} + \underline{\beta}}{\underline{\alpha} + \underline{\beta} + n}\right] + \underbrace{\frac{\sum y_i}{n}}_{\hat{\theta}_{ML}}\left[\frac{n}{\underline{\alpha} + \underline{\beta} + n}\right] \tag{51}$$

Note que a forma como $\mathbb{E}[\theta|y]$ está escrita na equação (51) indica que a média de θ, após observar os dados, nada mais é que uma média ponderada entre a média da priori, $\mathbb{E}[\theta]$, e o estimador de máxima verossimilhança, $\hat{\theta}_{ML}$. Os pesos desta média são justamente $\underline{\alpha} + \underline{\beta}$ para a priori e o tamanho amostral n para a componente dos dados. Se o tamanho amostral for muito superior à soma dos hiperparâmetros, então a média da posteriori será praticamente igual ao estimador de máxima verossimilhança. De forma geral:

$$\lim_{n\to\infty}[\mathbb{E}(\theta|y)] = \hat{\theta}_{ML}$$

Suponha que observemos a seguinte sequência: $\underset{\sim}{Y} = \{1, 1, 1\}$. Baseado no estimador de máxima verossimilhança, temos:

$$\hat{\theta}_{ML} = \frac{\sum y_i}{n} = \frac{1+1+1}{3} = \frac{3}{3} = 1$$

Logo, a probabilidade de ocorrência de coroa estimada por máxima verossimilhança é 0.

Sob o enfoque bayesiano, usando uma priori não informativa $Beta(1, 1) = \text{Uniforme}(0, 1)$, temos:

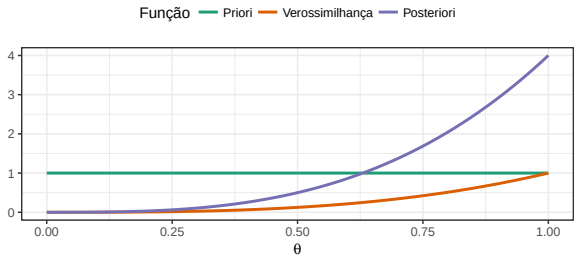
$$\begin{aligned}
\bar{\alpha} &= (\underline{\alpha} + \sum y_i) = (1+3) = 4 \quad , \quad \bar{\beta}(1+3-3) = 1 \\
\mathbb{E}[\theta|y] &= \frac{\bar{\alpha}}{\bar{\alpha}+\bar{\beta}} = \frac{4}{4+1} = \frac{4}{5} = 0.8
\end{aligned}$$

Logo, a estimativa bayesiana para a probabilidade de ocorrência de coroa é 0.2.

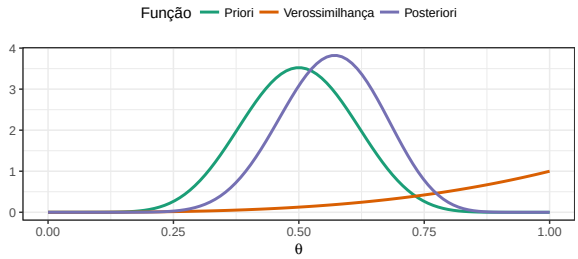
Observe que, pelo enfoque frequentista, o fato de nunca temos observado uma coroa, implica que sua chance de ocorrência é zero! O estimador de MV atribui probabilidade 0 aos eventos que não ocorreram na amostra. Já para a estimativa Bayesiana, por mais que não tenhamos observado nenhuma coroa, o número de repetições observadas ainda não é suficiente para dizer que é impossível obter uma coroa.

Esse problema é conhecido como *excesso de ajuste* do estimador de máxima verossimilhança: ele tenta escolher os parâmetros de forma a maximizar o evento que foi encontrado nos dados, então, quando algo não acontece, esse evento acabar tendo probabilidade zero. No exemplo das moedas isso fica claro: se forem observadas 3 caras, o valor de θ que maximiza a chance dessa sequência ter vindo de uma moeda na qual a chance de ocorrência de cara é θ é 1, e, portanto, a chance estimada de se observar coroa é zero.

Uma ferramenta útil para comparar priori, posteriori e verossimilhança dos dados é o *triplot*, que é um gráfico que contém as três funções plotadas no mesmo plano[8]. A Figura (14) contém os triplots para duas situações: na primeira (Figura 14a) são reproduzidos os valores do exemplo 2.0.3 onde a priori é uma distribuição Beta com parâmetros $\underset{\sim}{\alpha} = \underset{\sim}{\beta} = 1$, os valores para a verossimilhança Binomial são $n = \sum y = 3$ e a posteriori é uma densidade Beta com parâmetros $\bar{\alpha} = 4$ e $\bar{\beta} = 1$. Na Figura (14b) usa-se a mesma configuração da verossimilhança, porém a priori utilizada é mais informativa, com valores $\underset{\sim}{\alpha} = \underset{\sim}{\beta} = 10$ (caracterizando o que seria uma confiança maior de que a moeda é não viciada), o que produz uma posteriori Beta com parâmetros $\bar{\alpha} = 23$ e $\bar{\beta} = 10$.



(a) Priori Beta(1,1) e posteriori Beta(4,1)



(b) Priori Beta(10,10) e posteriori Beta(23,10)

Figura 14: Triplots para duas prioris diferentes no exemplo 2.0.3

É possível observar que na Figura (14a), utilizando uma priori não informativa, o formato da posteriori é o mesmo da verossimilhança. Embora a distribuição Beta(1, 1) tenha valor esperado igual a $1/2$, o fato desta densidade ser uniforme para todos valores entre 0 e 1 implica que a crença à priori na moeda ser não viesada é "fraca". Com isso, os dados da verossimilhança irão ter um peso maior na posteriori. Por outro lado, ao utilizar uma priori mais informativa, cuja média ainda é $1/2$ mas os valores estão mais concentrados em torno desta média, a crença à priori passa a ter mais "peso" na determinação da posteriori, o que é acentuado pelo baixo tamanho amostral. De fato, a média da distribuição Beta(23, 10) é aproximadamente 0.56, um valor bastante próximo da média a priori.

À medida que o tamanho amostral aumenta, a verossimilhança começa a afetar mais a posteriori. Na Figura (15) está o triplot para a situação onde a priori é uma distribuição Beta(10, 10) e a verossimilhança é a de uma distribuição binomial com parâmetros $n = 100$ e $p = 0.75$ (isto é, $\sum y_i = 75$). A densidade a posteriori agora terá parâmetros $\bar{\alpha} = 85$ e $\bar{\beta} = 35$ e seu valor esperado é igual a $^{85}/_{120} \approx 0.71$, bastante próximo do valor esperado dos dados.

[8]Uma vez que a função de verossimilhança não é uma densidade de probabilidade (e portanto não integra 1), é comum multiplicá-la por uma constante de maneira que fique na mesma escala que a priori e a posteriori no triplot. Este procedimento não afeta a análise dos resultados, uma vez que as estimativas obtidas não são afetadas por transformações monotônicas da verossimilhança, pela propriedade de invariância do EMV (Casella and Berger, 1990).

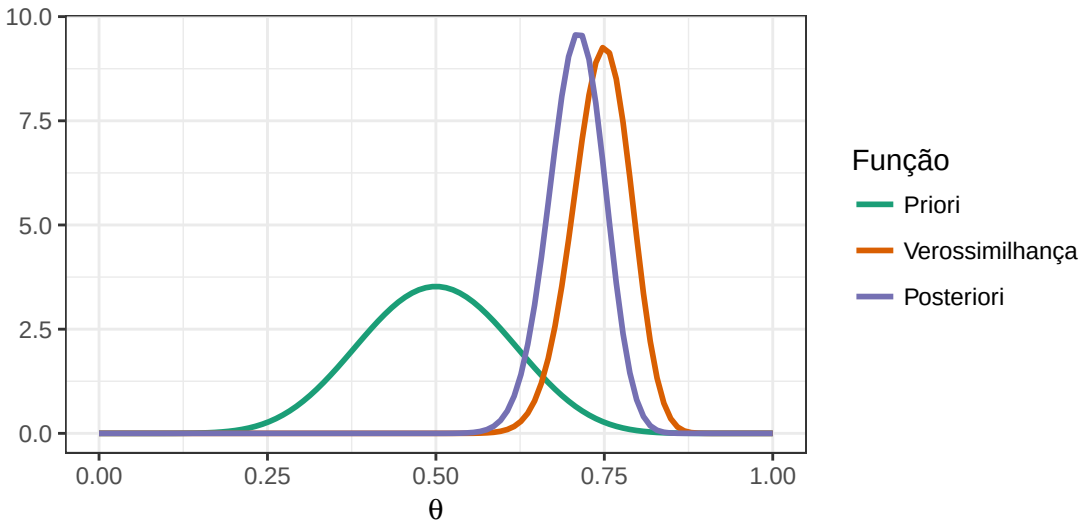


Figura 15: Triplot para priori Beta(10,10) e verossimilhança Binomial com $n = 100$ e $\sum y = 75$.

Os conceitos de núcleo de uma densidade, bem como de sua constante de integração, serão importantes em diversas derivações, pois facilitam o cálculo de várias integrais, como vimos acima. Portanto, vale uma pequena digressão nesse ponto.

Definição 2.0.4. Núcleo de uma densidade

O núcleo de uma densidade $\mathbb{P}(\theta)$ é a função $k(\theta)$ tal que:

$$\mathbb{P}(\theta) = \frac{k(\theta)}{\int k(\theta)\, d\theta} \tag{52}$$

sendo que $\chi = 1/\int k(\theta)\, d\theta$ é chamada de constante de integração

O núcleo de uma densidade contém todos os fatores de $\mathbb{P}(\theta)$ que dependem da variável aleatória θ. Portanto, a razão $k(\theta)/\mathbb{P}(\theta)$ é constante em relação a θ. Note que $k(\theta)$ não é único, pois $k(\theta)$ multiplicado por qualquer constante ou valor independente de θ também satisfaz a definição 2.0.4.

2.1 Um pouco de H.P.E.

Por volta de 1740, o Reverendo Thomas Bayes (1701-1762) desenvolveu seu trabalho intitulado *"Um Ensaio para Solucionar um Problema na Doutrina das Chances"*[9], que buscava uma solução matemática para o problema que na época era conhecido como problema da probabilidade inversa. Bayes nunca publicou seu ensaio e o mesmo foi descoberto apenas após seu falecimento por Richard Price que, após revisar as notas, o submeteu para publicação.

O ensaio de Bayes era baseado em um experimento mental desenvolvido por ele e que buscava desenvolver um método para chegar em resultados a partir de novas observações de um determinado fenômeno. Seu experimento consistia em imaginar uma mesa totalmente plana e sobre ela uma bola; o objetivo era obter um palpite adequado sobre a posição da bola estando de costas para a mesa. Assumindo que uma bola que fosse lançada na mesa teria probabilidade igual de parar em

[9]Do original em inglês *"An Essay toward Solving a Problem in the Doctrine of Chances"*.

qualquer posição, outra pessoa lança bolas e informa se ela parou à esquerda ou à direita da primeira bola. Por exemplo, se todas as bolas estiverem à direita da bola inicial, é seguro afirmar que a primeira bola deve estar na extremidade esquerda da mesa. Com este raciocínio, Bayes desenvolveu um método que permitia, com base em dados atuais, fazer afirmações sobre uma condição inicial desconhecida, em um processo que poderia ser constantemente atualizado à medida que novas informações estivessem disponíveis. No caso do experimento, as novas informações seriam as posições das novas bolas lançadas.

Sem conhecer o trabalho de Bayes, Pierre Simon Laplace (1749-1827) desenvolveu uma primeira versão do que hoje é conhecido como o Teorema de Bayes. Sua descoberta surgiu a partir da necessidade de ferramentas matemáticas para seus trabalhos na área de astronomia e, no ano de 1774, através de Richard Price, Laplace incorporou as ideias de Bayes em seu método. Nos dias atuais, o teorema de Bayes está consolidado como parte da teoria das probabilidades e deu origem aos chamados métodos bayesianos de inferência, porém, ao longo da história, surgiram diversas polêmicas e debates teóricos sobre a validade de tais métodos. Para uma história a respeito do desenvolvimento dos métodos Bayesianos, recomenda-se a leitura de McGrayne (2011).

2.2 Métodos Bayesianos em Econometria

A respeito do uso de métodos bayesianos em econometria, Geweke (2001) e Zellner (1985) citam autores da metade do século XIX como precursores no desenvolvimento do assunto, sendo que um dos primeiros livros textos de econometria bayesiana é a obra de Zellner (1971), intitulada "*An introduction to bayesian inference in econometrics*". Daquela época até os dias atuais, com o advento da computação, métodos bayesianos se tornaram cada vez mais acessíveis e inúmeros trabalhos foram desenvolvidos na área. O artigo de Zellner (1983) traz algumas aplicações de econometria bayesiana desenvolvidas até a década de 80.

Atualmente, existem diversos modelos macroeconométricos onde os métodos bayesianos são amplamente utilizados, como por exemplo os modelos de vetores aleatórios (VAR), desenvolvidos por Christopher Sims no início da década de 80. Doan et al. (1984) estimaram pela primeira vez um VAR bayesiano (BVAR) enquanto que a generalização do modelo considerando parâmetros variando no tempo (TVP-VAR), desenvolvida por Cogley and Sargent (2001), foi estimada utilizando o amostrador de Gibbs, um método bayesiano. Mais tarde, em 2005, os mesmos autores incluíram volatilidade estocástica multivariada no TVP-VAR e novamente estimaram o modelo utilizando o amostrador de Gibbs. De acordo com Geweke et al. (2011), os modelos dinâmicos estocásticos de equilíbrio geral (DSGE), utilizados amplamente em Bancos Centrais, são predominantemente estimados utilizando inferência bayesiana. De fato, desde 2011 o Banco Central do Brasil (BACEN) utiliza o SAMBA (modelo analítico estocástico com abordagem bayesiana[10]) para auxiliar na condução da política macroeconômica no país. Caldeira et al. (2015) fazem uma comparação de modelos, incluindo um TVP-VAR bayesiano com uso de priori de Minnesota para previsão de dados macroeconômicos do Brasil.

Na área de finanças, modelos bayesianos também tem sido utilizados, em particular pela capacidade de lidar com modelos altamente parametrizados além da possibilidade de incorporar conhecimentos prévios de mercado. Por exemplo, Philipov and Glickman (2006) propõem um modelo com volatilidade estocástica Wishart para otimização de portfólios. Kastner et al. (2017) fazem a estimação bayesiana eficiente de matrizes de covariância dinâmicas em modelos de series temporais multivariadas, aplicando o modelo para dados de taxas de câmbio.

Recomenda-se a leitura de Geweke et al. (2011) para métodos bayesianos nas áreas de microeconomia, macroeconomia, finanças e marketing, além das demais obras citadas neste texto.

[10]Tradução livre de *Stochastic Analytical Model with a Bayesian Approach*.

3 Parte 3 - Inferência Bayesiana no modelo normal clássico de regressão linear

Neste capítulo iremos estudar o modelo de regressão linear múltipla com priori conjugada natural, como descrito no capítulo 3 de Koop (2003). Note que o capítulo 2 foi "pulado" (é o modelo de regressão com apenas uma variável independente), porém é aconselhável que as principais contas dele sejam resolvidas. O que muda, do capítulo 2 para o capítulo 3, é o uso intensivo de álgebra matricial (AM), que simplifica muito as contas. O apêndice *A* de Koop (2003) tem uma revisão de AM, mas isso pode ser encontrado em outros livros, incluindo de econometria clássica, como Greene (2003). Além destas referências, pode ser de bastante ajuda o texto Petersen et al. (2008). Note que ele não é um livro formal (está mais para apostila) e contém apenas os resultados, sem as demonstrações. É bom sempre dar uma verificada nas fontes originais caso você tenha alguma dúvida se uma identidade é ou não válida[11].

O modelo de regressão é uma das principais ferramentas na econometria. Além de ser utilizado individualmente, acaba aparecendo como componente em uma série de outros modelos mais complexos. No modelo linear de regressão, é possível modelar o relacionamento de uma variável dependente y com k variáveis explicativas, x_1, x_2, ..., x_k através da seguinte forma funcional:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \ldots + \beta_k \boldsymbol{x_k} + \varepsilon \tag{53}$$

Onde:

1. $\boldsymbol{y}$ é um vetor com N entradas da variável dependente, isto é, para cada unidade amostral é observado um valor de y, em que N é o tamanho total da amostra;
2. $\boldsymbol{x_i}$, $i \in \{1, \ldots, k\}$ são vetores $N \times 1$ de variáveis independentes, também observáveis;
3. β_i, $i \in \{1, \ldots, k\}$ são os parâmetros do modelo, que deseja-se estimar;
4. ε é um vetor $N \times 1$ de erros não observáveis.

A solução clássica de modelos como o descrito na Equação (53) envolve encontrar estimativas $\hat{\beta}_1, \ldots, \hat{\beta}_k$ para os parâmetros populacionais $\beta_1, \ldots, \beta_k$, de forma que novas observações das variáveis independentes (que usualmente serão de mais fácil obtenção do que a variável dependente) possam ser utilizadas para uma aproximação de quais seriam os valores de y. Além disso, os coeficientes $\hat{\beta}_1, \ldots, \hat{\beta}_k$ estimados fornecem uma interpretação direta sobre a magnitude e a direção dos efeitos que os regressores x_i têm sobre y.

O modelo (53) pode ser escrito matricialmente, definindo os vetores $N \times 1$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

juntamente com o vetor $k \times 1$ β e a matriz $N \times k$ X, dados por

[11] Por exemplo, propriedades com determinantes não podem ser utilizadas para matrizes quadradas e isso não está explícito no Petersen et al. (2008) - isso pode induzir uma pessoa mais desatenta a cometer erros na demonstração por usar uma propriedade em um caso onde ela não é válida.

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad X = \begin{bmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \dots & x_{Nk} \end{bmatrix}.$$

Então, a representação matricial de (53) será dada por:

$$Y = X\beta + \varepsilon \tag{54}$$

em que, de forma similar ao que foi definido para (53), temos:

- Y é um vetor de tamanho $N \times 1$ da variável explicada (endógena ou dependente);
- X é uma matriz $N \times k$ cuja primeira coluna é toda formada de números 1 e as demais colunas representam entradas de variáveis explicativas (exógenas ou independentes);
- β é um vetor $k \times 1$ de parâmetros;
- ε é um vetor $N \times 1$ de erros.

Hipóteses a respeito de ε e X são o que definem a função de verossimilhança, sendo que para os modelos (53) e (54) iremos considerar inicialmente as seguintes hipóteses:

1. Os erros tem distribuição normal multivariada[12] com média zero e matriz de variância e covariância homocedástica, isso é, $\varepsilon \sim N(0_N, \mathbb{I}h^{-1})$, onde $\mathbb{I}$ é a matriz identidade $n \times n$, 0_N denota que a média é um vetor de zeros com N entradas e $h^{-1} = \sigma^2$. O parâmetro h é chamado de *precisão* e representa o inverso da variância (i.e., $h = 1/\sigma^2$);
2. Os elementos de X são fixos ou, caso sejam variáveis aleatórias, são independentes de ε com f.d.p. $p(X|\lambda)$, onde λ é um vetor de parâmetros que não depende de β ou h.

Nas notas, tentaremos[13] seguir a seguinte notação: letras maiúsculas como X denotam matrizes[14] e letras minúsculas como y ou x_i representam escalares ou vetores. Neste último caso, sempre que for importante distinguir vetor de escalar, isto será feito no texto.

3.1 A função de verossimilhança

A função de verossimilhança será determinada pela função densidade de probabilidade conjunta de Y e X condicional aos parâmetros desconhecidos, ou seja $p(Y, X|\beta, h, \lambda)$. Como os X são independentes dos erros ε, podemos reescrever da seguinte maneira:

$$p(Y, X|\beta, h, \lambda) = \underbrace{p(Y|X, \beta, h)}_{\substack{\text{Só essa parte é} \\ \text{relevante para} \\ \text{inferência a respeito} \\ \text{de } \beta \text{ e } h}} \underbrace{p(X|\lambda)}_{\substack{\text{independente} \\ \text{de } \beta \text{ e } h}}$$

[12] Consulte o anexo das notas de aula ou o anexo B de Koop (2003) para ver a definição e propriedades da Normal Multivariada.
[13] Às vezes pode ter algum problema de digitação, mas a princípio será tentado seguir uma convenção.
[14] Nas notas de probabilidade as letras maiúsculas eram utilizadas para variáveis aleatórias (a função) e as letras minúsculas para uma *realização* de uma v.a.. Agora, em econometria bayesiana, uma vez que essa distinção não é mais o foco, essa convenção probabilística será deixada de lado.

Observe que $p(X|\lambda)$ não depende nem de β nem de h. Então, para fins de cálculo da posteriori, ela será uma constante que poderá ser omitida. Iremos omitir também o X de $p(Y|X,\beta,h)$, escrevendo simplesmente $p(Y|\beta,h)$ (mas lembre-se que X estará implicitamente presente nessas condicionais).

Usando as hipóteses a respeito do modelo, pode-se demonstrar[15] que:

1. $p(Y|\beta,h)$ é normal;
2. $\mathbb{E}[Y|\beta,h] = X\beta$;
3. $Var[Y|\beta,h] = h^{-1}I$.

Combinando os resultados acima com a definição da distribuição normal multivariada, chega-se na função de verossimilhança:

$$p(y \mid \beta,h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left[-\frac{h}{2}(Y-X\beta)'(Y-X\beta)\right] \tag{55}$$

É interessante reescrever a função de verossimilhança de maneira que em sua equação apareçam as quantidades de mínimos quadrados ordinários (MQO), o que, ao final do processo, possibilitará uma interpretação analítica mais clara da posteriori. As seguintes quantidades de MQO serão utilizadas para reescrever (55):

$$\nu = N - k \quad \text{(graus de liberdade)}, \tag{56}$$

$$\hat{\beta} = (X'X)^{-1}(X'Y) \quad \text{(estimador de MQO)}, \tag{57}$$

$$s^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{\nu} = \frac{SQR}{\nu} \quad \text{(estimador não viesado de } \sigma^2\text{)}. \tag{58}$$

Antes de continuarmos, vamos relembrar a definição de núcleo de distribuições.

Definição 3.1.1. O núcleo de uma densidade $p(y)$ é uma função $k(y)$ tal que:

$$p(y) = \frac{k(y)}{\int k(y)dy} \tag{59}$$

na qual $1/\int k(y)dy = \chi$ é chamada de *constante de integração*. Se $p(y)$ é uma densidade, então $\int p(y)dy = 1$. Integrando 59, temos:

$$\int p(y)dy = \int \frac{k(y)}{\int k(y)dy}dy = \int \chi k(y)dy = \chi \int k(y)dy = \chi \cdot \frac{1}{\chi} = 1 \tag{60}$$

O núcleo contém todos os fatores de y que estão na densidade, isto é, todos os fatores de $p(y)$ que dependem de y estarão no núcleo $k(y)$.

Exemplo 3.1.2. Núcleo da Normal
Considere $Y \sim N(X,V)$. Então, sua densidade é dada por:

[15] Fica de sugestão pensar a respeito.

$$p(y) = \frac{1}{(2\pi)^{\frac{K}{2}}}|V|^{-1}\exp\left[-\frac{1}{2}(Y-X)'V^{-1}(Y-X)\right]$$

E um núcleo $k(y)$ pode ser:

$$k(y) = \exp\left[-\frac{1}{2}(Y-X)'V^{-1}(Y-X)\right] \tag{61}$$

$$= \exp\left[-\frac{1}{2}\left(Y'Y - 2Y'V^{-1}X + \underbrace{X'V^{-1}X}_{\substack{\text{não depende}\\ \text{de } Y \text{ - poderia}\\ \text{sair do } k(y)}}\right)\right] \tag{62}$$

O último termo no parênteses em (61) nos mostra que o núcleo de uma densidade não é único, pois $\kappa k(y)$ também é um núcleo para $p(y)$ para qualquer κ independente de y. A equação (60) nos ajuda a encontrar as constantes de integração para o caso de densidades conhecidas, pois, usando 59, a integral será calculada como $\int k(y)dy = \frac{k(y)}{p(y)}$.

Utilizando (56), (57) e (58), a verossimilhança descrita em (55) pode ser escrita como produto do núcleo de uma densidade Normal para $\beta|h$ vezes o núcleo de uma densidade Gama para h:

$$p(Y|\beta,h) = \underbrace{\left[\frac{1}{(2\pi)^{\frac{N}{2}}}h^{\frac{k}{2}}\exp\left[-\frac{h}{2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right]}_{\text{Núcleo de uma Normal para } \beta|h} \underbrace{\left[h^{\frac{\nu}{2}}\exp\left[-\frac{h}{2}\frac{\nu}{s^{-2}}\right]\right]}_{\text{Núcleo de uma Gama para } h} \tag{63}$$

Demonstração. Os próximos passos mostram que de fato a função de verossimilhança dada na equação (55) é o produto de uma densidade Normal para $\beta|h$ vezes o núcleo de uma densidade gama para h. **Atenção:** esta não é a mesma demonstração vista em sala de aula.

O primeiro passo é reescrever o termo $(Y-X\beta)'(Y-X\beta)$ de (55) somando e subtraindo $X\hat{\beta}$:

$$\begin{aligned}(Y-X\beta)'(Y-X\beta) = &\\ &= (Y-X\beta+X\hat{\beta}-X\hat{\beta})'(Y-X\beta+X\hat{\beta}-X\hat{\beta})\\ &= ((Y-X\hat{\beta})-(X\beta-X\hat{\beta}))'((Y-X\hat{\beta})-(X\beta-X\hat{\beta}))\\ &= ((Y-X\hat{\beta})'-(\beta-\hat{\beta})'X')((Y-X\hat{\beta})-X(\beta-\hat{\beta}))\end{aligned} \tag{64}$$

Então, faz-se a distributiva para obter:

$$(Y-X\hat{\beta})'(Y-X\hat{\beta}) - (Y-X\hat{\beta})'X(\beta-\hat{\beta}) - (\beta-\hat{\beta})'X'(Y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \tag{65}$$

Os termos centrais de (65) são iguais a zero pois a expressão $(Y-X\hat{\beta})'X$ é igual a zero (consequentemente, seu transposto também é). Isto decorre da hipótese de ortogonalidade entre o vetor de resíduos ε e a matriz X e pode ser verificada algebricamente da seguinte forma:

$$
\begin{aligned}
(Y - X\hat{\beta})'X &= (Y - X(X'X)^{-1}X'Y)'X \\
&= (Y' - (X(X'X)^{-1}X'Y)')X \\
&= (Y' - (X'Y)'(X(X'X)^{-1})')X \\
&= (Y' - Y'X((X'X)^{-1'}X'))X \\
&= (Y' - Y'X(X'X)^{-1}X')X \\
&= Y'X - Y'X\underbrace{(X'X)^{-1}X'X}_{\text{Identidade}} \\
&= Y'X - Y'X \\
&= 0,
\end{aligned}
$$

em que foram utilizados os seguintes resultados de álgebra matricial: se A e B são matrizes, então $(A + B)' = A' + B'$ e se o seu produto matricial está bem definido, vale também que $(AB)' = B'A'$.

Então:

$$
(Y - X\beta)'(Y - X\beta) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \tag{66}
$$

O próximo passo é mostrar que o termo $(Y - X\hat{\beta})'(Y - X\hat{\beta})$, que é a soma de quadrado dos resíduos de MQO (denotada por SQR), pode ser escrito como $SQR = Y'Y - \hat{\beta}X'X\hat{\beta}$. Primeiro, aplica-se a distributiva para obter $(Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - Y'X\hat{\beta} - (X\hat{\beta})'Y + (X\hat{\beta})'(X\hat{\beta})$. Uma vez que $Y'X\hat{\beta}$ é um produto interno e pode ser denotado por $\langle Y, X\hat{\beta}\rangle$, valem as propriedades de comutatividade, distributividade, bilinearidade, multiplicação por escalar e associatividade no produto por escalar.

Utilizando a comutatividade, tem-se:

$$
\begin{aligned}
-Y'X\hat{\beta} - (X\hat{\beta})'Y &= -\langle Y, X\hat{\beta}\rangle - \langle X\hat{\beta}, Y\rangle \\
&= -\langle X\hat{\beta}, Y\rangle - \langle X\hat{\beta}, Y\rangle \\
&= -Y'X\hat{\beta} - Y'X\hat{\beta} \\
&= -2Y'X\hat{\beta}
\end{aligned} \tag{67}
$$

De forma que:

$$
(Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \tag{68}
$$

O último termo pode ser reescrito da seguinte forma:

$$
\begin{aligned}
\hat{\beta}X'X\hat{\beta} = & \\
& = \hat{\beta}X'X\overbrace{(X'X)^{-1}X'Y}^{\hat{\beta}} \\
& = \hat{\beta}'X'Y \\
& = (X\hat{\beta})'Y \\
& = Y'X\hat{\beta}
\end{aligned} \tag{69}
$$

Fazendo a substituição em (68), tem-se:

$$
\begin{aligned}
(Y - X\hat{\beta})'(Y - X\hat{\beta}) = & \\
& = Y'Y - 2Y'X\hat{\beta} + Y'X\hat{\beta} \\
& = Y'Y - Y'X\hat{\beta} \\
& = \underbrace{Y'Y - \hat{\beta}'X'X\hat{\beta}}_{SQR}
\end{aligned} \tag{70}
$$

Logo, pode-se escrever a equação (66) como:

$$
(Y - X\beta)'(Y - X\beta) = Y'Y - \hat{\beta}'X'X\hat{\beta} + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \tag{71}
$$

E a verossimilhança descrita em (55) será:

$$
p(Y|\beta, h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} exp\left\{-\frac{h}{2}\left[\nu s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\} \tag{72}
$$

Utilizando o fato de que $\nu = N - k$, pode-se escrever $h^{\frac{N}{2}} = h^{\frac{\nu+k}{2}} = h^{\frac{\nu}{2}} \cdot h^{\frac{k}{2}}$. Além disso, pode-se escrever s^2 como $\frac{1}{s^{-2}}$ para reorganizar (72) e chegar na equação (63). □

3.2 Função de densidade a priori para β e h

A priori deve ser estabelecida de forma a representar a crença que se tem a respeito dos parâmetros *antes* (ou *a priori*) de observar os dados. Em geral, prioris podem assumir qualquer forma funcional, podendo ser, inclusive, constantes. A priori escolhida, usualmente, além de refletir a crença sobre o que se deseja inferir, terá também como característica desejável a facilidade de incorporação ao modelo, simplificando os cálculos e, algumas vezes, levando a soluções em fórmula analítica fechada[16].

Quando a priori escolhida não é uma constante e sim uma densidade de probabilidade, seus hiperparâmetros serão auxiliares para a "calibragem" da priori. No exemplo (2.0.3) foi utilizada como priori uma distribuição Beta(1, 1). Apesar

[16]Em aplicações onde não é possível chegar em uma posteriori com fórmula analítica fechada, métodos como o de Monte Carlo são utilizados para inferência a posteriori.

desta densidade ter média 0.5, que seria indicativa de uma moeda honesta, a "certeza" atribuída a esta média é baixa: como exemplificado na Figura (13), este é um caso particular da distribuição uniforme e por isso qualquer valor entre 0 e 1 é igualmente provável. Em geral, variâncias maiores na distribuição a priori estão associadas com um grau de incerteza maior e vice-versa.

Uma priori é dita *conjugada natural* quando, além de produzir uma posteriori de mesma distribuição (como no caso da priori conjugada), a priori também tem a mesma distribuição que a verossimilhança. Koop (2003) destaca que o uso de uma conjugada natural tem a vantagem adicional de que a informação à priori pode ser interpretada de maneira similar à da verossimilhança, isto é, a priori pode ser vista como proveniente de um conjunto de dados fictícios gerados por um mesmo processo gerador dos dados.

Definição 3.2.1. Priori Conjugada
A densidade a priori é dita *conjugada* quando ela combinada com a função de verossimilhança gera uma f.d.p. a posteriori que pertence à mesma classe da distribuição da priori.

Quando usamos uma priori conjugada, ao coletar novos dados basta fazer a atualização dos parâmetros. No exemplo das moedas, no caso onde $\bar{\alpha} = 4$ e $\bar{\beta} = 1$, caso observássemos mais 5 lançamentos da moeda e neles observássemos duas caras e três coroas, bastaria atualizar os parâmetros para $\bar{\alpha} = 4 + 2 = 6$ e $\bar{\beta} = 1 + 5 - 2 = 4$, conforme a equação (50).

Definição 3.2.2. Priori conjugada natural
A densidade a priori conjugada natural tem a propriedade adicional de possuir a mesma forma funcional da verossimilhança.

O modelo de regressão descrito em (54) tem dois parâmetros desconhecidos, β e h, e portanto a priori pode ser denotada como uma densidade conjunta $p(\beta, h)$. Utilizando a definição de densidade condicional, a priori pode ser escrita como $p(\beta, h) = p(\beta|h)p(h)$. Esta notação permite pensar em uma priori para β condicional a h e outra para a precisão h. A Equação (63) sugere que a f.d.p a priori conjugada natural é o produto de uma normal condicional para $\beta|h \sim N(\underline{\beta}, h^{-1}\underline{V})$ e uma gama para $h \sim G(\underline{\nu}, \underline{s}^{-2})$, que resulta em uma distribuição Normal-Gama (consulte o apêndice de Koop (2003) para a definição e propriedades desta distribuição). Isto é, $p(\beta, h) = \underbrace{p(\beta|h)}_{normal}\ \underbrace{p(h)}_{gama}$, de forma que:

$$p(\beta, h) = \underbrace{\frac{h^{\frac{k}{2}}}{(2\pi)^{\frac{k}{2}}}|\underline{V}|^{-\frac{1}{2}} \exp\left[-\frac{h}{2}(\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta})\right]}_{\text{Normal para } \beta|h} \underbrace{\left[\left(\frac{2\underline{s}^{-2}}{\underline{\nu}}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)\right]^{-1} h^{\frac{\underline{\nu}-2}{2}} \exp\left[-\frac{h}{2}\underline{\nu}\,\underline{s}^{-2}\right]}_{\text{Gama para } h} \tag{73}$$

Para $0 < h < \infty$, onde $\Gamma(\cdot)$ é a função Gama[17]. As quantidades $\underline{\beta}$, $\underline{V}$, $\underline{s}^{-2}$ e $\underline{\nu}$ são os hiperparâmetros da densidade a priori conjunta e devem caracterizar o conhecimento do pesquisador a respeito de β e h antes de observar os dados. Observe que, ao contrário da verossimilhança que é apenas uma *função*, aqui estamos trabalhando com uma densidade e por isso todas as constantes aparecem corretamente.

[17]

Definição 3.2.3. $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$ é chamada de função Gama e é tal que:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

A função Gama ($\Gamma(\cdot)$) satisfaz as seguintes propriedades:

1. Para $a \geq 1$, $\Gamma(a) = (a-1)\Gamma(a-1)$
2. Se $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$

O termo $\underline{\nu}\underline{s}^2$ de (75) é uma soma de quadrados dos resíduos a priori. Reescrevendo (75) (não é difícil, mas é recomendado que todo mundo tente fazer a conta), podemos jogar todas constantes em um único termo χ e obter:

$$p(\beta, h) = \chi h^{\frac{k+\nu-2}{2}} \exp\left[-\frac{h}{2}\left(\underline{\nu}\underline{s}^2 + (\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta})\right)\right] \tag{74}$$

Em que $\chi = (2\pi)^{-\frac{k}{2}}\left[\left(2\frac{\underline{s}^{-2}}{\underline{\nu}}\right)^{\frac{\underline{\nu}}{2}}\Gamma\left(\frac{\underline{\nu}}{2}\right)\right]^{-1}|\underline{V}|^{-\frac{1}{2}}$. É utilizada a notação $p(\beta, h) \sim \mathcal{NG}(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu})$.

Uma outra notação usual é simplesmente omitir a constante e utilizar o símbolo de proporcionalidade:

$$p(\beta, h) \propto h^{\frac{k+\nu-2}{2}} \exp\left[-\frac{h}{2}(\underline{\nu}\underline{s}^2 + (\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta}))\right] \tag{75}$$

3.2.1 A densidade a posteriori

A densidade posterior irá resumir todas as informações disponíveis a respeito dos parâmetros. Isso inclui tanto informações amostrais quanto informações não amostrais (priori). A f.d.p. posterior é proporcional à função de verossimilhança (63) multiplicada pela f.d.p. a priori (75), isto é

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} = \frac{p(y|\theta)p(\theta)}{p(y)} \propto \overbrace{p(y|\theta)}^{\text{verossimilhança}}\overbrace{p(\theta)}^{\text{priori}}.$$

Enquanto a notação da priori era apenas $p(\beta, h)$ para mostrar a não influência dos dados, para a posteriori usa-se a notação $p(\beta, h|y)$ para deixar claro que é uma densidade após a inclusão da informação da verossimilhança. Combinando a verossimilhança e a priori definida na seção anterior, tem-se que a posteriori do modelo será dada por:

$$p(\beta, h|y) \propto h^{\frac{\bar{\nu}+\underline{k}-2}{2}} \exp\left[-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})\right)\right] \tag{76}$$

O lado direito da Equação 76 é o núcleo de uma densidade Normal-Gama com parâmetros:

- $\bar{V} = (\underline{V}^{-1} + X'X)^{-1}$, o que significa que a variância da posteriori depende da precisão da priori e da precisão da verossimilhança.
- $\bar{\beta} = \bar{V}(\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta})$. Isso implica que o parâmetro $\bar{\beta}$, que representa a média da posteriori conjunta, é uma média ponderada do parâmetro da priori $\underline{\beta}$ com o estimador de MQO $\hat{\beta}$. Se a precisão $\underline{V}^{-1}$ for baixa (ou seja, a variância à priori for alta), a contribuição de $\underline{\beta}$ para a posteriori será pequeno. Por outro lado, perceba que $X'X$ é uma soma de N termos e, portanto, é proporcional a N, ou seja, quanto maior N, maior o peso de $\hat{\beta}$ em $\bar{\beta}$. Se N for muito grande, a média da distribuição a posteriori ficará próxima do estimador de MQO $\hat{\beta}$ e, caso N seja muito pequeno, ficará próxima do hiperparâmetro da priori, $\underline{\beta}$.
- $\bar{\nu} = \underline{\nu} + N$, ou seja, os graus de liberdade da priori são adicionados ao tamanho amostral para fornecer os graus de liberdade da posteriori. Podemos pensar em $\underline{\nu}$ como "o tamanho da amostra a priori". Em uma aplicação prática onde a priori vem de um modelo, como por exemplo um DSGE usado como priori para um VAR, podemos pensar que o número de observações geradas pelo DSGE será o valor usado em $\underline{\nu}$.
- $\bar{s}$ é definido implicitamente através de $\bar{\nu}\bar{s}^2 = \underline{\nu}\underline{s}^2 + \nu s^2 + (\hat{\beta} - \underline{\beta})'[\underline{V} + (X'X)^{-1}]^{-1}(\hat{\beta} - \underline{\beta})$, em que $\underline{\nu}\underline{s}^2$ é a SQR da priori, νs^2 é a SQR da verossimilhança e o último termo penaliza quanto maior a diferença entre o valor de $\underline{\beta}$ e o estimador

de MQO, $\hat{\beta}$.

Demonstração. **Observação:** *Esta demonstração pode conter diferenças em relação ao que foi indicado na aula.*

A densidade a posteriori, por definição, é proporcional ao produto da verossimilhança e da priori. Utilizando (39) junto com as expressões da verossimilhança (63) e da priori (75), tem-se:

$$\begin{aligned} p(\beta, h|y) &\propto p(Y|\beta, h)p(\beta, h) \\ &= \left[\frac{1}{(2\pi)^{\frac{N}{2}}} h^{\frac{k}{2}} \exp\left[-\frac{h}{2}\underbrace{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}_{(*)}\right]\right]\left[h^{\frac{\nu}{2}}\exp\left[-\frac{h}{2}\frac{\nu}{s^{-2}}\right]\right] \\ &\quad \cdot \chi h^{\frac{k+\nu-2}{2}} \exp\left[-\frac{h}{2}(\underline{\nu}\underline{s}^2 + \underbrace{(\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta})}_{(**)})\right] \end{aligned} \tag{77}$$

O termo χ é a constante da distribuição Normal-Gama, dada por:

$$\chi = (2\pi)^{-\frac{k}{2}}\left[\left(2\frac{\underline{s}^{-2}}{\underline{\nu}}\right)^{\frac{\underline{\nu}}{2}}\Gamma\left(\frac{\underline{\nu}}{2}\right)\right]^{-1}|\underline{V}|^{-\frac{1}{2}}$$

em que $\Gamma(\cdot)$ é a função gama.

Uma vez que os termos $(*)$ e $(**)$ de (77) são expoentes de mesma base, podem ser somados:

$$\begin{aligned} &(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + (\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta}) = \\ &= \beta'X'X\beta - \beta'X'X\hat{\beta} - \hat{\beta}'X'X\beta + \hat{\beta}'X'X\hat{\beta} + \beta'\underline{V}^{-1}\beta - \beta'\underline{V}^{-1}\underline{\beta} - \underline{\beta}'\underline{V}^{-1}\beta + \underline{\beta}'\underline{V}^{-1}\underline{\beta} \\ &= \underbrace{\beta'\left[\underline{V}^{-1}+X'X\right]\beta - \beta'\left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right] - \left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right]'\beta}_{(***)} + \underbrace{\hat{\beta}'X'X\hat{\beta}+\underline{\beta}'\underline{V}^{-1}\underline{\beta}}_{(\bigstar)}. \end{aligned} \tag{78}$$

O termo $(***)$ é uma forma quadrática do tipo $\beta'A\beta - \beta'b - b'\beta$, onde $A = [\underline{V}^{-1} + X'X]$ e $b = \left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right]$. Somando e diminuindo $b'A^{-1}b$ e ignorando temporariamente o termo negativo, $-b'A^{-1}b$, pode-se escrever $(***)$ como:

$$\begin{aligned} &\beta'\left[\underline{V}^{-1}+X'X\right]\beta - \beta'\left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right] - \left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right]'\beta \\ &\quad + \left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right]'\left[\underline{V}^{-1}+X'X\right]^{-1}\left[\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}\right]. \end{aligned} \tag{79}$$

Observe que (79) é um termo do tipo $\beta'A\beta - \beta'b - b'\beta + b'A^{-1}b$. Como A é a soma de uma matriz de variância de posto completo e $X'X$ é simétrica (também de posto completo), vale que $A^{-1}A = AA^{-1} = I$ e também $A' = A$. Portanto:

$$\begin{aligned} &\beta'A\beta - \beta'b - b'\beta + b'A^{-1}b = \\ &\quad \beta'A\beta - \beta'AA^{-1}b - b'A^{-1}A\beta + b'A^{-1}AA^{-1}b \end{aligned} \tag{80}$$

Agora, considerando $\bar{V} = A^{-1}$ e $\bar{\beta} = A^{-1}b$, a expressão (80) será $\beta'\bar{V}^{-1}\beta - \beta'\bar{V}\bar{\beta} - \bar{\beta}'\bar{V}\beta + \bar{\beta}'\bar{V}\bar{\beta}$, que pode ser escrito como $(\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})$, onde $\bar{V} = A^{-1} = [\underline{V}^{-1} + X'X]^{-1}$ e $\bar{\beta} = A^{-1}b = \bar{V}\left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right]$.

Resta trabalhar com (★) de (78) junto com o termo $-b'A^{-1}b$ que havia sido desconsiderado até então. O objetivo dos próximos quatro passos é escrever $\hat{\beta}'X'X\hat{\beta} + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - b'A^{-1}b$ como $(\hat{\beta} - \beta)'[\underline{V} + (X'X)^{-1}]^{-1}(\hat{\beta} - \beta)'$. Utilizando a definição de b, tem-se:

$$\begin{aligned}&\hat{\beta}'X'X\hat{\beta} + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - b'A^{-1}b = \\ &\quad \hat{\beta}'X'X\hat{\beta} + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \underbrace{\left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right]'\left[\underline{V}^{-1} + X'X\right]^{-1}\left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right]}_{(a)}\end{aligned} \tag{81}$$

Passo 1: Reescrever o termo (a) da equação (81)
Usando a definição de $\bar{V}$ em (a) e fazendo a distributiva, obtem-se:

$$\begin{aligned}&\left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right]'\bar{V}\left[\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right] = \\ =&\beta'\left(\underline{V}^{-1}\right)'\bar{V}\underline{V}^{-1}\underline{\beta} + \hat{\beta}'X'X\bar{V}\underline{V}^{-1}\underline{\beta} + (\underline{V}^{-1}\underline{\beta})'\bar{V}X'X\hat{\beta} + \hat{\beta}'X'X\bar{V}X'X\hat{\beta}\end{aligned} \tag{82}$$

Então, (81), após rearranjar os termos e substituir (a) por (82) será:

$$\hat{\beta}'\underbrace{\left[X'X - X'X\bar{V}X'X\right]}_{(b)}\hat{\beta} + \underline{\beta}'\underbrace{\left[\underline{V}^{-1} - \underline{V}^{-1}\bar{V}\underline{V}^{-1}\right]}_{(c)}\underline{\beta} - \underline{\beta}\underline{V}^{-1}\bar{V}X'X\hat{\beta} - \hat{\beta}'X'X\bar{V}\underline{V}^{-1}\underline{\beta} \tag{83}$$

Onde foi usado o fato de que $\underline{V}$ é uma matriz simétrica (pois é uma matriz de covariância), de maneira que $\underline{V}'^{-1} = \underline{V}^{-1}$.

Passo 2: Mostrar que o termo (b) de (83) é igual a $X'X\bar{V}\underline{V}^{-1}$.

$$\begin{aligned}\mathbb{I} &= [\underline{V}^{-1} + X'X]^{-1}[\underline{V}^{-1} + X'X] \\ \mathbb{I} &= [\underline{V}^{-1} + X'X]^{-1}\underline{V}^{-1} + [\underline{V}^{-1} + X'X]^{-1}X'X \\ \mathbb{I} - [\underline{V}^{-1} + X'X]^{-1}X'X &= [\underline{V}^{-1} + X'X]^{-1}\underline{V}^{-1} \\ \mathbb{I} - \bar{V}X'X &= \bar{V}\underline{V}^{-1} \\ X'X(\mathbb{I} - \bar{V}X'X) &= X'X(\bar{V}\underline{V}^{-1})\end{aligned} \tag{84}$$

Com $\mathbb{I}$ correspondendo à matriz identidade.

Passo 3: Mostrar que $\underline{V}^{-1} - \underline{V}^{-1}\bar{V}\underline{V}^{-1} = \underline{V}^{-1}\bar{V}X'X$

$$
\begin{aligned}
\mathbb{I} = \overline{V}\overline{V}^{-1} &= \\
&= \overline{V}\left[\underline{V}^{-1} + X'X\right] \\
&= \overline{V}\underline{V}^{-1} + \overline{V}X'X
\end{aligned}
$$

Juntando as duas extremidades, tem-se que:

$$
\begin{aligned}
\mathbb{I} &= \overline{V}\underline{V}^{-1} + \overline{V}X'X \\
\mathbb{I} - \overline{V}\underline{V}^{-1} &= \overline{V}X'X \\
\underline{V}^{-1}\left[\mathbb{I} - \overline{V}\underline{V}^{-1}\right] &= \underline{V}^{-1}\overline{V}X'X \\
\underline{V}^{-1} - \underline{V}^{-1}\overline{V}\underline{V}^{-1} &= \underline{V}^{-1}\overline{V}X'X
\end{aligned}
$$

Passo 4: Mostrar que o termo $X'X\overline{V}\underline{V}^{-1}$ (do passo 3) é igual a $(\underline{V} + (X'X)^{-1})^{-1}$.

$$
\begin{aligned}
X'X\overline{V}\underline{V}^{-1} &= X'X\left[\underline{V}^{-1} + X'X\right]^{-1}\underline{V}^{-1} = \\
&= X'X\left[\underline{V}^{-1} + ((X'X)^{-1})^{-1}\right]^{-1}\underline{V}^{-1} \\
&= X'X(X'X)^{-1}(\underline{V} + (X'X)^{-1})^{-1}\underline{V}\underline{V}^{-1} \\
&= (\underline{V} + (X'X)^{-1})^{-1}
\end{aligned}
$$

Na penúltima etapa do passo 4 foi utilizado no termo entre colchetes o fato que se as inversas estão bem definidas, então $(A^{-1} + B^{-1})^{-1} = A(A + B)^{-1}B = B(A + B)^{-1}A$ (onde $A = \underline{V}$ e $B = (X'X)^{-1}$) (Searle, 1982). Isso pode ser verificado multiplicando um termo pelo inverso do outro, para obter a matriz identidade:

$$
\begin{aligned}
(A^{-1} + B^{-1})A(A + B)^{-1}B = & \\
&= A^{-1}A(A + B)^{-1}B + B^{-1}A(A + B)^{-1}B \\
&= (A + B)^{-1}B + B^{-1}A(A + B)^{-1}B \\
&= (\mathbb{I} + B^{-1}A)(A + B)^{-1}B \\
&= (B^{-1}B + B^{-1}A)(A + B)^{-1}B \\
&= B^{-1}(B + A)(A + B)^{-1}B \\
&= B^{-1}\mathbb{I}B \\
&= B^{-1}B = \mathbb{I}
\end{aligned}
$$

Usando o mesmo resultado, conclui-se que $\underline{V}^{-1}\left[\underline{V}^{-1} + X'X\right]^{-1}X'X = \left[\underline{V} + (X'X)^{-1}\right]^{-1}$.

Juntando os passos 1, 2, 3 e 4, pode-se reescrever a equação (83) como:

$$
\begin{aligned}
\hat{\beta}' \overbrace{X'X - X'X\bar{V}X'X}^{X'X\bar{V}\underline{V}^{-1}} \hat{\beta} + \underline{\beta}' \overbrace{\underline{V}^{-1} - \underline{V}^{-1}\bar{V}\underline{V}^{-1}}^{\underline{V}^{-1}\bar{V}X'X} \underline{\beta} - \underline{\beta}\underline{V}^{-1}\bar{V}X'X\hat{\beta} - \hat{\beta}'X'X\bar{V}\underline{V}^{-1}\underline{\beta} = \\
= \hat{\beta}' \left[\underline{V} + (X'X)^{-1}\right]^{-1} \hat{\beta} + \underline{\beta}' \left[\underline{V} + (X'X)^{-1}\right]^{-1} \underline{\beta} + \underline{\beta}' \left[\underline{V} + (X'X)^{-1}\right]^{-1} \hat{\beta} + \hat{\beta}' \left[\underline{V} + (X'X)^{-1}\right]^{-1} \underline{\beta} \\
= \left(\hat{\beta} - \underline{\beta}\right)' \left[\underline{V} + (X'X)^{-1}\right]^{-1} (\hat{\beta} - \underline{\beta})
\end{aligned}
$$

Com os resultados acima, o núcleo da densidade a posteriori conjunta para β e h dado y pode ser escrita como:

$$
p(\beta, h|y) \propto \bar{\chi} \cdot \exp\left\{-\frac{h}{2}\left[(\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta}) + (\hat{\beta} - \underline{\beta})'\left[\underline{V} + (X'X)^{-1}\right]^{-1}(\hat{\beta} - \underline{\beta}) + \nu s^2 + \underline{\nu}\underline{s}^2\right]\right\} \tag{85}
$$

Onde $\bar{\chi} = \frac{h^{\frac{\nu+k}{2}}}{(2\pi)^{\frac{N}{2}}}\chi h^{\frac{k+\nu-2}{2}}$ e χ foi definido previamente. □

A expressão final para a densidade a posteriori de β e h é dada por:

$$
p(\beta, h|y) = \underbrace{\frac{|\bar{V}|^{-\frac{1}{2}}}{\left(\frac{2\bar{s}^{-2}}{\bar{\nu}}\right)^{\frac{\bar{\nu}}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)(2\pi)^{\frac{k}{2}}}}_{\text{Constante}} \underbrace{h^{\frac{\bar{\nu}+k-2}{2}} \exp\left[-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})\right)\right]}_{\text{Equação 76}} \tag{86}
$$

Uma vez que foi utilizada a priori conjugada do modelo, é fácil fazer atualização da posteriori após a coleta de novas informações. Sabendo-se que a distribuição a posterior é uma normal-gama, posteriori obtida pode ser utilizada como uma nova priori e é feita a atualização dos seus parâmetros com as informações provenientes da verossimilhança dos novos dados.

A Equação (86) é uma f.d.p. conjunta para β e h, porém, normalmente estamos interessados apenas em β. Por esse motivo, o termo h muitas vezes é chamado de *de parâmetro incômodo* (Bauwens et al., 2003). Da definição da distribuição normal-gama, sabe-se que a marginal para β é dada por uma distribuição t multivariada, isso é, $\beta \sim t(\bar{\beta}, \bar{s}^2\bar{V}, \bar{\nu})$ (Koop, 2003). A verificação pode ser feita calculando a integral de (86) em relação a h.

Demonstração.

$$
\int p(\beta, h|y)dh = \bar{\chi}\int h^{\frac{\bar{\nu}+k-2}{2}} \exp\left[-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})\right)\right]dh \tag{87}
$$

Definindo $\nu^* \equiv \bar{\nu} + k$ e $\mu^* \equiv \nu^* \cdot \left(\bar{\nu}\bar{s}^2 + (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})\right)^{-1}$, (87) pode ser escrita como o núcleo de uma densidade gama com vetor de parâmetros (ν^*, μ^*) vezes a constante $\bar{\chi}$:

$$
\bar{\chi}\int h^{\frac{\nu^*-2}{2}} \exp\left(-\frac{h}{2}\left[\frac{\nu^*}{\mu^*}\right]\right)dh \tag{88}
$$

Sabendo que $\int \mathcal{G}(\nu^*, \mu^*)dh = 1$ e multiplicando apenas o conteúdo da integral em (88) pelo inverso da constante de integração da densidade gama, obtem-se:

$$
\int \underbrace{\left[\left(\frac{2\mu^*}{\nu^*}\right)^{\frac{\nu^*}{2}}\Gamma\left(\frac{\nu^*}{2}\right)\right]^{-1}}_{a} h^{\frac{\nu^*-2}{2}} \exp\left[-\frac{h}{2}\frac{\nu^*}{\mu^*}\right]dh = 1 \tag{89}
$$

Logo,

$$\int \underbrace{h^{\frac{\nu^*-2}{2}} \exp\left[-\frac{h}{2}\nu^*\right]}_{\text{Núcleo da gama}} dh = \frac{1}{a^{-1}} = a \tag{90}$$

Isto é:

$$\int h^{\frac{\nu^*-2}{2}} \exp\left[-\frac{h}{2}\nu^*\right] = \left[\left(\frac{2\mu^*}{\nu^*}\right)^{\frac{\nu^*}{2}} \Gamma\left(\frac{\nu^*}{2}\right)\right] \tag{91}$$

Usando (91) em (88), obtem-se:

$$p(\beta|y) = \bar{\chi}\left[\left(\frac{2\mu^*}{\nu^*}\right)^{\frac{\nu^*}{2}} \Gamma\left(\frac{\nu^*}{2}\right)\right] = \frac{(2\pi)^{-\frac{k}{2}}|\bar{V}|^{-\frac{1}{2}}}{\Gamma\left(\frac{\bar{\nu}}{2}\right)/\Gamma\left(\frac{\nu^*}{2}\right)}\left[\frac{\left(\frac{2}{\bar{\nu}\bar{s}^2+\bar{Q}}\right)^{\frac{\bar{\nu}+k}{2}}}{\left(\frac{2}{\bar{\nu}\bar{s}^2}\right)^{\frac{\bar{\nu}}{2}}}\right] \tag{92}$$

Onde $\bar{Q} = (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta})$.

Trabalhando apenas a parcela dentro de colchetes na expressão (92), tem-se:

$$\left[\frac{\left(\frac{2}{\bar{\nu}\bar{s}^2+\bar{Q}}\right)^{\frac{\bar{\nu}+k}{2}}}{\left(\frac{2}{\bar{\nu}\bar{s}^2}\right)^{\frac{\bar{\nu}}{2}}}\right] = \left[\frac{\left(\frac{1}{\bar{\nu}\bar{s}^2+\bar{Q}}\right)^{\frac{\bar{\nu}+k}{2}}(2)^{\frac{\bar{\nu}+k}{2}}}{\left(\frac{1}{\bar{\nu}\bar{s}^2}\right)^{\frac{\bar{\nu}}{2}}(2)^{\frac{\bar{\nu}}{2}}}\right]$$

Juntando a última expressão com o termo $(2\pi)^{-\frac{k}{2}}$, (92) pode ser escrita como:

$$\frac{\pi^{-\frac{k}{2}}\Gamma\left(\frac{\bar{\nu}+k}{2}\right)}{|\bar{V}|^{\frac{1}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)}\left[\frac{(\bar{\nu}\bar{s}^2)^{\frac{\bar{\nu}}{2}}}{(\bar{\nu}\bar{s}^2+\bar{Q})^{\frac{\bar{\nu}+k}{2}}}\right] = \frac{\pi^{-\frac{k}{2}}\Gamma\left(\frac{\bar{\nu}+k}{2}\right)}{|\bar{V}|^{\frac{1}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)}\left[\frac{(\bar{s}^2)^{\frac{\bar{\nu}}{2}}(\bar{\nu})^{\frac{\bar{\nu}}{2}}}{(\bar{s}^2)^{\frac{\bar{\nu}+k}{2}}\left(\bar{\nu}+\frac{\bar{Q}}{\bar{s}^2}\right)^{\frac{\bar{\nu}+k}{2}}}\right]$$

Que é igual a:

$$\frac{\bar{\nu}^{\frac{\bar{\nu}}{2}}\Gamma\left(\frac{\bar{\nu}+k}{2}\right)}{\pi^{\frac{k}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)}|\bar{s}^2\bar{V}|^{-\frac{1}{2}}\left[\bar{\nu}+(\beta-\bar{\beta})'(\bar{s}^2\bar{V})^{-1}(\beta-\bar{\beta})\right]^{-\frac{\bar{\nu}+k}{2}} \tag{93}$$

E a expressão acima é a de uma variável aleatória que segue uma distribuição $t(\bar{\beta}, \bar{s}^2\bar{V}, \bar{\nu})$. □

Como a densidade marginal a posteriori possui infinitos valores, é comum que a média a posteriori seja utilizada como estimativa pontual para β:

$$\mathbb{E}[\beta|y] = \int\int \beta \underbrace{\overbrace{p(\beta,h|y)}^{\text{f.d.p. conjunta de } \beta \text{ e } h}\, dh}_{\text{marginal posterior de } \beta} d\beta = \int \beta\left[\int p(\beta,h|y)dh\right]d\beta = \int \beta \underbrace{p(\beta|y)}_{\substack{\text{Densidade}\\ \text{marginal}\\ \text{de } \beta}} d\beta \tag{94}$$

Usando as propriedades da distribuição t multivariada, a esperança dada em (94) será $\mathbb{E}[\beta|y] = \bar{\beta}$ e $Var[\beta|y] = \frac{\bar{\nu}\bar{s}^2}{\bar{\nu}-2}\bar{V}$. Isto significa que, diferente das estimativas clássicas, não é necessária a estimativa da variância h^{-1} para saber informações do parâmetro β.

Integrando (86) em relação a β, podemos obter a f.d.p. posterior marginal de h, que é dada por $p(h|Y) \sim Gamma(\bar{s}^2, \bar{\nu})$, portanto:

$$\mathbb{E}[h|y] = \bar{s}^{-2} \quad \text{e} \quad Var[h|y] = \frac{2\bar{s}^{-4}}{\bar{\nu}} \tag{95}$$

Demonstração. Para encontrar a marginal de h, integramos (86) em relação a β:

$$\begin{aligned}
p(h|y) &= \int p(\beta, h|y)d\beta = \bar{\chi} h^{\frac{\bar{\nu}+k-2}{2}} \int \exp\left\{-\frac{h}{2}\left[\bar{\nu}\bar{s}^2 + \left(\beta-\bar{\beta}\right)'\bar{V}^{-1}\left(\beta-\bar{\beta}\right)\right]\right\} d\beta \\
&= \bar{\chi} h^{\frac{\bar{\nu}+k-2}{2}} \cdot \exp\left\{\frac{-h\bar{\nu}}{2\bar{s}^{-2}}\right\} \underbrace{\int \exp\left\{-\frac{h}{2}\left(\beta-\bar{\beta}\right)'\bar{V}^{-1}\left(\beta-\bar{\beta}\right)\right\} d\beta}_{\text{Integral do núcleo da normal}} \\
&= \bar{\chi} \cdot h^{\frac{\bar{\nu}+k-2}{2}} \cdot \exp\left\{\frac{-h\bar{\nu}}{2\bar{s}^{-2}}\right\} \cdot (2\pi)^{\frac{k}{2}} \cdot |\bar{V}|^{\frac{1}{2}} \cdot h^{-\frac{k}{2}} \\
&= \frac{\cancel{|\bar{V}|^{\frac{1}{2}}}}{\left(\frac{2\bar{s}^{-2}}{\bar{\nu}}\right)^{\frac{\bar{\nu}}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)\cancel{(2\pi)^{\frac{k}{2}}}} \cdot h^{\frac{\bar{\nu}+\cancel{k}-2}{2}} \cdot \exp\left\{\frac{-h\bar{\nu}}{2\bar{s}^{-2}}\right\} \cdot \cancel{(2\pi)^{\frac{k}{2}}} \cdot \cancel{|\bar{V}|^{\frac{1}{2}}} \cdot \cancel{h^{\frac{k}{2}}} \\
&= \left[\left(\frac{2\bar{s}^{-2}}{\bar{\nu}}\right)^{\frac{\bar{\nu}}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)\right]^{-1} \cdot h^{\frac{\bar{\nu}-2}{2}} \cdot \exp\left\{\frac{-h\bar{\nu}}{2\bar{s}^{-2}}\right\}.
\end{aligned} \tag{96}$$

E (96) implica que $h|y \sim \mathcal{G}(\bar{s}^{-2}, \bar{\nu})$ e podemos usar os resultados já existentes da distribuição para encontrar sua média e variância. □

Podemos pensar na priori como uma informação prévia. Por exemplo, imagine que você queira estimar a regra de Taylor para o Brasil e dispõe de dados de artigos para outros países. Uma vez que estamos usando a priori como tendo a mesma distribuição da verossimilhança, podemos interpretar a priori como uma regressão que foi feita anteriormente. Se, neste caso, a revisão de literatura envolve estimativas bem próximas para todos países, podemos dar uma precisão grande para a priori. Por outro lado, se as informações disponíveis forem muito conflitantes, pode-se diminuir o peso da priori alterando sua precisão (maior variância ⇒ menor precisão).

Lembre-se que as seguintes quantidades são usadas na posteriori:

$$\bar{\beta} = \bar{V}(\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}, \quad \bar{V} = (\underline{V}^{-1} + X'X)^{-1}, \quad \bar{\nu} = \underline{\nu} + N\bar{\nu} \quad \text{e} \quad \bar{s}^2 = \underline{\nu}\underline{s}^2 + \nu s^2 + (\hat{\beta}-\underline{\beta})'[\underline{V} + (X'X)^{-1}]^{-1}(\hat{\beta}-\underline{\beta}).$$

Então, para minimizar a influência da informação a priori, podemos escolher $\underline{\nu} \to 0$, ou seja, a "amostra a priori" contém 0 observações[18], além de escolher $\underline{V}^{-1} = c\mathbb{I}_k$ usando $c \to 0$ (ambas precisões de β e h indo para 0). Neste caso, a posteriori seria $p(\beta, h|Y) \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}\bar{\nu})$, na qual:

$$\bar{V}\Big|_{c\to 0} = (X'X)^{-1}, \quad \bar{\beta}\Big|_{c\to 0} = \hat{\beta}, \quad \bar{\nu}\Big|_{c\to 0} = N \quad \text{e} \quad \bar{\nu}\bar{s}^2\Big|_{c\to 0} = \nu s^2 \tag{97}$$

[18] Lembre-se que já argumentamos sobre a relação dos graus de liberdade da priori como sendo uma medida do "número de observações" da priori.

Que são as quantidades de MQO. Isto significa que quando usamos uma a priori não informativa, a esperança da densidade posterior colapsa nos valores de MQO. Isto é, nesse caso, as estimativas bayesianas são iguais às de MQO. Essa priori não informativa pode ser escrita como $p(\beta, h) \propto 1/h$. Entretanto, essa priori não é uma densidade e é o que chamamos de *priori imprópria*. Prioris impróprias implicam abrir mão do uso de densidades de probabilidade e podem acarretar problemas na comparação de modelos.

3.3 Comparação de modelos

Muitas vezes estamos interessados em comparar modelos para saber se devemos ou não incluir uma variável a mais, por exemplo.

3.3.1 Caso de restrições de igualdade

Existem 2 tipos de comparação de modelos com restrições de igualdade:

1. O caso no qual queremos comparar o modelo M_1, que impõe $R\beta = r$ com o modelo M_2, que não possui tais restrições. Neste caso, dizemos que M_1 é *aninhado* a M_2;
2. O caso no qual $M_1 : Y = X_1\beta_1 + \varepsilon_1$ e $M_2 : Y = X_2\beta_2 + \varepsilon_2$, nos quais X_1 e X_2 são matrizes contendo diferentes variáveis explicativas. Esse é o caso de modelos não aninhados, pois M_1 não é caso particular de M_2.

Ambos casos podem ser tratados escrevendo:

$$M_j : Y_j = X_j\beta_{(j)} + \varepsilon_j \tag{98}$$

No qual $j = 1, 2$ indica o modelo, X_j é de tamanho $N \times K_j$ e contém as variáveis explicativas do j-ésimo modelo, $\beta_{(j)}$ é um vetor $K_j \times 1$ de coeficientes do modelo j, ε_j é um vetor $N \times 1$ de erros com distribuição $N(0, h_j^{-1}\mathbb{I}_N)$ e Y_j define a variável dependente.

Vamos denotar a priori normal-gama por:

$$\beta_{(j)}, h_{(j)}|M_j \sim \mathcal{NG}(\underline{\beta}_j, \underline{V}_j, \underline{s}_j^{-2}, \underline{\nu}_j) \tag{99}$$

ou ainda $\beta_{(j)}|h_{(j)}, M_j \sim N(\underline{\beta}_{(j)}, \bar{V}_j h_j^{-1})$, $h_j \sim G(\underline{s}^{-2}, \underline{\nu}_j)$. Com isso, denotamos a f.d.p. posterior por:

$$\beta_{(j)}, h_j|Y_j \sim \mathcal{NG}(\bar{\beta}_{(j)}, \bar{V}_j, \bar{s}_j^{-2}, \bar{\nu}_j), \tag{100}$$

em que $\bar{\beta}_{(j)}, \bar{V}_j, \bar{s}_j^{-2}, \bar{\nu}_j$ foram definidos anteriormente (abaixo da equação (76)).

A ferramenta básica para comparação de modelos na abordagem bayesiana é a *razão de chances* (*odds ratio*). Lembre que $P_j(\theta|Y, M_j) = \frac{P(Y|\theta,M_j)P(\theta,M_j)}{\int P(Y|\theta)P(\theta)d\theta} = \frac{P(Y|\theta,M_j)P(\theta,M_j)}{P(Y,M_j)}$ e queremos justamente encontrar esse denominador para colocar na expressão abaixo

$$PO_{12} = \frac{\mathbb{P}(Y_1|M_1)\mathbb{P}(M_1)}{\mathbb{P}(Y_2|M_2)\mathbb{P}(M_2)}$$

A probabilidade a priori do modelo (j), $\mathbb{P}(M_j)$, precisa ser determinada antes de observar a amostra. Uma escolha usual é $\mathbb{P}(M_1) = \mathbb{P}(M_2) = 0.5$. A função de verossimilhança marginal $\mathbb{P}(Y, M_j)$ é calculada por:

$$\underbrace{\mathbb{P}(Y|M_j)}_{\substack{\text{Prob. dos dados}\\ \text{terem vindo do}\\ \text{modelo j}}} = \int\int \underbrace{\mathbb{P}(Y|\beta_j, h_j)}_{\text{f. verossimilhança}} \underbrace{\mathbb{P}(\beta_j, h_j)}_{\substack{\text{Priori de}\\ \beta,h}} d\beta_j dh_j \tag{101}$$

Diferentemente da maior parte dos modelos, para o MNRL com a priori conjugada natural, é possível calcular (101) analiticamente. Observe que na equação (101) temos uma "verossimilhança marginalizada", isto é, ao integrar $\mathbb{P}(Y_j|M_j)$ em função dos parâmetros nós estamos removendo o efeito que β_j e h_j exercem sobre Y_j para deixar apenas o efeito da estrutura do modelo. A expressão (101), como estamos integrando em relação tanto a β_j como h_j, resulta em uma constante, mas que, dependendo de M_j, assume valores diferentes.

Usando as expressões para a função de verossimilhança e para a priori com os cálculos de (76), temos:

$$\begin{aligned}
\mathbb{P}(Y|M_j) &= \int\int \chi_1 h_j^{\frac{\bar{\nu}_j+k-2}{2}} \exp\left\{-\frac{h_j}{2}\left(\bar{\nu}_j\bar{s}_j^2 + (\beta_j-\bar{\beta}_j)'\bar{V}_j^{-1}(\beta_j-\bar{\beta}_j)\right)\right\} d\beta_j dh_j \\
\text{Onde } \chi_1 &= (2\pi)^{-\frac{N}{2}}\chi \\
&= (2\pi)^{-\frac{N}{2}}(2\pi)^{-\frac{k}{2}}\left[\left(2\frac{\underline{s}^{-2}}{\underline{\nu}}\right)^{\frac{\underline{\nu}}{2}}\Gamma\left(\frac{\underline{\nu}}{2}\right)\right]^{-1}|\underline{V}|^{-\frac{1}{2}}
\end{aligned}$$

E podemos organizar como um produto dos núcleos de uma normal com uma gama:

$$\mathbb{P}(Y|M_j) = \chi_1 \int \underbrace{h_j^{\frac{\bar{\nu}_j+k-2}{2}} \exp\left\{-\frac{h_j}{2}\bar{\nu}_j\bar{s}_j^2\right\}}_{\text{Núcleo de uma Gama para } h_j} \int \underbrace{\exp\left\{-\frac{h_j}{2}(\beta_j-\bar{\beta}_j)'\bar{V}_j^{-1}(\beta_j-\bar{\beta}_j)\right\}}_{\text{Núcleo de uma Normal para } \beta_j|h_j} d\beta_j dh_j. \tag{102}$$

Da distribuição normal multivariada, sabemos que

$$\int \frac{1}{(2\pi)^{\frac{k}{2}}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Y-X)'\Sigma^{-1}(Y-X)\right\} dy = 1.$$

Então:

$$\int \exp\left\{-\frac{1}{2}(Y-X)'\Sigma^{-1}(Y-X)\right\} dy = (2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}} \tag{103}$$

em que Σ representa a matriz de variância e covariância da v.a. Y.

Observe que o lado direito da expressão 103 escrito em termos da distribuição normal para $\beta_j|h_j$ descrita em 102, será:

$$(2\pi)^{\frac{k}{2}}|\bar{V}h^{-1}|^{\frac{1}{2}} = (2\pi)^{\frac{k}{2}}|\bar{V}|^{\frac{1}{2}}h^{-\frac{k}{2}},$$

de maneira que a integral do núcleo da normal em 102 pode ser substituído para obtermos:

$$\begin{aligned}\mathbb{P}(Y|M_j) &= \varkappa_1 \int h_j^{\frac{\bar{\nu}_j+k-2}{2}} \exp\left\{-\frac{h_j}{2}\bar{\nu}_j\bar{s}_j^2\right\}\left(\frac{2\pi}{h_j}\right)^{\frac{k}{2}} |\bar{V}|^{\frac{1}{2}} dh_j \\ &= \varkappa_1 (2\pi)^{\frac{k}{2}} |\bar{V}|^{\frac{1}{2}} \int \underbrace{h_j^{-\frac{k}{2}} \cdot h_j^{\frac{\bar{\nu}_j+k-2}{2}} \exp\left\{-\frac{h_j\bar{\nu}_j}{2\bar{s}_j^2}\right\}}_{\text{Núcleo de uma } Gama(\bar{s}_j^2,\bar{\nu}_j)} dh_j\end{aligned}$$

Fazendo o mesmo processo de substituir a integral do núcleo pelo inverso da constante de integração da distribuição, teremos:

$$\begin{aligned}\mathbb{P}(Y|M_j) &= \varkappa_1 (2\pi)^{\frac{k}{2}} |\bar{V}|^{\frac{1}{2}} \left(\frac{2\bar{s}_j^{-2}}{\bar{\nu}_j}\right)^{\frac{\bar{\nu}}{2}} \Gamma\left(\frac{\bar{\nu}_j}{2}\right) \\ &= \frac{|\bar{V}_j|^{\frac{1}{2}} \left(\frac{2\bar{s}_j^{-2}}{\bar{\nu}_j}\right)^{\frac{\bar{\nu}}{2}} \Gamma\left(\frac{\bar{\nu}_j}{2}\right)}{(2\pi)^{\frac{N}{2}} |\underline{V}_j|^{\frac{1}{2}} \left(\frac{2\underline{s}_j^{-2}}{\underline{\nu}_j}\right)^{\frac{\underline{\nu}_j}{2}} \Gamma\left(\frac{\underline{\nu}_j}{2}\right)} \\ &= \frac{\Gamma\left(\frac{\bar{\nu}_j}{2}\right)(\underline{\nu}_j\underline{s}_j^2)^{-\frac{\underline{\nu}_j}{2}}}{\pi^{\frac{N}{2}}\Gamma\left(\frac{\underline{\nu}_j}{2}\right)} \left(\frac{|\bar{V}_j|}{|\underline{V}_j|}\right)^{\frac{1}{2}} (\bar{\nu}_j\bar{s}_j^2)^{\frac{-\bar{\nu}_j}{2}}\end{aligned}$$

Que pode ser escrito como:

$$\mathbb{P}(Y|M_j) = C_j \left(\frac{|\bar{V}_j|}{|\underline{V}_j|}\right)^{\frac{1}{2}} (\bar{\nu}_j\bar{s}_j^2)^{\frac{-\bar{\nu}_j}{2}} \tag{104}$$

Note que se estivéssemos no caso com a priori imprópria, onde $\underline{V}_j = c\mathbb{I}_K$ com $c \to \infty$, o determinante que está no denominador vai a zero e portanto a razão explode. Isso faz com que nossa priori nem seja uma densidade, e por isso que a razão de chances não funciona. Mas isso não é um problema da razão de chances é um problema decorrente de usarmos uma priori imprópria!

Não precisamos calcular a probabilidade a posteriori pois podemos calcular a razão de chances a partir da probabilidade de cada modelo. Logo, a razão de chances para comparar M_1 e M_2 é:

$$PO_{12} = \frac{\mathbb{P}(Y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(Y|M_2)\mathbb{P}(M_2)} = \frac{C_1\left(\frac{|\bar{V}_1|}{|\underline{V}_1|}\right)^{\frac{1}{2}}(\bar{\nu}_1\bar{s}_1^2)^{\frac{-\bar{\nu}_1}{2}}\mathbb{P}(M_1)}{C_2\left(\frac{|\bar{V}_2|}{|\underline{V}_2|}\right)^{\frac{1}{2}}(\bar{\nu}_2\bar{s}_2^2)^{\frac{-\bar{\nu}_2}{2}}\mathbb{P}(M_2)} \tag{105}$$

Mas de onde veio a equação (105)? Note que a probabilidade do modelo M_1 é:

$$\mathbb{P}(M_1|Y) = \frac{\mathbb{P}(Y|M_1)\mathbb{P}(M_1)}{\int \mathbb{P}(Y|M_1)\mathbb{P}(M_1)dM_1} = \frac{\mathbb{P}(Y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(Y)}$$

Assim, a razão de chances se torna:

$$PO_{12} = \frac{\mathbb{P}(M_1|Y)}{\mathbb{P}(M_2|Y)} = \frac{\mathbb{P}(Y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(Y)} \times \frac{\mathbb{P}(Y)}{\mathbb{P}(Y|M_2)\mathbb{P}(M_2)} = \frac{\mathbb{P}(Y|M_1)\mathbb{P}(M_1)}{\mathbb{P}(Y|M_2)\mathbb{P}(M_2)}$$

De maneira que não precisamos mais determinar $\mathbb{P}(Y)$ para o cálculo de PO_{12}.

Para diversos modelos, teremos:

$$\mathbb{P}(M_1|Y) + \mathbb{P}(M_2|Y) + \ldots + \mathbb{P}(M_m|Y) = 1$$

No caso de $m = 2$, temos que:

$$\mathbb{P}(M_1|Y) + \mathbb{P}(M_2|Y) = 1 \quad \text{e} \quad PO_{12} = \frac{P(M_1|Y)}{P(M_2|Y)}$$

Logo,

$$\mathbb{P}(M_1|Y) = PO_{12}\overbrace{[1 - \mathbb{P}(M_1|Y)]}^{\mathbb{P}(M_2|Y)} = PO_{12} - PO_{12}\mathbb{P}(M_1|Y)$$
$$\Rightarrow \mathbb{P}(M_1|Y)[1 + PO_{12}] = PO_{12} \Rightarrow \mathbb{P}(M_1|Y) = \frac{PO_{12}}{1 + PO_{12}}$$

No entanto, como dito anteriormente, quando usamos a razão de chances para comparar modelos, só será aceitável utilizar a priori não informativa para parâmetros comuns a todos os modelos. A priori própria e informativa será necessária para os outros parâmetros. Perceba em 104 e 105 que se $\nu_1 \to 0$ e $\nu_2 \to 0$ à mesma taxa $c_1 = c_2$ e 105 simplificando, mas manterá a interpretação envolvendo o ajuste dos dados e a coerência da priori e da verossimilhança.Caso $k_1 \neq k_2$, não podemos fazer $V_j^{-1} = c\mathbb{I}_j$ e deixar $c \to 0$, pois $|V_j| = c^{-k_j}$ para $k_j = max\{k_1, k_2\}$ irá explodir mais rápido. Logo, se $k_1 < k_2$, $PO_{12} \to \infty$, enquanto se $k_1 > k_2$, $PO_{12} \to 0$. Em outras palavras, a razão de chances dará suporte infinito do modelo mais parcimonioso independente do que dizem os dados.

3.3.2 Intervalos (ou regiões) de maior densidade posterior (HPDI)

As técnicas bayesianas de comparação de modelos estão baseadas na ideia de que $p(M_j|y)$ resume todo nosso conhecimento e incerteza sobre M_j após ver os dados. No entanto, o cálculo de probabilidades a posteriori para modelos tipicamente irá envolver o uso de prioris informativas.

Antes de comparar modelos, vamos ver algumas definições básicas. Estas definições estarão considerando um vetor de parâmetros β no modelo linear de regressão normal, porém podem ser estendidas para parâmetros de qualquer modelo. Suponha que os elementos do vetor de coeficientes da regressão, β, podem assumir qualquer valor real, i.e. $\beta \in \mathbb{R}^k$. Seja $\omega = g(\beta)$ um vetor de funções m-dimensional de β que está definido em uma região Ω, com $m \leq k$. Defina ainda C como uma região de Ω, denotada por $C \subseteq \Omega$.

Definição 3.3.1. Conjunto crível
O conjunto $C \subseteq \Omega$ é um conjunto de credibilidade $100(1 - \alpha)\%$ com relação a $\mathbb{P}(\omega, y)$ se

$$\mathbb{P}(\omega \in C|y) = \int_C \mathbb{P}(\omega|y)d\omega = 1 - \alpha$$

Supondo que $\omega = g(\beta) = \beta_j$, o intervalo de credibilidade 95% para β_j é qualquer intervalo $[a, b]$ tal que:

$$\mathbb{P}(a \leq \beta_j \leq b \mid y) = \int_a^b \mathbb{P}(\beta_j \mid y)d\beta_j = 0.95$$

Existem inúmeros intervalos que irão satisfazer a definição. Por exemplo, se $\beta_j|Y \sim \mathcal{N}(0, 1)$, então tanto $[-1.96; 1.96]$ quanto $[-1.64; +\infty)$ e $[-1.75; 2.33]$ são intervalos de 95% de credibilidade para β_j, como representado na Figura (16). Para escolher um deles, usualmente escolhe-se o menor intervalo.

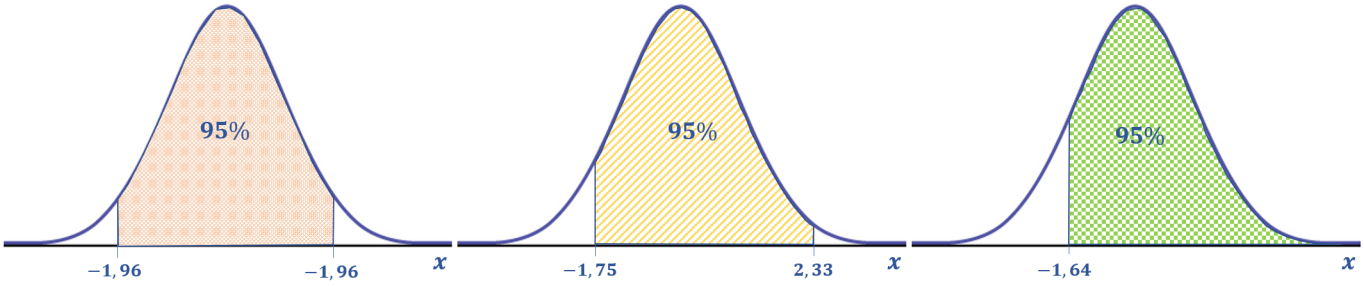


Figura 16: Exemplos de intervalos que correspondem a uma área de 95% na distribuição normal padrão.

Definição 3.3.2. Intervalo de maior densidade a posteriori (HPDI[19])
Um intervalo de credibilidade $100(1-\alpha)\%$ para ω é chamado de *HPDI* se ele possuir a menor área[20] entre todos os intervalos de credibilidade $100(1 - \alpha)\%$.

É comum em aplicações bayesianas apresentar o HPDI em complemento à estimativas pontuais. Por exemplo, a pesquisadora pode apresentar a média à posteriori junto com o HPDI de 95% para β_j, sendo que, neste caso, ela tem uma crença de 95% de que β_j pertence ao intervalo. É tentador pensar em intervalos de credibilidade como análogos aos intervalos de confiança da inferência clássica, porém eles **não** são a mesma coisa. Aqui na inferência bayesiana estamos calculando um intervalo para o qual β_j tem 95% de probabilidade de pertencer, enquanto que na inferência clássica o parâmetro tem sempre probabilidade 0 ou 1 de pertencer a um dado intervalo.

3.4 Previsão

Suponha que queiramos prever T observações do MNRL que ainda não foram observadas. Intuitivamente, podemos querer prever uma observação, 10 observações, etc. A hipótese mais óbvia é que Y^* é não observável porém a variável explicativa X^* é observável. Note que isso não acontece quando estamos num modelo de série de tempo autoregressivo, $Y_t = \rho Y_{t-1} + \varepsilon_t$ - se queremos Y_{t+10} precisaríamos ter Y_{t+9}. Mas, podemos fazer a previsão 1 passo a frente nesse modelo. Em modelos de cross section é mais fácil termos os valores de X^*. Formalmente, queremos encontrar Y^*:

$$Y^* = X^*\beta + \varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}(0, h^{-1}\mathbb{I}_T) \tag{106}$$

[19] Do inglês *Highest Probability Density Interval*.
[20] Note que como estamos lidando com variáveis aleatórias de dimensão maior que 1, os "intervalos" definem uma *região*, onde a probabilidade está no "volume" e a área representa o equivalente ao comprimento do intervalo no caso unidimensional.

Sendo que Y^* é não observável, todas as outras hipóteses continuam válidas e X^* é uma matriz $T \times k$ de *dados observáveis*. Logo, precisamos calcular[21]:

$$\mathbb{P}(Y^*|Y) = \int\int \underbrace{p(Y^*|Y,\beta,h,X^*)}_{\text{veio da eq. 106}} p(\beta,h|Y)d\beta dh \tag{107}$$

Estamos interessados na quantia $\mathbb{P}(Y^*|Y)$ que significa a densidade dos dados não-observados condicionada aos dados que já observamos.

Como ε^* é independente de Y, então Y^* é independente de Y também e:

$$\mathbb{P}(Y^*|Y,\beta,h) = \mathbb{P}(Y^*|\beta,h)$$

No resto das demonstrações, iremos suprimir X^* para economia de notação, assim como foi feito nas aulas anteriores e no livro texto.

A densidade de Y^* condicionada aos parâmetros então será dada por[22]:

$$\mathbb{P}(Y^*|\beta,h) = \frac{h^{\frac{T}{2}}}{(2\pi)^{\frac{T}{2}}} \exp\left\{-\frac{h}{2}(Y^* - X^*\beta)'(Y^* - X^*\beta)\right\} \tag{108}$$

Multiplicando (108) pela f.d.p. a posteriori conjunta de β e h (86) e integrando em relação a β e h temos uma densidade de uma variável aleatória t multivariada (consulte o apêndice de Koop (2003) ou o apêndice das notas de aula para a definição e propriedades) com os seguintes parâmetros:

$$Y^*|Y \sim t\left(X^*\bar{\beta}, \bar{s}^2[\mathbb{I}_T + X^*\bar{V}X^{*'}], \bar{\nu}\right) \tag{109}$$

Demonstração. Usando o mesmo procedimento já apresentado no texto, a verossimilhança dos novos dados pode ser escrita como:

$$\begin{aligned} p(y^*|\beta,h) &= \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{h}{2}(y^* - X^*\beta)'(y^* - X^*\beta)\right\} \\ &= \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{h}{2}\left[\nu s^2 - (\beta-\hat{\beta})'X^{*'}X^*(\beta-\hat{\beta})\right]\right\} \end{aligned} \tag{110}$$

Enquanto que a posteriori dada em (76) pode ser escrita como:

[21] Se fosse previsão para Y_{t+10} no modelo autoregressivo, apareceria uma integral em $Y_{t+i}, i = 1,\ldots,9$ ali na equação (107).

[22] Observe que como Y^* ainda não foram observados, essa expressão é sim uma densidade para a variável aleatória Y^* e não podemos pensar nela como uma verossimilhança

$$p(\beta,h|y) = \underbrace{\frac{|\bar{\nu}|^{-1/2}}{\left(\frac{2\bar{s}^{-2}}{\bar{\nu}}\right)^{\bar{\nu}^2/2}\Gamma\left(\frac{\bar{\nu}}{2}\right)(2\pi)^{k/2}}}_{\bar{\chi}} h^{\frac{\bar{\nu}+k-2}{2}} \exp\left\{-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right)\right\}$$

$$= \bar{\chi}\ \underbrace{h^{\frac{\bar{\nu}+k-2}{2}} \exp\left\{-\frac{h}{2}\frac{\bar{\nu}}{\bar{s}^{-2}}\right\}}_{\textbf{(a)}} \underbrace{\exp\left\{-\frac{h}{2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right\}}_{\text{(b)}} \tag{111}$$

O termo (a) de (111) é quase[23] o núcleo de uma distribuição Gama($\bar{s}^2$, $\bar{\nu}$) para h e o termo em (b) é o núcleo de uma densidade Normal com parâmetros $\bar{\beta}$ e $h\bar{V}$ para $\beta|h$.

Os próximos passos irão mostrar que $y^*|y \sim t\left(X^*\bar{\beta}, \bar{s}^2\left\{\mathbb{I}_N + X^*\bar{V}X^{*'}\right\}, \bar{\nu}\right)$ (onde $\mathbb{I}_N$ é a matriz identidade de dimensão N). Por definição, isso significa que:

$$p(y^*|y) = \frac{1}{C_t}\left|\bar{s}^2\left(\mathbb{I}_N + X^*\bar{V}X^{*'}\right)\right|^{-\frac{1}{2}} \left[\bar{\nu} + (y^* - X^*\bar{\beta})'\left(\bar{s}^2\left(\mathbb{I}_N + X^*\bar{V}X^{*'}\right)\right)^{-1}(y^* - X^*\bar{\beta})\right]^{-\left(\frac{\bar{\nu}+N}{2}\right)} \tag{112}$$

Onde $C_t = \frac{\pi^{K/2}\Gamma\frac{\bar{\nu}}{2}}{\bar{\nu}^{\bar{\nu}/2}\Gamma\left(\frac{\bar{\nu}+N}{2}\right)}$, que é a constante da distribuição t.

Utilizando (110) e (111) em (107), tem-se que:

$$\begin{aligned} p(y^*|y) &= \int\int \frac{h^{N/2}}{(2\pi)^{N/2}}\bar{\chi}\exp\left\{-\frac{h}{2}(y^* - X^*\beta)'(y^* - X^*\beta)\right\} h^{\frac{\bar{\nu}+k-2}{2}} \\ &\quad \exp\left\{-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right)\right\} dh\, d\beta \\ &= \int\int \frac{h^{\frac{N+\bar{\nu}+k-2}{2}}}{(2\pi)^{N/2}}\bar{\chi}\exp\left\{-\frac{h}{2}\bar{\nu}\bar{s}^2\right\} \\ &\quad \exp\left\{-\frac{h}{2}\underbrace{\left[(y^* - X^*\beta)'(y^* - X^*\beta) + (\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right]}_{(*)}\right\} dh\, d\beta \end{aligned} \tag{113}$$

Trabalhando com o termo (*) de (113), obtem-se:

$$\beta'\underbrace{(\bar{V}^{-1} + X^{*'}X)}_{(A)}\beta - \beta'\underbrace{(X^{*'}y^* + \bar{V}^{-1}\bar{\beta})}_{(b)} - (y^{*'}X^* + \bar{\beta}'\bar{V}^{-1})\beta + \underbrace{y^{*'}y^* + \bar{\beta}'\bar{V}^{-1}\bar{\beta}}_{(r)} \tag{114}$$

[23] Integrando a posteriori em relação a β resta um termo $h^{-\frac{k}{2}}$ que é o termo faltante para completar a densidade Gama para h.

Que pode ser visto como $\beta' A\beta - \beta' b - b'\beta + r$, onde o termo r não depende de β. Somando e diminuindo um termo $b'A^{-1}b$ chega-se em:

$$\begin{aligned}
&\beta' A\beta - \beta' b - b'\beta + b'A^{-1}b \underbrace{-b'A^{-1}b + r}_{(c)} = \\
&= \beta' A\beta - \beta' b - b'\beta + b'A^{-1}b + c \\
&= \beta' A\beta - \beta' AA^{-1}b - b'AA^{-1}\beta + b'A^{-1}AA^{-1}b
\end{aligned} \tag{115}$$

Considerando $\bar{\Sigma} = A^{-1}$ e $\bar{\mu} = A^{-1}b$, pode-se reescrever (115) como:

$$\begin{aligned}
&\beta\bar{\Sigma}^{-1}\beta - \beta'\bar{\Sigma}^{-1}\bar{\mu} - \bar{\mu}'\bar{\Sigma}^{-1}\beta + \bar{\mu}'\bar{\Sigma}^{-1}\bar{\mu} + c \\
&= (\beta - \bar{\mu})'\,\bar{\Sigma}^{-1}(\beta - \bar{\mu}) + c
\end{aligned} \tag{116}$$

Então, (113) é:

$$\begin{aligned}
p(y^*|y) = &\int\int \frac{h^{\frac{N+\bar{\nu}+k-2}{2}}}{(2\pi)^{N/2}}\bar{X}\exp\left\{-\frac{h}{2}\frac{\bar{\nu}}{\bar{s}^{-2}}\right\} \\
&\exp\left\{(\beta - \bar{\mu})'\,\bar{\Sigma}^{-1}(\beta - \bar{\mu})\right\}\exp\left\{-\frac{h}{2}c\right\}dh\,d\beta \\
= &\int \frac{h^{(N+\bar{\nu}+k-2)/2}}{(2\pi)^{N/2}}\bar{X}\exp\left\{-\frac{h}{2}\left(\frac{\bar{\nu}}{\bar{s}^{-2}} + c\right)\right\}\int \underbrace{\left[\exp\left\{-\frac{h}{2}(\beta - \bar{\beta})'\bar{\Sigma}^{-1}(\beta - \bar{\mu})\right\}d\beta\right]}_{\text{Núcleo de uma normal para } \beta|h} dh
\end{aligned} \tag{117}$$

Utilizando o resultado de que a integral do núcleo é o inverso da constante de uma densidade Bauwens et al. (2003), a integral em relação a β que aparece em (117) será simplificada:

$$\begin{aligned}
\int \exp\left\{-\frac{h}{2}(\beta - \bar{\beta})'\bar{\Sigma}^{-1}(\beta - \bar{\mu})\right\}d\beta = & \\
&= \frac{\frac{1}{h^{k/2}}}{(2\pi)^{k/2}|\bar{\Sigma}|^{1/2}} \\
&= \frac{(2\pi)^{k/2}|\bar{\Sigma}|^{1/2}}{h^{k/2}}
\end{aligned} \tag{118}$$

De maneira que (117) é:

$$p(y^*|y) = \int \frac{h^{(N+\bar{\nu}-2)/2}}{(2\pi)^{N/2}} \bar{\chi} \frac{h^{k/2}}{h^{k/2}} (2\pi)^{k/2} |\bar{\Sigma}|^{1/2} \exp\left\{-\frac{h}{2}\left(\bar{\nu}\bar{s}^{-2} + c\right)\right\} dh$$

$$= \frac{\bar{\chi}(2\pi)^{k/2}}{(2\pi)^{N/2}} |\bar{\Sigma}^{1/2}| \int h^{(N+\bar{\nu}-2)/2} \exp\left\{-\frac{h}{2}(\bar{\nu}\bar{s}^2 + c)\right\} dh$$

Definindo $\nu^* = N + \bar{\nu}$ e $\frac{\nu^*}{\mu^*} = \bar{\nu}\bar{s}^2 + c$ e substituindo acima, tem-se:

$$p(y^*|y) = \frac{\bar{\chi}(2\pi)^{k/2}}{(2\pi)^{N/2}} |\bar{\Sigma}^{1/2}| \int \underbrace{h^{(\nu^*-2)/2} \exp\left\{-\frac{h}{2}\frac{\nu^*}{\mu^*}\right\}}_{\text{Núcleo de uma Gama}(s^*, \nu^*)} dh \tag{119}$$

Usando novamente o resultado de que a integral do núcleo é o inverso da constante, obtem-se:

$$\int h^{(\nu^*-2)/2} \exp\left\{-\frac{h}{2}\frac{\nu^*}{\mu^*}\right\} dh =$$

$$= \left(2\frac{\mu^*}{\nu^*}\right)^{\nu^*/2} \Gamma\left(\frac{\nu^*}{2}\right)$$

$$= \left(\frac{2}{\bar{\nu}\bar{s}^2 + c}\right)^{\frac{N+\bar{\nu}}{2}} \Gamma\left(\frac{N+\bar{\nu}}{2}\right) \tag{120}$$

Substituindo em (119), as integrais são eliminadas e chega-se em:

$$p(y^*|y) = \frac{\bar{\chi}(2\pi)^{k/2}}{(2\pi)^{N/2}} |\bar{\Sigma}|^{1/2} \left(\frac{2}{\bar{\nu}\bar{s}^2 + c}\right)^{\frac{N+\bar{\nu}}{2}} \Gamma\left(\frac{N+\bar{\nu}}{2}\right) \tag{121}$$

com $\bar{\chi} = \frac{|\bar{V}|^{-1/2}}{\left(2\frac{\bar{s}^2}{\bar{\nu}}\right)^{\bar{\nu}/2} \Gamma\left(\frac{\bar{\nu}}{2}\right)(2\pi)^{k/2}}$.

A expressão em (121) não depende mais dos parâmetros β e h. Os passos seguintes manipulam a expressão para que esteja com a mesma organização que a densidade da distribuição t. O primeiro passo é verificar quem é o produto $|\bar{V}|^{-1/2} \cdot |\bar{\Sigma}|^{1/2}$, em que o termo $\bar{\Sigma} \equiv A^{-1}$ é dado por $\left(\bar{V}^{-1} + X^{*'}X^*\right)^{-1}$.

$$
\begin{aligned}
|\bar{V}|^{-1/2} \cdot |\bar{\Sigma}|^{1/2} &= |\bar{V}|^{-1/2} \cdot |\left(\bar{V}^{-1} + X^{*'}X^{*}\right)^{-1}|^{1/2} \\
&= |\bar{V}|^{-1/2} \cdot |\left(\bar{V}^{-1} + X^{*'}X^{*}\right)|^{-1/2} \\
&= |\bar{V}\left(\bar{V}^{-1} + X^{*'}X^{*}\right)|^{-1/2} \\
&= |\bar{V}\bar{V}^{-1} + \bar{V}X^{*'}X^{*}|^{-1/2} \\
&= |\mathbb{I}_k + \bar{V}X^{*'}X^{*}|^{-1/2} = |\mathbb{I}_N + X^{*}\bar{V}X^{*'}|^{-1/2}
\end{aligned}
$$

Onde na última igualdade foi utilizado o teorema de Sylvester para determinantes (Pozrikidis, 2014). Basta tomar $A = \bar{V}X^{*'}eB = X^{*}$. A é $k \times n$ e B é $n \times k$, de forma que teremos pelo teorema que $\det(\mathbb{I}_k + AB) = det(\mathbb{I}_N + BA)$. Referência adicional: https://terrytao.wordpress.com/tag/sylvester-determinant-identity/.

Assim, $p(y^{*}|y)$ é:

$$
p(y^{*}|y) = \frac{(2\pi)^{k/2}}{(2\pi)^{N/2}(2\pi)^{k/2}} \frac{\Gamma\left(\frac{N+\bar{\nu}}{2}\right)}{\Gamma\left(\frac{\bar{\nu}}{2}\right)} \left|\mathbb{I}_N + X^{*}\bar{V}X^{*'}\right|^{-1/2} \underbrace{\left(\frac{2}{\bar{\nu}\bar{s}^2 + c}\right)^{\frac{N+\bar{\nu}}{2}} \frac{1}{\left(2\frac{\bar{s}^{-2}}{\bar{\nu}}\right)^{\bar{\nu}/2}}}_{\mathcal{A}} \tag{122}
$$

Trabalhando apenas com o termo $\mathcal{A}$ de (122):

$$
\begin{aligned}
\mathcal{A} &= \frac{2^{N/2}2^{\bar{\nu}/2}}{(\bar{\nu}\bar{s}^2 + c)^{\frac{N+\bar{\nu}}{2}}} \frac{1}{2^{\bar{\nu}/2}(\bar{\nu})^{-\bar{\nu}/2}\bar{s}^{-\frac{2\bar{\nu}}{2}}} \\
&= \frac{2^{N/2}\bar{\nu}^{\bar{\nu}/2}\bar{s}^{\bar{\nu}}}{(\bar{\nu}\bar{s}^2 + c)^{\frac{N+\bar{\nu}}{2}}} \\
&= \frac{2^{N/2}\bar{\nu}^{\bar{\nu}/2}\bar{s}^{\bar{\nu}}}{(\bar{s}^2(\bar{\nu} + c\bar{s}^{-2}))^{\frac{N+\bar{\nu}}{2}}} \\
&= \frac{2^{N/2}\bar{\nu}^{\bar{\nu}/2}\bar{s}^{\bar{\nu}}}{\bar{s}^{2\left(\frac{N+\bar{\nu}}{2}\right)}(\bar{\nu} + c\bar{s}^{-2})^{\frac{N+\bar{\nu}}{2}}} \\
&= 2^{N/2}\ \bar{\nu}^{\bar{\nu}/2}\ \bar{s}^{-N}(\bar{\nu} + c\bar{s}^{-2})^{-\frac{N+\bar{\nu}}{2}}
\end{aligned} \tag{123}
$$

Inserindo (123) em (122):

$$
p(y^{*}|y) = \frac{2^{N/2}\ \bar{\nu}^{\bar{\nu}/2}}{2^{N/2}\ \pi^{N/2}} \frac{\Gamma\left(\frac{N+\bar{\nu}}{2}\right)}{\Gamma\left(\frac{\bar{\nu}}{2}\right)} \left|\bar{s}^2\ (\mathbb{I}_N + X^{*}\bar{V}X^{*'})\right|^{-1/2}\ (\bar{\nu} + c\bar{s}^{-2})^{-\frac{N+\bar{\nu}}{2}} \tag{124}
$$

O termo c de (124) foi definido em (114) e é dado por:

$$
\begin{aligned}
c &= r - b'A^{-1}b = y^{*'}y^* + \bar{\beta}'\bar{V}^{-1}\bar{\beta} - b'A^{-1}b \\
&= y^{*'}y^* + \bar{\beta}'\bar{V}^{-1}\bar{\beta} - (X^{*'}y^* + \bar{V}^{-1}\bar{\beta})'\underbrace{(\bar{V}^{-1} + X^{*'}X^*)^{-1}}_{Q^{-1}}(X^{*'}y^* + \bar{V}^{-1}\bar{\beta}) \\
&= y^{*'}y^* + \bar{\beta}'\bar{V}^{-1}\bar{\beta} - X^{*'}y^*Q^{-1}X^{*'}y^* - (\bar{V}^{-1}\bar{\beta})'Q^{-1}X^{*'}y^* \\
&\quad - y^{*'}X^{*'}Q^{-1}\bar{V}^{-1}\bar{\beta} - (\bar{V}^{-1}\bar{\beta})'Q^{-1}\bar{V}^{-1}\bar{\beta} \\
&= y^{*'}(y^* - X^*Q^{-1}X^{*'}y^*) + \bar{\beta}'(\bar{V}^{-1} - \bar{V}^{-1}Q^{-1}\bar{V}^{-1})\bar{\beta} - 2\bar{\beta}'\bar{V}^{-1}Q^{-1}X^{*'}y^* \\
&= y^{*'}\underbrace{(\mathbb{I}_N - X^*(\bar{V} + X^{*'}X)^{-1}X^{*'})}_{(\mathcal{B})}y^* + \bar{\beta}'\underbrace{(\bar{V}^{-1} - \bar{V}^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}\bar{V}^{-1})}_{(C)}\bar{\beta} \\
&\quad - 2\bar{\beta}'\underbrace{\bar{V}^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'}}_{(\mathcal{D})}y^*
\end{aligned} \tag{125}
$$

Os termos destacados em (125) podem ser reescritos da seguinte maneira:

- $(\mathcal{B})$ será $\mathbb{I}_N - X^*(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'} = (\mathbb{I}_N + X^*\bar{V}X^{*'})$, usando o resultado de Searle (1982) que estabelece que $(\mathbb{I} + AB)^{-1} = \mathbb{I} - A(\mathbb{I} + BA)^{-1}B$. Para verificar, basta tomar $A = X^*\bar{V}$ e $B = X'$:

$$
\begin{aligned}
\mathbb{I} - X^*\bar{V}(\mathbb{I} + X^{*'}X^*\bar{V})^{-1}X^{*'} &= \\
&= \mathbb{I} - X^*\bar{V}(\bar{V}^{-1}\bar{V} + X^{*'}X^*\bar{V})^{-1}X^{*'} \\
&= \mathbb{I} - X^*\bar{V}((\bar{V}^{-1} + X^{*'}X^*)\bar{V})^{-1}X^{*'} \\
&= \mathbb{I} - X^*\bar{V}(\bar{V})^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^* \\
&= \mathbb{I} - X^*(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'}
\end{aligned} \tag{126}
$$

- (C) é $(\bar{V}^{-1} - \bar{V}^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}\bar{V}^{-1}) = X^{*'}(\mathbb{I} + X^*\bar{V}X^{*'})^{-1}X^*$, pois:

$$
\begin{aligned}
(\bar{V}^{-1} - \bar{V}^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}\bar{V}^{-1}) &= \\
&= X^{*'}X^* - X^{*'}X^*(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'}X^* \\
&= X^{*'}(X^* - X^*(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'}X) \\
&= X^{*'}(\mathbb{I} - X^*(\bar{V}^{-1} + X^{*'}X^*)^{-1})X^* \\
&= X^{*'}(\mathbb{I} + X^*\bar{V}X^{*'})^{-1}X^*
\end{aligned}
$$

  Onde na primeira igualdade foi utilizado o resultado de Searle (1982) para matrizes dado por $A - A(A + B)^{-1}A = B - B(A + B)^{-1}B$, considerando $A = \bar{V}^{-1}$ e $B = X^{*'}X^*$. Adicionalmente, na última igualdade, foi utilizado o resultado calculado em $(\mathcal{B})$.

- $(\mathcal{D})$ é $\bar{V}^{-1}(\bar{V}^{-1} + X^{*'}X^*)^{-1}X^{*'} = X^*(\mathbb{I} + X^*\bar{V}X^{*'})$ pois:

$$
\begin{aligned}
\overline{V}^{-1}(\overline{V}^{-1} + X^{*'}X^{*})^{-1}X^{*'} = & \\
&= \overline{V}^{-1}(\overline{V}^{-1} + X^{*'}X^{*}\overline{V}\overline{V}^{-1})^{-1}X^{*'} \\
&= \overline{V}^{-1}\left((\mathbb{I} + X^{*'}X^{*}\overline{V})\overline{V}^{-1}\right)^{-1} \\
&= \overline{V}^{-1}(\overline{V}^{-1})^{-1}(\mathbb{I} + X^{*'}X^{*}\overline{V})^{-1}X^{*'} \\
&= (\mathbb{I} + X^{*'}X^{*}\overline{V})^{-1}X^{*'} \\
&= X^{*'}(\mathbb{I} + X^{*}\overline{V}X^{*'})^{-1}
\end{aligned}
$$

Na última igualdade foi utilizado o resultado $(\mathbb{I} + AB)^{-1}A = A(\mathbb{I} + BA)^{-1}$ (Searle, 1982).

Substituindo os termos $(\mathcal{B})$, (C) e $(\mathcal{D})$ na equação (125), tem-se que:

$$
\begin{aligned}
c &= y^{*'}(\mathbb{I} + X^{*}\overline{V}X^{*'})^{-1}y^{*} + \bar{\beta}'X^{*'}(\mathbb{I} + X^{*}\overline{V}X^{*'})^{-1}X^{*}\bar{\beta} - 2\bar{\beta}'X^{*'}(\mathbb{I} + X^{*}\overline{V}X^{*'})^{-1}y^{*} \\
&= (y^{*} - X^{*}\bar{\beta})'(\mathbb{I} + X^{*}\overline{V}X^{*'})^{-1}(y^{*} - X^{*}\bar{\beta})
\end{aligned} \tag{127}
$$

Por fim, substituindo (127) em (124), é obtida a forma final da densidade preditiva a posteriori:

$$
\begin{aligned}
p(y^{*}|y) = \frac{\bar{\nu}^{\bar{\nu}/2}\Gamma\left(\frac{N+\bar{\nu}}{2}\right)}{\pi^{N/2}\Gamma\left(\frac{\bar{\nu}}{2}\right)} & \left|\bar{s}^{2}\,(\mathbb{I}_{N} + X^{*}\overline{V}X^{*'})\right|^{-1/2} \\
& \cdot\left[\bar{\nu} + (y^{*} - X^{*}\bar{\beta})'\left(\bar{s}^{-2}(\mathbb{I} + X^{*}\overline{V}X^{*'})\right)^{-1}(y^{*} - X^{*}\bar{\beta})\right]^{-\frac{N+\bar{\nu}}{2}}
\end{aligned} \tag{128}
$$

□

Logo, sabemos pelas propriedades da f.d.p. t, que:

$$
\mathbb{E}[Y^{*}|Y] = X^{*}\bar{\beta} \quad \text{e} \quad Var[Y^{*}|Y] = \bar{s}^{2}[\mathbb{I}_{T} + X^{*}\bar{V}X^{*'}]\cdot\frac{\bar{\nu}}{\bar{\nu} - 2}
$$

A variância da previsão depende de duas fontes de variação: tanto a incerteza vinda da estimativa de β como a incerteza oriunda dos termos de erro ε que ainda não foram observados. Observe que o termo $X^{*}\bar{V}X^{*'}$ remete ao fato de β ser uma variável aleatória. Já o termo I_{T}, quando multiplicado por $\bar{s}^{2}$ irá remeter aos choques aleatórios dos erros. O mais legal é que não temos apenas uma previsão pontual e sim uma densidade. Podemos então calcular qualquer função de Y^{*}.

Exemplo 3.4.1. Modelo AR

Considere o seguinte modelo:

$$
Y_{t} = \alpha + \rho Y_{t-1} + \varepsilon_{t}
$$

Que pode ser escrito como:

$$
Y = X\beta + \varepsilon
$$

E assuma que $\rho \sim t(\bar{\beta}, \bar{V}, \bar{\nu})$.

Podemos representar nosso vetor Y da seguinte forma:

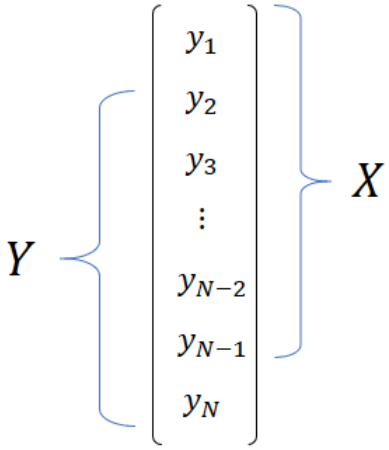


Figura 17: O vetor Y do exemplo 3.4.1

Note que as variáveis dependentes (Y) serão as observações a partir do tempo 2 e as variáveis independentes X serão as primeiras $N-1$ observações.

Caso queiramos fazer previsão para $Y_{N+1} = X^*\beta + \varepsilon^*$, nosso X^* será igual ao Y_N, de forma que:

$$Y_{N+1} = X^*\beta + \varepsilon^*$$

Utilizando o conhecimento visto em aula, sabemos que $Y_{N+1} \sim t(Y_N\bar{\beta}, \bar{s}^2[1 + Y_N\bar{V}Y_N], \bar{\nu})$.

3.5 Média Bayesiana de Modelos (*Bayesian Model Average*)

Como vimos, é possível calcular probabilidade a posteriori para modelos, denotada por $\mathbb{P}(M_j|Y)$ para $j = 1, 2, \ldots, n$. Essas probabilidades podem ser usadas para calcular qual modelo deve ser usado para previsão. Porém, dado a incerteza a respeito do melhor modelo, pode ser interessante "não colocar todos os ovos na mesma cesta". A média Bayesiana de modelos nos permite gerar previsões levando em conta a incerteza relativa ao melhor modelo. Utilizando as regras de probabilidade, podemos escrever:

$$\mathbb{P}(Y^*|Y) = \mathbb{P}(Y^*|Y, M_1)\mathbb{P}(M_1|Y) + \mathbb{P}(Y^*|Y, M_2)\mathbb{P}(M_2|Y) + \ldots + \mathbb{P}(Y^*|Y, M_n)\mathbb{P}(M_n|Y) \tag{129}$$

Usando a propriedade do valor esperado e o teorema do estatístico inconsciente, temos:

$$\mathbb{E}[g(Y^*)|Y] = \mathbb{E}[g(Y^*)|Y, M_1]\mathbb{P}(M_1|Y) + \mathbb{E}[g(Y^*)|Y, M_2]\mathbb{P}(M_2|Y) + \ldots + \mathbb{E}[g(Y^*)|Y, M_n]\mathbb{P}(M_n|Y) \tag{130}$$

que pode ser usada para calcular previsões ponderadas e funções delas.

A ideia da média bayesiana de modelos é a seguinte: vimos na seção de comparação de modelos que podemos calcular probabilidades condicionais de cada modelo dado o que observamos nos dados. Utilizando esse processo podemos escolher o modelo com a maior probabilidade para gerar nossas previsões. Só que, como podemos ter vários modelos com probabili-

dades similares, estamos correndo um risco ao escolher apenas um modelo. A média bayesiana de modelos vai permitir que não seja necessário escolher um único modelo para gerar previsões e possamos considerar todos eles de forma ponderada (com as probabilidades atuando como pesos). Geramos então uma *média ponderada* de previsões, onde modelos com maior chance de serem o melhor modelo terão peso maior. Observe que como temos $g(Y^*)$, podemos trabalhar com funções das não-observadas, como por exemplo as funções de impulso resposta.

4 Parte 4 - MNRL com a priori Normal-Gama independente

Até agora, trabalhamos com uma priori conjugada, fazendo com que a distribuição à priori de β fosse condicional a h. Porém, nem sempre isso será desejável, pois em alguma aplicações iremos querer que os dois parâmetros sejam tratados de forma independente. É possível estabelecer uma priori que comporte a independência, mas o "preço" será não ter mais uma forma fechada para a posteriori.

A densidade a priori conjugada natural do capítulo 3 de Koop (2003) era o produto de $\beta|h \sim \mathcal{N}(\underline{\beta}, \underline{V}h^{-1})$ com $h \sim \mathcal{G}(\underline{s}^{-2}, \underline{\nu})$, ou seja, havia *dependência* a priori entre β e h. Essa dependência implica que alguns tipos de informação a priori não podem ser descritas pela priori conjugada natural. Muitas vezes quando temos o MNRL, temos um palpite inicial para os β's, às vezes baseados em outros estudos, ao mesmo tempo que é mais difícil ter uma informação para h e por isso seria mais interessante manter uma priori difusa ou pouco informativa para este último. Porém, no caso do modelo conjugado, isso implica que a falta de informação de h será carregada para os β's pela estrutura condicional da priori $\beta|h$ utilizada até o momento. Esta "pequena" alteração na priori irá implicar mudanças drásticas na posteriori, como veremos adiante.

Note que a densidade priori marginal para β é $t(\underline{\beta}, \underline{s}^2\underline{V}, \underline{\nu})$, o que nos dá $Var(\beta) = \frac{\underline{\nu}\underline{s}^2}{\underline{\nu}}\underline{V}$, isto é, a variância à priori de β é afetada tanto pelos graus de liberdade da distribuição de h, $\underline{\nu}$ (lembre-se que os graus de liberdade à priori podem ser vistos como o tamanho amostral dos dados da priori), como também por $\underline{s}^{-2}$.Uma priori não informativa para o parâmetro incômodo h significa que sua precisão é pequena e portanto $\frac{2}{\underline{\nu}\underline{s}^4} \to 0$. Isso significa que $\frac{\underline{\nu}\underline{s}^2}{\underline{\nu}} \to \infty$ e com isso a variância de β será muito grande, deixando o parâmetro pouco informativo *independentemente do que escolhermos para* $\underline{V}$. Resumindo, na priori conjugada natural, se fizermos $\underline{\nu}$, $\underline{s}^{-2} \to 0$ para obtermos uma priori não informativa para h, a priori de β também será não informativa. Isso significa que *não é possível ser informativo a respeito de* β *ao mesmo tempo que se é pouco informativo a respeito de* h com as ferramentas vistas no Capítulo 3. Uma forma de resolver esse problema é utilizar uma priori Normal para β que seja independente da priori para h, ou seja, $\beta \sim \mathcal{N}(\underline{\beta}, \underline{V})$ e $h \sim \mathcal{G}(\underline{s}^{-2}, \underline{\nu})$. Porém, note que isso irá implicar que nossa verossimilhança não terá a mesma forma da priori e com isso não teremos mais a posteriori em fórmula fechada.

Iremos continuar com as hipóteses assumidas para o modelo descrito em (54), de forma que a verossimilhança em (63) será a mesma e portanto não iremos repeti-la aqui. O restante deste capítulo terá a priori Normal-Gama independente e iniciará a discussão da posteriori para motivar a introdução de métodos numéricos (no livro do Koop (2003) na seção 3.8 há uma pequena explicação sobre integração usando Monte Carlo que pulamos e vamos trazer agora). A seguir, retomamos ao MNRL com priori normal-gama independente e iremos fazer a simulação a posteriori usando o amostrador de Gibbs.

4.1 A priori Normal-Gama independente

O MNRL foi definido no capítulo 3 e depende dos parâmetros β e h. Inicialmente utilizamos uma priori conjugada natural onde $p(\beta|h)$ é uma densidade Normal e $p(h)$ é uma densidade Gama. Nesta seção iremos utilizar uma estrtura parecida, porém assumindo independência dos parâmetros. Se β e h são independentes a priori, então a $f.d.p.$ conjunta a priori será dada por $p(\beta, h) = p(\beta)p(h)$, pois independência implica que a densidade conjunta pode ser escrita como o produto das marginais.

Usando $\beta \sim \mathcal{N}(\underline{\beta}, \underline{V})$ e $h \sim \mathcal{G}(\underline{\nu}, \underline{s}^{-2})$:

$$p(\beta) = (2\pi)^{-\frac{K}{2}}|\underline{V}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta})\right\} \tag{131}$$

e

$$p(h) = \chi_G^{-1} h^{\frac{\underline{\nu}-2}{2}} \exp\left\{-\frac{h}{2}\frac{\underline{\nu}}{\underline{s}^{-2}}\right\} \tag{132}$$

sendo que $\chi_{\mathcal{G}}$ é a constante de integração da distribuição Gama e é dada por:

$$\chi_G = \left(\frac{2\underline{s}^{-2}}{\underline{\nu}}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)$$

Por simplicidade, o livro adota a mesma notação dos capítulos anteriores.

4.2 A posteriori

A f.d.p. a posteriori é dada pelo produto das f.d.p. a priori (131)-(132) pela função de verossimilhança, dada em (63), o que resulta em:

$$p(\beta, h|y) \propto \chi_{\mathcal{G}}[\exp\{-\frac{1}{2}[\overbrace{h(y-X\beta)'(y-X\beta)}^{\substack{\text{Núcleo de uma}\\ \text{normal para } \beta|h\\ \text{vindo da verossimilhança}}} + \overbrace{(\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta})}^{\substack{\text{Núcleo de uma}\\ \text{normal para } \beta\\ \text{vindo da priori}}}]\}] \overbrace{h^{\frac{N+\underline{\nu}-2}{2}} \exp\left\{-\frac{h}{2}\frac{\underline{\nu}}{\underline{s}^{-2}}\right\}}^{\substack{\text{Combinação do núcleo}\\ \text{gama da priori p/ } h \text{ com}\\ \text{o da verossimilhança}}} \tag{133}$$

A expressão (133) não é o núcleo de nenhuma distribuição conhecida e, portanto, não pode ser vista de forma direta para inferência a posterior. Por exemplo, não existe fórmula analítica para a f.d.p. marginal posterior de β, impossibilitando o cálculo de estimativas pontuais de β em forma fechada. Nestes casos, teremos que usar métodos de simulação para realizar a inferência, comparação de modelos e previsão.

Apesar da f.d.p. posterior conjunta de β e h em (133) não possuir fórmula fechada, as f.d.p. condicionais para $\beta|h, y$ e $h|\beta, y$ terão. Sabemos que $p(\beta|h, y) = \frac{p(\beta,h|y)}{p(h|y)}$ e, como $p(h|y)$ não depende de β, a f.d.p. conjunta $p(\beta, h|y)$ dá o núcleo de $p(\beta|h, y)$, uma vez que $p(h|y)$ é apenas uma constante para $p(\beta|h, y)$. Então, fazendo manipulações similares às feitas para a conjugada natural, temos:

$$p(\beta|h, y) \propto \exp\left\{-\frac{1}{2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right\} \exp\left\{-\frac{1}{2}Q\right\} h^{\frac{N+\underline{\nu}-2}{2}} \exp\left\{-\frac{h}{2}\frac{\underline{\nu}}{\underline{s}^{-2}}\right\} \tag{134}$$

em que

$$\bar{V} = (\underline{V}^{-1} + hX'X)^{-1}, \quad \bar{\beta} = \bar{V}(\underline{V}^{-1}\underline{\beta} + hX'y) \quad \text{e} \quad Q = hy'y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} + \bar{\beta}'\bar{V}^{-1}\bar{\beta}. \tag{135}$$

Por sua vez, podemos ignorar os termos que não dependem de β e escrever:

$$p(\beta|h, y) \propto \exp\left\{-\frac{1}{2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right\} \tag{136}$$

Demonstração. Vamos começar trabalhando com os seguintes termos de (133): $h(y-X\beta)'(y-X\beta) + (\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta})$, abrindo as multiplicações. Nos passos seguintes foi considerado que $\underline{V}^{-1}$ por ser uma matriz de variâncias e covariâncias é simétrica (e portanto é igual ao seu transposto). Além disso, utilizou-se que o transposto de uma constante é a própria constante. Também foi utilizado que $-(\beta'\underline{V}^{-1} + hy'X) = -(\beta'\underline{V}^{-1} + hy'X)'' = -(\underline{V}^{-1}\beta + hX'y)'$.

$$
\begin{aligned}
h(y-X\beta)'(y-X\beta)+(\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta}) &= \\
&= h\left[y-X\beta)'(y-X\beta)\right]+\left[\beta'\underline{V}^{-1}\beta-\beta'\underline{V}^{-1}\underline{\beta}-\underline{\beta}'\underline{V}^{-1}\beta+\underline{\beta}'\underline{V}^{-1}\underline{\beta}\right] \\
&= \left[hy'y-hy'X\beta-h\beta'X'y+h\beta'X'X\beta\right]+\left[\beta'\underline{V}^{-1}\beta-\beta'\underline{V}^{-1}\underline{\beta}-\underline{\beta}'\underline{V}^{-1}\beta+\underline{\beta}'\underline{V}^{-1}\underline{\beta}\right] \\
&= hy'y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}-\underbrace{hy'X\beta}_{3a}-\underbrace{h\beta'X'y}_{2a}+\underbrace{h\beta'X'X\beta}_{1a}+\underbrace{\beta'\underline{V}^{-1}\beta}_{1b}-\underbrace{\beta'\underline{V}^{-1}\beta}_{2b}-\underbrace{\beta'\underline{V}^{-1}\underline{\beta}}_{3b} \\
&= \underbrace{hy'y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}}_{(\bigstar)}+\underbrace{\beta'(\underline{V}^{-1}+hX'X)\beta}_{1=1a\text{ e }1b}-\underbrace{\beta'(\underline{V}^{-1}\underline{\beta}+hX'y)}_{2=2a\text{ e }2b}-\underbrace{(\underline{V}^{-1}\underline{\beta}+hX'y)'\beta}_{3=3a\text{ e }3b}
\end{aligned}
\tag{137}
$$

Os próximos passos irão mostrar como chegar de (137) em $(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})+Q$, com $\bar{\beta}$, $\bar{V}^{-1}$ e Q definidos em (135).

Note que ★ de (137) corresponde aos dois primeiros termos de Q (definido em (135). Então, vamos ver quem é o terceiro termo de Q:

$$
\begin{aligned}
-\bar{\beta}'\bar{V}^{-1}\bar{\beta} &= \\
&= -\underbrace{\left[\bar{V}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]'}_{\bar{\beta}'}\underbrace{(\underline{V}^{-1}+hX'X)}_{\bar{V}^{-1}}\underbrace{\left[\bar{V}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]}_{\bar{\beta}} \\
&= -(\underline{V}^{-1}\underline{\beta}+hX'y)'(\underline{V}^{-1}+hX'X)^{-1}(\underline{V}^{-1}+hX'X)(\underline{V}^{-1}+hX'X)^{-1}(\underline{V}^{-1}\underline{\beta}+hX'y) \\
&= -(\underline{V}^{-1}\underline{\beta}+hX'y)'(\underline{V}^{-1}+hX'X)^{-1}(\underline{V}^{-1}\underline{\beta}+hX'y) = -\mathcal{A}
\end{aligned}
\tag{138}
$$

Agora, vamos somar e diminuir $\mathcal{A}$ em (137)[24]:

$$
\begin{aligned}
hy'y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}+\beta'(\underline{V}^{-1}+hX'X)\beta-\beta'(\underline{V}^{-1}\underline{\beta}+hX'y)-(\underline{V}^{-1}\underline{\beta}+hX'y)'\beta+\mathcal{A}-\mathcal{A} &= \\
&= hy'y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}\underbrace{-\bar{\beta}'\bar{V}^{-1}\bar{\beta}}_{-\mathcal{A}}+\beta'(\underline{V}^{-1}+hX'X)\beta-\beta'(\underline{V}^{-1}\underline{\beta}+hX'y)-(\underline{V}^{-1}\underline{\beta}+hX'y)'\beta+\mathcal{A} \\
&= Q+\beta'(\underline{V}^{-1}+hX'X)\beta-\beta'(\underline{V}^{-1}\underline{\beta}+hX'y)-(\underline{V}^{-1}\underline{\beta}+hX'y)'\beta+(\underline{V}^{-1}\underline{\beta}+hX'y)'(\underline{V}^{-1}+hX'X)^{-1}(\underline{V}^{-1}\underline{\beta}+hX'y) \\
&= Q+\left[\beta'\left(\underline{V}^{-1}+hX'X\right)-\left(\underline{\beta}'\underline{V}^{-1}+hy'X\right)\right]\left[\beta-\left(\underline{V}^{-1}+hX'X\right)^{-1}(\underline{V}^{-1}\underline{\beta}+hX'y)\right] \\
&= \left(\beta'-\left(\underline{\beta}'\underline{V}^{-1}+hy'X\right)\left(\underline{V}^{-1}+hX'X\right)^{-1}\right)\left(\underline{V}^{-1}+hX'X\right)\left[\beta-\left(\underline{V}^{-1}+hX'X\right)^{-1}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]+Q \\
&= \left[\beta'-\left(\underline{V}^{-1}\underline{\beta}+hX'y\right)'\left(\underline{V}^{-1}+hX'X\right)^{-1}\right]\left(\underline{V}^{-1}+hX'X\right)\left[\beta-\bar{V}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]+Q \\
&= \left[\beta-\bar{V}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]'\left(\underline{V}^{-1}+hX'X\right)\left[\beta-\bar{V}(\underline{V}^{-1}\underline{\beta}+hX'y)\right]+Q \\
&= (\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})+Q
\end{aligned}
\tag{139}
$$

De forma que a igualdade do livro vale, isto é,

[24]Essa conta está "de trás para frente" aqui nas notas, mas é mais fácil começar a partir de $(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})$, isto é, começar pelo fim.

$$h(y - X\beta)'(y - X\beta) + (\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta}) = \\ = (\beta - \bar{\beta})'\bar{V}^{-1}(\beta - \bar{\beta}) + Q, \tag{140}$$

o que nos permite escrever (134). □

Note que (136) é o núcleo de uma densidade normal, ou seja,

$$\beta|h \sim N(\bar{\beta}, \bar{V}) \tag{141}$$

Observe que a expressão em 136 nos diz que β, quando conhecemos h, terá uma distribuição normal. Isso seria diferente caso tivéssemos feito o processo de *marginalização*: neste caso, esperaríamos uma distribuição com uma incerteza maior, penalizando nosso desconhecimento de h.

$p(h|\beta, Y)$ é obtida tratando 133 como função de h, logo:

$$p(h|y, \beta) \propto h^{\frac{N+\underline{\nu}-2}{2}} \exp\left\{-\frac{h}{2}\left[(y - X\beta)'(y - X\beta) + \underline{\nu}\underline{s}^2\right]\right\}$$

O que representa um núcleo de uma gama com parâmetros dados por

$$\bar{\nu} = N + \underline{\nu} \quad \text{e} \quad \bar{s}^2 = \frac{(y - X\beta)'(y - X\beta) + \underline{\nu}\underline{s}^2}{\bar{\nu}}, \tag{142}$$

ou seja,

$$h|\beta, y \sim \mathcal{G}(\bar{s}^{-2}, \bar{\nu}) \tag{143}$$

Observe que, apesar de termos fórmulas fechadas para as distribuições à posteriori condicionais, continuamos sem uma fórmula fechada para a posteriori conjunta de β e h. Estamos usando uma priori Normal-Gama independente, o que significa que a priori conjunta é o produto das duas a prioris marginais, mas isso não acontece para a posteriori, isto é, $p(\beta, h|y) \neq p(\beta|h, y)p(h|\beta, y)$. Este resultado decorre da influência da "base de dados", isto é, nosso vetor de observações da variável y utilizado na verossimilhança. Como y é o mesmo para ambos parâmetros, o resultado é que β e h estão "amarrados" um a outro, sendo que a "cordinha" que os une é y. Para ter uma intuição disso, pense no modelo $Y = \alpha + \beta x + \varepsilon$: no momento que o valor de Y está determinado, não podemos alterar α, β e ε indiscriminadamente, pois é necessário manter a igualdade.

Portanto, apesar de conseguirmos escrever analiticamente as posterioris condicionais, não é possível encontrar a posteriori conjunta para β, h, impossibilitando a inferência a respeito desses parâmetros. Entretanto, é possível utilizar métodos numéricos de integração para aproximar os momentos das distribuições posteriori conjunta e marginal. Iremos agora fazer um "parênteses" para estudar métodos de integração numérica e também alguns conceitos de simulação estocástica para poder resolver o problema da posteriori conjunta dada em (133).

4.3 Métodos de Integração

Métodos numéricos têm sido cada vez mais utilizados em aplicações econométricas, tanto na abordagem clássica[25] quanto na abordagem bayesiana. Uma das razões para a popularização destes métodos é o aumento da complexidade dos modelos, o que usualmente leva a soluções que não possuem fórmula analítica fechada e/ou soluções que requerem uma carga computacional intensiva para serem resolvidas. Esta seção é baseada principalmente no capítulo 2 de Moura (2010). As demais referências estão indicadas ao longo do texto.

Sob o enfoque clássico, muitos cálculos envolvem a maximização de uma função objetivo (por exemplo, da função de verossimilhança). Já na inferência bayesiana, estamos basicamente interessadas em obter informações a posteriori, como momentos ou intervalos de credibilidade. Por exemplo, pode ser do nosso interesse calcular a integral do produto de duas funções reais, g e f, em um conjunto $D \subseteq \mathbb{R}^n$:

$$I = \int_D g(x) \cdot f(x)\, dx \tag{144}$$

em que $\int_D f(x)dx = 1$. A função f funciona como uma "ponderação" para a função g, sendo que em algumas aplicações ela pode ser a função identidade (i.e., $f(x) \equiv 1, \forall x \in D$) e neste caso a integral em (144) representa a área sob a curva definida por g. No caso onde $f(x)$ representa a função densidade de probabilidade de uma v.a. contínua X cujo suporte é D, então (144) define a esperança de $g(x)$ (vide a Lei do Estatístico Inconsciente no Anexo 2).

Assumindo $f(x)$ conhecida, tanto nos casos onde $g(x)$ não possui primitiva ou é conhecida apenas em alguns pontos como também naqueles onde é mais fácil de aproximar $\int_D g(x) \cdot f(x)\, dx$ numericamente do que resolver analiticamente será necessário o uso de métodos numéricos. A integral em (144) pode ser aproximada como uma soma ponderada, isto é:

$$I = \int_D g(x) \cdot f(x)\, dx \approx \sum_{i=0}^{n} \omega_i \cdot g(x_i), \tag{145}$$

na qual ω_i são pesos e x_i são pontos no domínio de $g(\cdot)$. Tanto métodos determinísticos como estocásticos podem ser usados para aproximar integrais como foi feito em (145). Enquanto métodos determinísticos levam a resultados precisos para problemas de dimensão baixa, a complexidade computacional destes métodos aumenta exponencialmente com a dimensão da integral, o que é conhecido como *maldição da dimensionalidade*. Os métodos de Monte Carlo, por outro lado, apesar de convergirem mais lentamente que os métodos determinísticos quando usados em problemas unidimensionais, não dependem da dimensão da integral e com isso acabam sendo uma alternativa escalável para problemas de alta dimensionalidade.

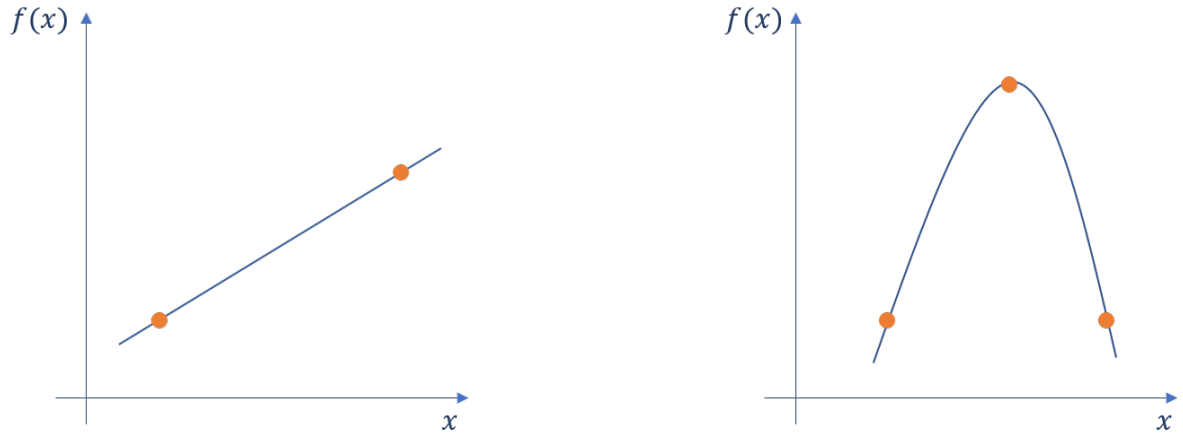
4.3.1 Métodos Determinísticos

O estudo de métodos determinísticos de integração numérica data da antiguidade, quando Arquimedes usou polígonos para encontrar um limite superior e um limite inferior para π. Esta área de pesquisa continuou ativa por muitos séculos, com nomes como Kepler, Newton, Euler, Gauss e outros contribuindo para este campo de pesquisa. Em linhas gerais, os métodos determinísticos se baseiam na teoria da interpolação para avaliar o integrando em um número finito de pontos. Estes pontos são usados para construir uma aproximação polinomial que, por sua vez, é integrada para aproximar I (da equação (144)). Dois teoremas principais são utilizados para dar base a estes métodos: o teorema da interpolação e teorema de Weierstrass.

O teorema da interpolação estabelece que é possível ajustar perfeitamente um polinômio de grau n a um conjunto de $n + 1$

[25] Embora nesta disciplina sejam tratadas as aplicações para a econometria bayesiana, é importante ter em mente que os métodos de integração descritos aqui também podem ser utilizados em aplicações frequentistas.

pontos. Por exemplo, dados dois pontos no plano, só existe uma reta (que é um polinômio de grau 1) passando por eles (Figura 18a). Analogamente, dados três pontos, só existe uma parábola, ou seja, um polinômio de grau 2, que interseciona os três pontos ao mesmo tempo (Figura 18b).



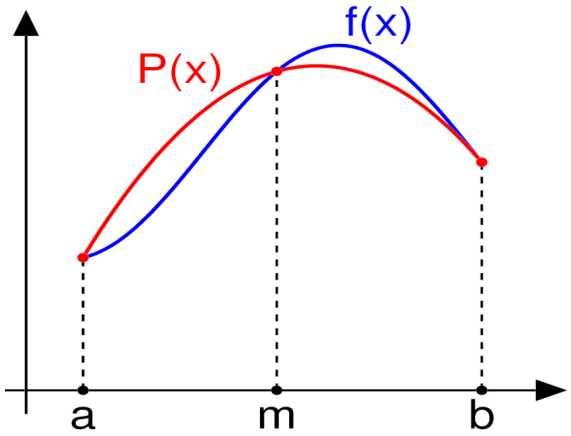
(a) Dois pontos definem uma reta.

(b) Três pontos definem uma parábola.

Figura 18: Exemplos do teorema da interpolação.

O teorema de Weierstrass, por sua vez, estabelece que é possível utilizar polinômios para aproximar qualquer função real contínua definida em um intervalo limitado a qualquer nível de precisão desejado. Os métodos determinísticos irão se diferenciar pelos polinômios que são utilizados para aproximação de $g(x) \cdot f(x)$. Se dizemos que um método é de grau m, quer dizer que ele é exato sempre que o integrando for um polinômio de grau menor ou igual a m, o que decorre dos dois teoremas mencionados.

A Figura (19a) ilustra a situação onde a integral da função $f(x)$ (curva azul) está sendo aproximada pela integral de um polinômio de grau 2, $P(X)$ (curva vermelha). Neste caso, o erro de aproximação será a soma das áreas que estão entre as duas curvas, tanto no intervalo $[a, m]$ quanto no intervalo $[m, b]$.



(a) Aproximação da $f(x)$ por um polinômio $P(x)$.

Figura 19: Exemplos de integração numérica via método de quadraturas.

Para mais detalhes sobre integração determinística, recomenda-se a leitura do capítulo 2 de Moura (2010) e o capitulo 9 de Justo et al. (2018).

4.3.2 Integração de Monte Carlo

A econometria bayesiana tipicamente envolve o uso intensivo de *simulação a posteriori*, sendo que alguns métodos disponíveis, além de integração de Monte Carlo, são amostragem por importância, amostrador de Gibbs e o algoritmo de Metropolis-Hastings. O termo Monte Carlo[26] foi utilizado pela primeira vez na literatura no trabalho de Metropolis e Ulan, em 1949[27]. Seu artigo trouxe um método para resolver problemas sem solução analítica através do uso de um problema estocástico cuja solução pudesse ser aproximada por experimentos estatísticos[28] (Moura, 2010). Uma das vantagens dos métodos de Monte Carlo é que eles não sofrem da maldição da dimensionalidade, diferentemente dos métodos determinísticos.

Reescrevendo o problema (144) com a notação que vínhamos adotando para o MNRL, temos:

$$I = \mathbb{E}[g(\theta)|y] = \int_{\Theta} g(\theta) \cdot p(\theta|y)\, d\theta \tag{146}$$

Logo,

- se $g(\theta) = \theta$, então (180) nos dá a média a posteriori de θ;
- se $g(\theta) = \mathbb{I}(\theta \geq 0)$, então (180) calcula a probabilidade de θ ser positivo;
- se $g(\theta) = p(y^*|y, \theta)$, então (180) calcula a densidade preditiva.

Sendo assim, um método que aproxime (180) nos dará quantidades a posteriori de interesse para o procedimento de inferência a respeito de θ.

A integração de Monte Carlo usa a Lei dos Grandes Números e no Teorema do Limite Central para aproximar uma média populacional por uma média amostral[29]. Para relembrar o que estes teoremas estabelecem, considere $\{X_i\}_{i=1}^N$ é uma a.a. de uma população com densidade qualquer com média μ e variância $\sigma^2 < \infty$ e sejam as quantidades $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ e $s_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X}_N)^2$. Então, $\mathbb{E}[\bar{X}_N] = \mu$ e, pela Lei Forte dos Grandes Números,

$$\bar{X}_N \xrightarrow{\text{a.s.}} \mu \qquad \text{e} \qquad s_N^2 \xrightarrow{\text{a.s.}} \sigma^2 \tag{147}$$

em que a.s. é sigla para *almost surely* e $a \xrightarrow{\text{a.s.}} b$ significa *a converge quase certamente para b* (para uma revisão dos tipos de convergência, consulte Ross (2010)). Além disso, o teorema do limite central garante que $\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$, isto é, $\sqrt{N}(\bar{X}_N - \mu)$ converge em distribuição para uma densidade normal com média 0 e variância σ^2, o que nos permite calcular o quão boa é a aproximação utilizada.

A ideia da integração de MC é avaliar uma integral como a que foi dada em (180) usando uma média amostral obtida a partir de v.a. simuladas da seguinte forma:

$$I = \mathbb{E}[g(\theta)|y] = \int_{\Theta} g(\theta) \cdot p(\theta|y)\, d\theta \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) = \hat{g}_N(\theta) = \bar{I}_N, \tag{148}$$

em que os valores θ_i são amostrados de $p(\theta|y)$ e $\hat{g}_N(\theta) = \bar{I}_N$ é uma v.a. que pode ser vista como um estimador de MC para

[26] O nome Monte Carlo foi escolhido em "homenagem" ao tio de Ulan que costumava apostar no Casino de Monte Carlo, localizado em Mônaco.

[27] Robert (2007) ressalta que não surpreende o fato dos primeiros algoritmos de Monte Carlo terem surgido concominantemente ao surgimento do primeiro computador, uma vez que métodos de MC não são viáveis sem o uso de computadores.

[28] Não confunda *Integração de Monte Carlo* (**MC**) com *Markov Chain Monte Carlo* (**MCMC**). Embora estes métodos se relacionem (como será visto ao longo do curso), essas siglas correspondem a coisas diferentes.

[29] Para ver como estes resultados funcionam em uma simulação, para diferentes tipos de distribuições, acesse: https://istats.shinyapps.io/sampdist_cont/.

$\mathbb{E}[g(\theta|y)]$. Observe que a aproximação feita em (183) tem como fonte de erro não mais a aproximação numérica e sim o fato de que somente uma amostra finita de tamanho n pode ser tomada (quando o resultado do teorema do limite central e da lei dos grandes números é assintótico, i.e., vale para $n \to \infty$). Por exemplo, no caso de $\theta \equiv \beta$ do MNRL, o método que usa MC para obter valores de uma função de β, usando (148), está descrito no Algoritmo (1).

Algoritmo 1: Integração de MC para β no MNRL com priori conjugada natural (adaptado de Koop (2003))

Entrada: Parâmetros da posteriori marginal de β: $\bar{\beta}, \bar{s}^2\bar{V}, \bar{\nu}$
Saída: $\hat{g}_N(\beta)$
início
  para $i = 1, \ldots, N$ **gere**
    1. $\beta^{(i)}$, a partir da posteriori marginal $p(\beta|y)$ dada em (93) - use um gerador de números aleatórios para amostrar de uma distribuição t multivariada;
    2. Calcule $g(\beta^{(i)})$ e armazene este resultado.
  fim
  Tome a média aritmética das N retiradas $g(\beta^{(1)}), g(\beta^{(2)}), \ldots, g(\beta^{(N)})$.
fim

Uma vez que $\bar{I}_N$ é uma variável aleatória, podemos obter informações a respeito da "qualidade" da aproximação de MC para I. Se $\mathbb{E}[g(\theta|y)] < \infty$, então pela lei dos grandes números e teorema do limite central temos que

$$\sqrt{N}\left(\bar{I}_N - \mathbb{E}[g(\theta|y)]\right) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_g^2),$$

para $N \to \infty$, sendo $\sigma_g = var[g(\theta|y)]$. Esta variância, por sua vez, pode ser estimada usando

$$\hat{\sigma}_g^2 = \frac{1}{N}\sum_{i=1}^{N}(g(\theta_i) - \bar{I}_N)^2.$$

Assim, o erro de aproximação da integração de MC é obtido pelas propriedades da distribuição normal:

$$\mathbb{P}\left[-1.96 \cdot \frac{\hat{\sigma}_g}{\sqrt{N}} \leq \bar{I}_N(\theta) - \mathbb{E}[g(\theta|y)] \leq 1.96 \cdot \frac{\hat{\sigma}_g}{\sqrt{N}}\right],$$

de maneira que podemos escolher N para o qual $\bar{I}_N(\theta) - \mathbb{E}[g(\theta|y)]$ é suficientemente pequena com alta probabilidade. A quantia $\hat{\sigma}_g/\sqrt{N}$ é uma estimativa para o *desvio padrão numérico*[30].

A implementação dos métodos de Monte Carlo depende da possibilidade de gerar sequências de variáveis aleatórias (no nosso exemplo, da densidade a posteriori $p(\theta|y)$. Na prática, o que fazemos é utilizar um gerador de números pseudo-aleatórios[31], uma vez que gerar números verdadeiramente aleatórios requer processos físicos que são lentos ou inconvenientes. Existe uma discussão a respeito das implicações do uso de números pseudo-aleatórios que não será abordada aqui, mas pode ser lida na seção 2.2.1 de Moura (2010).

A seguir, vamos ver dois exemplos de integração de Monte Carlo. A parte computacional destes e dos demais exemplos do capítulo estão no repositório da disciplina no Github.

Exemplo 4.3.1. Integrando uma função determinística
Suponha que o objetivo é integrar $I = \int_1^2 \exp(\theta)\, d\theta$ sem usar integração analítica. Reinterprete I como uma esperança em

[30] O desvio padrão numérico é dado por $\sigma_G/\sqrt{N}$ e, de acordo com Koop (2003), pode ser usada como uma medida aproximada do erro de MC.
[31] Neste link (`https://goo.gl/DsYSDx`) há uma breve explicação de como funcionam os geradores de números pseudo-aleatórios.

relação a $\theta \sim \mathcal{U}(1,2)$ (escolhida de maneira conveniente dentro dos intervalos de integração). Sabemos que a densidade de uma $\mathcal{U}[a,b]$ é $\frac{1}{b-a}$, de forma que $p_{\mathcal{U}}(\theta) = 1/(2-1)$. Obtemos então:

$$I = \int_1^2 \exp\{\theta\} d\theta = (2-1) \int_1^2 \exp\{\theta\} \frac{1}{2-1} d\theta = (2-1) \cdot \mathbb{E}_{\mathcal{U}}[\exp(\theta)]$$

Para aproximar a integral, simule N observações de $\theta \sim \mathcal{U}(1,2)$ e aproxime $\mathbb{E}_{\mathcal{U}}[\exp(\theta)]$ através da média amostral:

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} \exp\left(\theta^{(i)}\right)$$

Usando $N = 10.000$, obtemos 4.67057, que é uma aproximação razoável para o valor exato $I = 4.67077$.

Exemplo 4.3.2. A função densidade acumulada da distribuição Normal

A função densidade acumulada (f.d.a.) da distribuição normal padrão é dada por:

$$\Phi(\theta) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\theta^2/2}\, d\theta$$

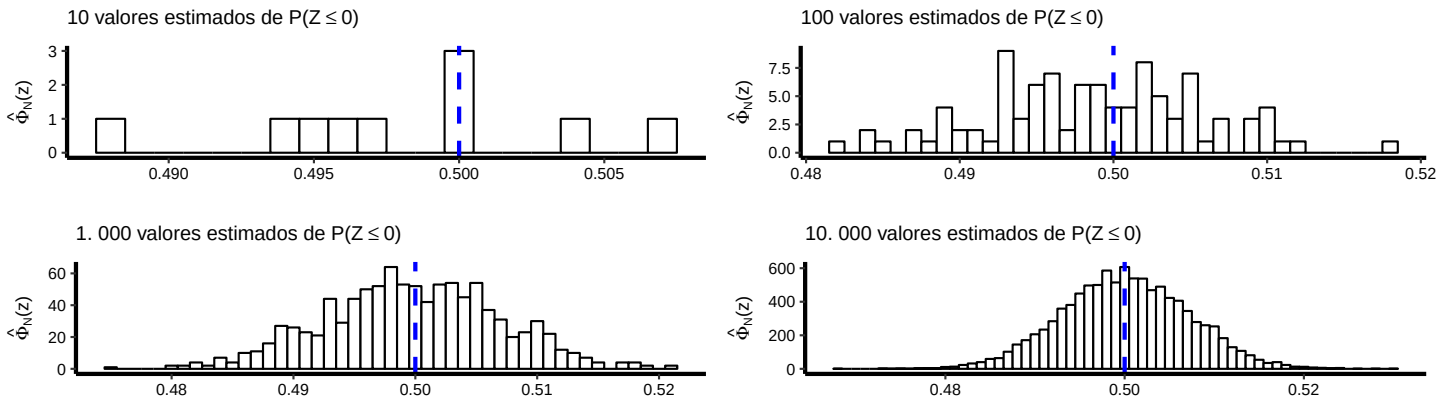
Esta expressão não possui fórmula analítica fechada, o que requer métodos numéricos para seu cálculo. Se forem amostrados i valores de uma normal padrão, isto é, amostrando $\theta^i \sim N(0,1)$, tem-se:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{\theta^2/2}\, d\theta \approx \widehat{\Phi}_N(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\theta^i \leq t)$$

A porção $\widehat{\Phi}_N(t)$ representa uma soma de variáveis aleatórias que seguem distribuição Bernoulli, logo sua variância é $\Phi(t)(1-\Phi(t))/N$ (faça a conta!). Comparando o valor teórico de 0.5 (uma vez que a distribuição normal é simétrica em torno de sua média, espera-se que o 0 divida a área sob a curva em duas parcelas iguais, no caso da normal padrão) com o valor aproximado de 0.498735 obtém-se uma diferença de 0.001265 entre esperado e aproximado. O valor de $\hat{\sigma}_g^2$ é de 1.2487 e portanto o desvio padrão numérico, dado por $\frac{\hat{\sigma}_g}{\sqrt{S}}$, é de 0.0025. Para ilustrar a convergência dos valores, é possível repetir o procedimento diversas vezes. Neste caso, é esperado encontrar uma distribuição de valores simétrica em torno do ponto 0.5, como na figura (20).

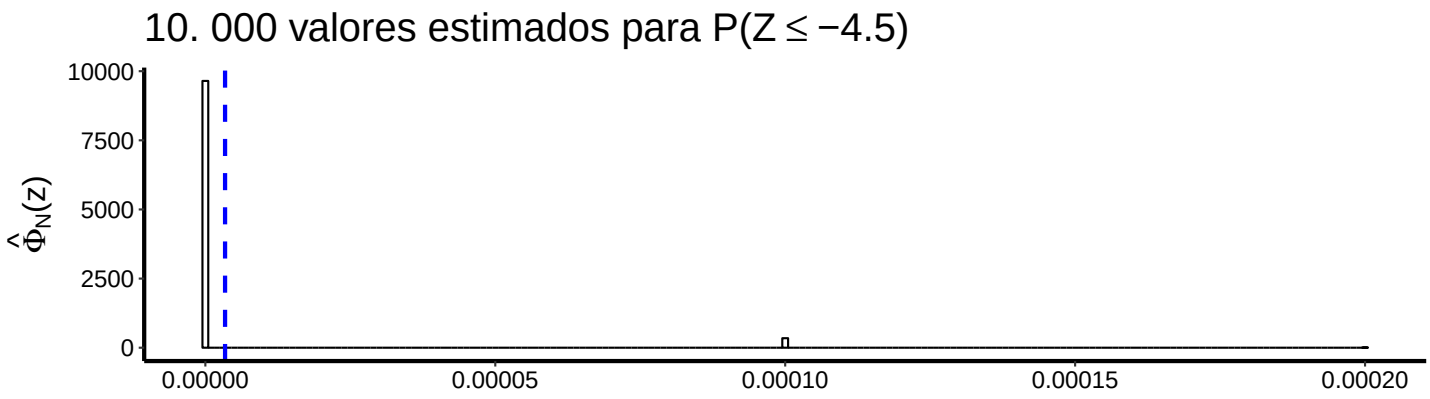
Para t menor que -4.5 precisaremos de muito mais observações ainda para conseguir uma estimativa acurada para esta probabilidade, pois amostrando de uma Normal padrão, raramente iremos amostrar $\theta^i < -4.5$ e, em amostras finitas, com frequência iremos estimar que essa probabilidade é 0. Embora a probabilidade seja baixa, ela não é igual a zero: $\mathbb{P}(X \leq -4.5) = 3.397673110^{-6}$ e por isso nosso resultado utilizando o MC tradicional não é confiável.

Podemos fazer o mesmo procedimento que anteriormente de repetir a simulação e verificar como ficam nossas estimativas. A figura (21) contém 10.000 valores gerados a partir de médias que, por sua vez, foram calculadas a partir de 10.000 valores. Isto é, foram gerados 10.000 valores de uma normal, criou-se uma variável dummy que assumiu valor 1 se o valor simulado era menor que -4.5 e calculou-se a frequência relativa desta variável. Este procedimento, por sua vez, foi repetido 10.000 vezes. Mesmo assim, nota-se que no histograma a maior parte dos valores está concentrada sobre o 0.



Cada valor do gráfico corresponde a uma média calculada a partir de 5.000 valores. A linha tracejada (azul) corresponde ao valor 0.5, que é o verdadeiro valor que se quer aproximar.

Figura 20: Distribuição de 10, 100, 1000 e 10000 valores aproximados de $\mathbb{P}(\theta) \leq 0$ quando $\theta \sim \mathcal{N}(0, 1)$ usando MC.



Cada valor do gráfico corresponde a uma média calculada a partir de 10.000 valores. A linha tracejada (azul) corresponde ao valor 3.397673^{-6}, que é o verdadeiro valor que se quer aproximar. Em 10.000 valores de média calculados, somente 352 não foram iguais a zero.

Figura 21: Distribuição de 10.000 valores aproximados de $\mathbb{P}(\theta) \leq -4.5$ quando $\theta \sim \mathcal{N}(0, 1)$ usando MC.

Sendo assim, o nosso método de MC "falha" para termos estimativas de uma probabilidade na cauda. Isso ocorre porque o domínio da função que estamos amostrando os valores de θ é toda a reta, enquanto que a região que estamos interessados, o suporte da função g, é bastante restrito.

Calcular probabilidade de eventos raros como $\Phi(-4.5)$ usando o método de MC "clássico" é difícil, pois muito raramente iremos amostrar θ^i tal que $\mathbb{I}(\theta^i \leq -4.5) = 1$, logo $\widehat{\Phi}_S(-4.5) = 0$ mesmo para um valor alto de N. Note, porém, que usando a regra de mudança de variáveis, podemos usar $v = \frac{1}{x}$:

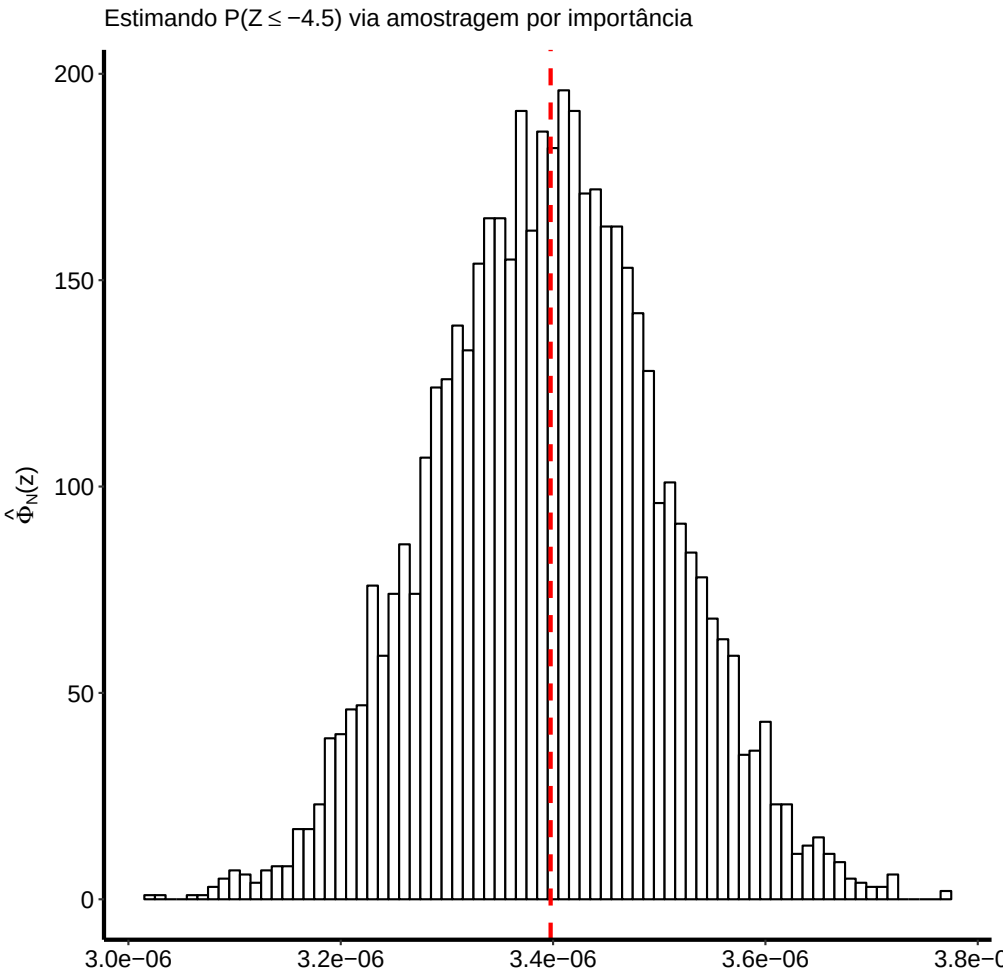
$$\int_{-\infty}^{-4.5} \frac{1}{\sqrt{2\pi}} e^{\theta^2/2}\, d\theta = \int_{\frac{-1}{4.5}}^{0} \frac{\phi(1/v)}{v^2} dv = \frac{1}{4.5} \int_{\frac{-1}{4.5}}^{0} \frac{\phi(1/v)}{v^2} p_U(v) dv.$$

Podemos amostrar $v_i \sim U(-1/4.5, 0)$, então:

$$\int_{-\infty}^{-4.5} \frac{1}{\sqrt{2\pi}} e^{\theta^2/2}\, d\theta \approx \widehat{\Phi}_S^U(-4.5) = \frac{1}{S} \sum_{i=1}^{S} \frac{\phi(1/v^i)}{4.5 v^{i^2}}$$

Note que a F.D.P. de v $p_U(v) = 4.5$ é usada no denominador para compensar o fato de que não amostramos da distribuição original, mas sim de uma distribuição alternativa. Novamente, podemos simular diversos valores e observar seu comporta-

mento. O histograma para 10.000 valores simulados está na figura (22). É possível perceber que agora os valores de fato estão concentrados em torno do verdadeiro valor médio de 3.397673^{-6}.



Histograma de 5.000 valores simulados, onde cada valor do gráfico corresponde a uma média calculada a partir de 10.000 valores. A linha tracejada (vermelha) corresponde ao valor 3.397673^{-6}, que é o verdadeiro valor que se quer aproximar.

Figura 22: Distribuição de 10.000 valores aproximados de $\mathbb{P}(\theta) \leq -4.5$ quando $\theta \sim \mathcal{N}(0,1)$ usando amostragem por importância.

Este exemplo é a motivação do método de Monte Carlo chamado *amostragem por importância*. A ideia principal é encontrar uma maneira de transformar o problema em outro que tenha a mesma solução, mas cujos valores sejam amostrados de uma densidade com mais massa de probabilidade na vizinhança do pico do produto $g(\cdot)f(\cdot)$. Note que, ao amostrar de $f(\cdot)$, estamos ignorando completamente o nosso conhecimento a respeito de $g(\cdot)$. O que a amostragem por importância se propõe a fazer é incorporar essas informações em um novo amostrador diferente de $f(\cdot)$, mas que incorpore as informações que temos tanto a respeito de $f(\cdot)$, quanto de $g(\cdot)$.

4.3.3 Amostragem por importância (Importance Sampling)

Como existem diversos estimadores de Monte Carlo, se torna um problema saber decidir qual das estimativas é a melhor. O critério para esta decisão será com base na variância do estimador. Embora o método de Monte Carlo clássico produza estimativas não-viesadas e seja de fácil implementação, a sua variância pode ser muito grande. Neste sentido, existem

métodos mais eficientes que o método de MC original, que ainda são não viesados e ao mesmo tempo focam na redução da variância das estimativas.

De acordo com Rubinstein (1981), a redução da variância pode ser vista como uma forma de utilizar conhecimento prévio sobre o problema. Em um extremo, quando não se sabe nada a respeito das densidades envolvidas, não é possível reduzir a variabilidade. Por outro lado, se temos total conhecimento do problema, a variância é zero e métodos de MC não seriam necessários. Em suas palavras: "*Variance reduction cannot be obtained from nothing; it is merely a way of not wasting information*".

Para fins de simplicidade, considere que o problema a ser resolvido é obter uma estimativa para a seguinte integral:

$$I = \int g(x)\, dx, \quad x \in D \subset \mathbb{R}^n \tag{149}$$

Supondo que $g \in L^2(D)$ (isso significa que g é uma função quadrado-integrável), isto é, $\int g^2(x)dx$ está bem definida. A ideia da amostragem por importância será concentrar a amostragem dos pontos, utilizando integração de Monte Carlo, nas regiões de D que tem mais "importância", ao invés de amostrar igualmente de toda a região.

Por exemplo (baseado neste site), considere a figura 23, onde comparamos uma distribuição $Beta(2, 2)$ (vermelho) com as uniformes $\mathcal{U}(0, 1)$ (azul) e $\mathcal{U}(0, 5)$ (verde). É claro que ambas uniformes podem ser utilizadas em um algorítmo para amostrar valores da *Beta*, porém a distribuição uniforme que varia entre 0 e 1 tem seus pontos mais concentrados onde a densidade da *Beta* assume seus valores. Para o caso de $X \sim \mathcal{U}(0, 1)$, aproximamos a integral $\int_0^1 g(x)dx = \mathbb{E}(g(X))$ por MC, usando $\frac{1}{n}\sum_1^n g(x_i)$. Como a densidade da Beta é 0 para valores abaixo de 0 e acima de 1, esta aproximação deve funcionar razoavelmente bem. Já se utilizarmos $X \sim \mathcal{U}(0, 5)$, temos $\int_0^1 g(x)dx = 5\mathbb{E}(g(X))$ e o estimador de MC será $\frac{5}{n}\sum_1^n g(x_i)$. Essa aproximação acaba não sendo interessante pois 80% dos valores desta uniforme estão fora do suporte da função $g(\cdot)$ original.

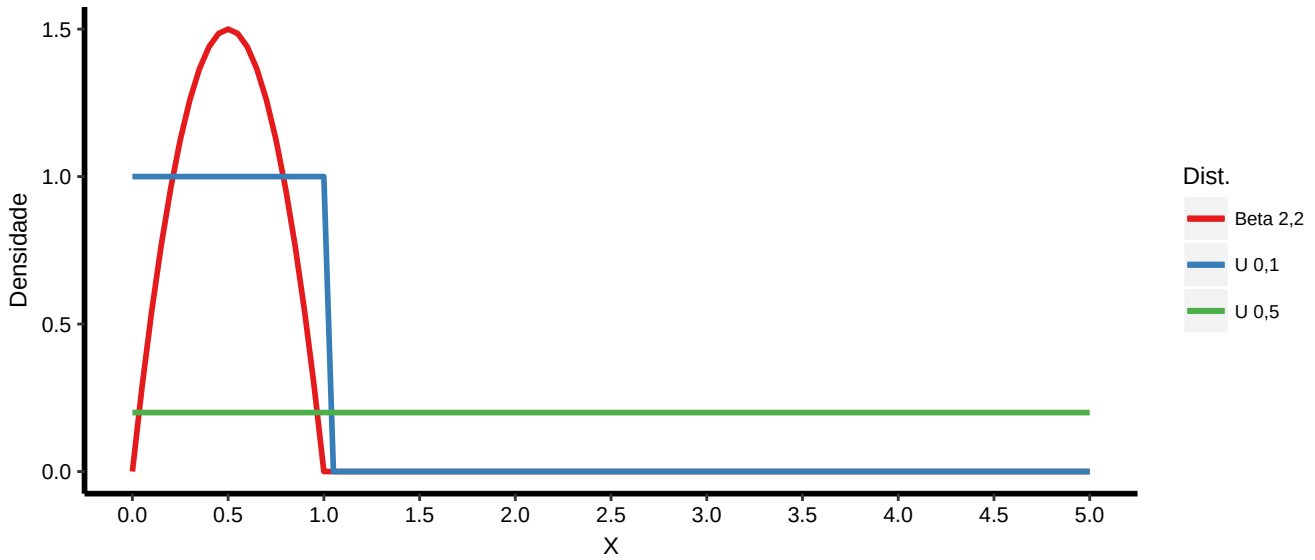


Figura 23: Comparação de uma densidade Beta(2,2) (linha vermelha) com a Uniforme(0,1) (linha azul) e a Uniforme(0,5) (linha verde).

Voltando a Rubinstein (1981), podemos escrever (183) como:

$$I = \int \frac{g(x)}{m_X(x)} m_X(x)dx = \mathbb{E}_m\left[\frac{g(x)}{m_X(x)}\right] \tag{150}$$

em que X é um vetor aleatório com densidade $m_X(\cdot)$ tal que $m_X(x) > 0 \ \forall \ x \in D$. A função $m_X(\cdot)$ é conhecida como *amostrador de importância* (*importance sampler*).

Considere o estimador:

$$\hat{I}_S = \frac{g(X)}{m_X(X)} \tag{151}$$

Ele é não viesado para (184) e sua variância é dada por:

$$Var[\hat{I}_S] = \int \frac{g^2(X)}{m_X(X)} dx - I^2 \tag{152}$$

Podemos então aproximar a integral dada em (184) pegando uma amostra aleatória $X_1, \ldots, X_n$ da densidade $m_X(x)$ e substituir seu valor na equação de média amostral:

$$\theta = \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{m_X(X_i)} \tag{153}$$

Mas voltamos ao problema do exemplo que compara a Beta com as Uniformes. Como escolher a densidade para X de forma a minimizar a variância de $\hat{I}_S$? Os próximos dois resultados nos ajudam a saber qual seria a variância mínima do estimador de MC.

Teorema 4.3.3. ***Variância mínima***

A variância mínima para $\hat{I}_S$ é dada por:

$$Var[\hat{I}_S] = \left(\int |g^2(X)| dx \right)^2 - I^2 \tag{154}$$

E ocorre quando a variável aleatória X tem densidade:

$$m_X(x) = \frac{|g(x)|}{\int |g(x)| dx} \tag{155}$$

Demonstração. (**Teorema 4.3.3**)

A equação (19) aparece quando substituímos (20) em (17).

$$Var[\theta] = \underbrace{\int \frac{g^2(X)}{m_X(X)} dx - I^2}_{(17)} = \int \underbrace{\frac{g^2(X)}{\frac{|g(x)|}{\int |g(x)| dx}}}_{(20)} dx - I^2 = \int \frac{g^2(X) \int |g(x)| dx}{|g(x)|} dx - I^2 = \left(\int |g(x)| dx \right)^2 - I^2 \tag{156}$$

Para verificar que podemos simplificar o quadrado com o módulo, observe que $a^2/|a| = |a|^2/|a| = |a|$. Na dúvida, só fazer para um número real qualquer. Já para demonstrar que $Var[\theta] \leq Var[\hat{I}_S]$, é suficiente mostrar que $\left(\int |g(x)| dx \right)^2 \leq \int \frac{g^2(X)}{m_X(X)} dx$. Este resultado é obtido utilizando a desigualdade de Cauchy-Schwarz:

$$\left(\int |g(x)|dx\right)^2 = \left(\int \frac{|g(x)|}{[m_X(x)]^{\frac{1}{2}}}\,[m_X(x)]^{\frac{1}{2}}\,dx\right)^2$$
$$\leq \int \frac{g^2(x)}{m_X(x)}dx \underbrace{\int m_X(x)dx}_{=1} = \int \frac{g^2(x)}{m_X(x)}dx$$

Pergunta: Por que pode separar as integrais em duas? **Resposta:** Pode-se fazer um produto interno em funções quadrado integráveis (que é o caso aqui pois $g(x) \in L^2$, equivalentemente, todo mundo tem variância finita), que é definido como $< f(x), g(x) >= \int f(x)g(x)dx$. Então, usando a desigualdade de Cauchy-Shwarz, temos:

$$< h(x), l(x) > \leq ||h(x)||||l(x)||$$
$$(< h(x), l(x) >)^2 \leq (||h(x)||)^2(||l(x)||)^2$$
$$(< h(x), l(x) >)^2 \leq < h(x), h(x) >< l(x), l(x) >$$
$$\left(\int h(x)l(x)\right)^2 \leq \int h^2(x)dx \int l^2(x)dx$$

Se tomarmos $h(x) = \frac{|g(x)|}{m_X(x)^{1/2}}$ e $l(x) = m_X(x)^{1/2}$ e substituirmos acima, veremos que o resultado vale e portanto (190) está ok. □

Corolário 4.3.4. *Se $g(x) > 0$, então a densidade ótima $m_X(x)$ será dada por:*

$$m_X(x) = \frac{g(x)}{I} \tag{157}$$

E teremos que $Var[\theta] = 0$.

Observe que este método não é prático, pois a densidade ótima requer o conhecimento de $\int |g(x)|dx$, que acaba sendo praticamente a mesma coisa que conhecer I (de fato serão iguais quando $g(x)$ não muda de sinal). Mas aí usar MC para uma coisa que já se conhece acaba sendo desnecessário. Embora a técnica não seja de valor prático, a ideia de minimizar a variância do estimador de Monte Carlo é bastante útil, conforme veremos no próximo exemplo.

Exemplo 4.3.5. (Adaptado de Robert and Casella (2010b))

Queremos estimar a probabilidade de que uma variável aleatória X, com distribuição de Cauchy de parâmetros (0,1), seja maior do que 2. Isto é, para $X \sim C(0, 1)$, queremos calcular $\mathbb{P}(X \geq 2)$:

$$p = \mathbb{P}(X \geq 2) = \int_2^\infty \frac{1}{\pi(1 + x^2)}dx \tag{158}$$

Imagine que os valores em (158) não sejam de fácil obtenção. Podemos utilizar as ideias de cadeias de markov e, para uma amostra aleatória $X_1, \cdots, X_m$ da distribuição de X, aproximar p de diferentes maneiras.

Método 1

$$p \approx \hat{p}_1 = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}_{X_j>2} \tag{159}$$

A variância do estimador $\hat{p}_1$ pode ser obtida da seguinte maneira:

$$Var[\hat{p}_1] = Var\left[\frac{1}{m}\sum_{j=1}^{m} \mathbb{I}_{X_j>2}\right] = \frac{1}{m^2}\sum_{j=1}^{m}\Big(Var[\mathbb{I}_{X_j>2]}\Big) = \frac{1}{m^2}mp(1-p) = \frac{p(1-p)}{m} \tag{160}$$

E uma vez que $\mathbb{P}(X \geq 2) = 0.15$, a variância do estimador em (24) será dada por $Var[\hat{p}_1] = 0.128/m$.

Método 2

Visando reduzir a variância de (159), podemos propôr outro estimador. Considerando que a distribuição de Cauchy$(0, 1)$ é simétrica em torno do zero, uma estimativa alternativa para p seria:

$$p \approx \hat{p}_2 = \frac{1}{2m}\sum_{j=1}^{m} \mathbb{I}_{|X_j|>2} \tag{161}$$

$$Var[\hat{p}_2] = Var\left[\frac{1}{2m}\sum_{j=1}^{m} \mathbb{I}_{|X_j|>2}\right] = \frac{1}{4m^2}\sum_{j=1}^{m}\Big(Var[\mathbb{I}_{|X_j|>2]}\Big) = \frac{1}{4m^2}\cdot 2mp(1-2p) = \frac{p(1-2p)}{2m} \tag{162}$$

E, novamente usando o fato que $\mathbb{P}(X \geq 2) = 0.15$, a variância do estimador dado em (161) será dada por $Var[\hat{p}_2] = 0.052/m$.

Método 3

Os dois métodos apresentados anteriormente tem uma ineficiência relativa devida à geração de valores fora do domínio de interesse, que neste caso é $[2, +\infty)$. Estes termos "extras" são irrelevantes para a aproximação de p.

Sabendo que $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X < 2)$ e que $\mathbb{P}(X > 2|X > 0) = \frac{1}{2} - \mathbb{P}(0 < X < 2)$, podemos pensar em escrever p como:

$$p = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)}dx \tag{163}$$

Considere agora uma v.a. $X \sim \mathcal{U}(0, 2)$. Sabemos que $f_X(x) = \frac{1}{2-0} = \frac{1}{2}$. Então, multiplicando a integral em (163) por $\frac{2}{2}$, teremos:

$$p = \frac{1}{2} - \int_0^2 \overbrace{\frac{2}{\pi(1+x^2)}}^{h(x)} \underbrace{\frac{1}{2}}_{\text{fdp de } X} dx = \frac{1}{2} - \int_0^2 h(x)f_X(x)dx = \frac{1}{2} - \mathbb{E}[h(X)] \tag{164}$$

A integral em (164) pode ser vista como uma esperança de função de X, isto é, utilizando o lema do estatístico inconsciente podemos enxergar p como uma esperança populacional. Isso significa que ele vai poder ser aproximado por uma média amostral:

$$\hat{p}_3 = \frac{1}{2} - \frac{1}{m}\sum_{j=1}^{m} h(U_j) = \frac{1}{2} - \frac{1}{m}\sum_{j=1}^{m} \frac{2}{\pi}(1 + U_j^2)$$

Onde $U_j \sim \mathcal{U}(0, 2)$. Para calcular a variância de $\hat{p}_3$, utilizamos:

$$\begin{aligned} Var(\hat{p}_3) &= 0 - Var\left(\frac{1}{m}\sum_{j=1}^{m} h(U_j)\right) \\ &= \frac{1}{m^2}\sum_{j=1}^{m} Var(h(U_j)) \\ &= \frac{1}{m^2} \cdot mVar(h(U_j)) \\ &= \frac{1}{m} Var(h(U_j)) \end{aligned}$$

Então, podemos utilizar a forma $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ na expressão acima para obter:

$$Var(\hat{p}_3) = \frac{1}{m}\mathbb{E}(h^2(U)) - \mathbb{E}(h(U)) \tag{165}$$

Como $U \sim \mathcal{U}(0, 2)$, estas esperanças são calculadas utilizando integrais. As integrais são obtidas usando integrais de funções trigonométricas. Lembrando que $\int 1/(a^2 + x^2) = (1/a)tan^{-1}(x/a) + c$, temos que a segunda integral será dada por:

$$\begin{aligned} \mathbb{E}[h(U)] &= \int_0^2 \underbrace{\frac{2}{\pi(1^2 + u^2)}}_{h(U)} \underbrace{\frac{1}{2}}_{\text{fdp de } U} du \\ &= \frac{1}{\pi}\int_0^2 \frac{1}{\pi(1^2 + u^2)} du \\ &= \frac{1}{\pi}(tg^-1(u))\Big|_0^2 \\ &= \frac{1}{\pi} tg^{-1}(2) \end{aligned}$$

Logo, temos que $\mathbb{E}[h(U)] = 0.3524$ e portanto $(\mathbb{E}[h(U)])^2 = 0.1242$.

De maneira similar,

$$\mathbb{E}[h^2(U)] = \int_0^2 \underbrace{\left(\frac{2}{\pi(1^2 + u^2)}\right)^2}_{h^2(U)} \underbrace{\frac{1}{2}}_{\text{fdp de } U} du = \frac{2 + 5tg^{-1}(2)}{5\pi^2}$$

Assim, $\mathbb{E}[h^2(U)] = 0.1527$ e temos $Var(\hat{p}_3) = \frac{1}{m}\mathbb{E}(h^2(U)) - \mathbb{E}(h(U)) = 0.0285/m$.

Método 4

Considere agora uma v.a. $Y \sim \mathcal{U}(0, 1/2)$. Sabemos que $f_Y(y) = \frac{1}{1/2-0} = \frac{1}{1/2} = 2$. Podemos fazer uma transformação de variáveis na expressão (158) utilizando $Y = \frac{1}{X}$, de forma que:

$$
\begin{aligned}
x &= \frac{1}{y} \\
dx &= -\frac{1}{y^2} = -y^{-2} \\
x = 1/2 &\Rightarrow y = 2 \\
x \to \infty &\Rightarrow y = 0
\end{aligned}
$$

Como os limites de integração precisarão trocar de lugar, a integral ganha um sinal de menos que irá cancelar com o sinal negativo do dx, de forma que (158) será:

$$
p = \mathbb{P}(X \geq 2) = \mathbb{P}(0 < Y < 1/2) = \int_0^{\frac{1}{2}} \frac{y^{-2}}{\pi(1 + y^{-2})} dy
$$

Observe ainda que $\frac{y^{-2}}{(1+y^{-2})} = \frac{1}{y^2(1+y^{-2})} = \frac{1}{y^2+y^0} = \frac{1}{1+y^2}$ e portanto a expressão acima pode ser escrita como:

$$
p = \int_0^{\frac{1}{2}} \frac{1}{\pi(1 + y^2)} dy.
$$

Tome $h(Y) = \frac{2}{\pi(1+y^2)}$. Então, $\frac{1}{4}h(Y) = \frac{2}{4\pi(1+y^2)} = \frac{1}{2}\frac{1}{\pi(1+y^2)}$, que é a expressão de p. Portanto:

$$
p = \int_0^{\frac{1}{2}} \frac{1}{\pi(1 + y^2)} dy = \int_0^{\frac{1}{2}} \frac{1}{\pi(1 + y^2)} \underbrace{\frac{2}{2}}_{\text{fdp de } Y} dy = 2 \cdot \mathbb{E}\left(\frac{1}{4}h(Y)\right) = \frac{1}{2}\mathbb{E}(h(Y)) \tag{166}
$$

A esperança em (166) pode ser aproximada por uma média amostral:

$$
\hat{p}_4 = \frac{1}{4m} \sum_{j=1}^{m} h(Y_j) \tag{167}
$$

Usando o mesmo método, calculamos a variância de $\hat{p}_4$:

$$
Var[\hat{p}_4] = \frac{1}{16m^2} \sum_{j=1}^{m} Var[h(Y_j)] = \frac{m}{16m^2} Var[h(Y_j)] = \frac{Var[h(Y_j)]}{16m}
$$

Uma vez que $Var[h(Y_j)] = \mathbb{E}[h^2(Y_j)] - \mathbb{E}[h(Y_j)]^2$, teremos que calcular cada um dos termos, também utilizando integração por partes.

$$\mathbb{E}[h(Y_j)] = \frac{4}{\pi} tg^{-1}(1/2)$$
$$\mathbb{E}[h^2(Y_j)] = \frac{4(2 + 5tg^{-1}(1/2))}{5\pi^2}$$

Então, $Var[h(Y_j)] = \mathbb{E}[h^2(Y_j)] - \mathbb{E}[h(Y_j)]^2 = 9.37510^{-5}/m$.

Formalmente, a amostragem por importância é um método de monte carlo que visa reduzir a variância (das estimativas) ao amostrar de uma densidade mais apropriada do que a f.d.p. original. Suponha que você deseje calcular a esperança de uma $g(\theta)$ cuja densidade é dada por $p(\theta|y)$, porém esta expressão é desconhecida ou é difícil obter amostras de seus valores. Podemos então utilizar um truque matemático combinado com a ideia de MC para obter uma aproximação para esta esperança, como descrito na definição (4.3.6).

Definição 4.3.6. Amostragem por Importância
O método da *amostragem por importância*[32] consiste em avaliar (180) baseado numa amostra $\theta_1, \theta_2, \ldots, \theta_n$ de uma dada distribuição m.

O primeiro passo é reescrever I

$$I = \int_{\Theta} g(\theta)p(\theta|y)d\theta = \int_{\Theta} \frac{g(\theta)p(\theta|y)}{m(\theta)} m(\theta)d\theta = \mathbb{E}_m\left[\frac{g(\theta)p(\theta|y)}{m(\theta)}\right] \tag{168}$$

e então aproximamos a expressão em (168) por:

$$I \approx \hat{I}_s(\theta) = \frac{1}{S}\sum_{i=1}^{S} \frac{g(\theta^i)f(\theta^i)}{m(\theta^i)} \tag{169}$$

O estimador em (169) converge para (168) pelos mesmos motivos que o estimador tradicional de Monte Carlo, qualquer que seja a escolha da função m (contanto que o suporte de m seja comum ao suporte de $f \cdot g$).

A amostragem de $p(\theta|y)$ ignora as informações contidas em $g(\theta)$, tornando o processo ineficiente, pois se o suporte comum das duas distribuições é pequeno[33] o processo se torna ineficiente. Um bom amostrador por importância $m(\theta)$ deve concentrar as observações θ^i nas regiões de maior importância de $g(\theta)p(\theta|y)$. A divisão por $m(\theta)$ em $\hat{I}_S$ leva em consideração as diferenças entre $m(\theta)$ e $p(\theta|y)$.

A variância do estimador em (169) será dada por:

$$Var\left[\hat{I}_s(\theta)\right] = \mathbb{E}_m\left[\hat{I}_s{}^2\right] - \mathbb{E}_m\left[\hat{I}_s\right]^2 = \int g^2(\theta)\frac{f^2(\theta)}{m(\theta)}d\theta - I^2 \tag{170}$$

Então, para que a variância de $\hat{I}_S$ seja finita, precisamos que $\int g^2(\theta)\frac{f^2(\theta)}{m(\theta)}d\theta < \infty$, isto é, a densidade $m(\cdot)$ deve ter caudas mais pesadas que $f(\cdot)$. Um estimador por importância considerado bom será aquele que minimiza a quantidade em (169). Isto ocorre quando $m(\cdot)$ se assemelha ao comportamento do produto $g(\cdot)f(\cdot)$.

[32] Do inglês *Importance Sampling*.
[33] Por exemplo, se o suporte de $p(\theta|y)$ varia em toda a reta real e $g(\theta)$ está concentrada em um intervalo como $[0, 1]$, precisaríamos de muitos valores da primeira até obter alguns poucos que caiam dentro do suporte da segunda.

Exemplo 4.3.7. A distribuição t de Student (adaptado de Robert and Casella (2010b)[34])

Suponha que queremos amostrar $\theta \sim t(\nu, 0, 1)$ para calcular:

$$\int_{2.1}^{\infty} \theta^5 p(\theta|y) d\theta$$

em que

$$p(\theta|y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\theta^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Note que podemos enxergar a quantia de interesse como uma esperança de $h(\cdot)$, quando $h(\theta) = \theta^5 \, \mathbb{I}_{[2.1,\infty)}(\theta)$. É possível amostrar diretamente de $p(\theta|y)$, uma vez que ela é a razão de uma normal padrão pela raiz quadrada de uma gama com ambos parâmetros iguais a $\nu/2$ (isto é, uma v.a. que segue uma qui-quadrado com ν graus de liberdade). No entanto, esse método pode ser muito custoso para valores de ν muito altos.

Vamos comparar diferentes amostradores por importância:

1. Cauchy: $C(0, 1)$;
2. Normal: $\mathcal{N}(0, 1)$;
3. Uniforme: $\mathcal{U}(0, 1/2.1)$.

Note que $h(\theta) = \theta^5 \, \mathbb{I}_{[2.1,\infty)}(\theta)$ tem um domínio restrito e uma forma de lidar com isso é considerar distribuições candidatas que levem essa informação em consideração. Uma distribuição uniforme com parâmetros 0 e $1/2.1$ é uma escolha boa, pois neste caso poderemos usar uma transformação de variáveis em que $x = 1/u$. Detalhadamente, queremos calcular

$$\int_{2.1}^{\infty} \theta^5 p(\theta|y) d\theta$$

Para isso, tome $\theta = \frac{1}{u}$. Assim, $d\theta = -\frac{1}{u^2} du$. Para saber quem são os limites de integração com relação a nova variável, perceba que $\theta = \frac{1}{u} \to u = \frac{1}{\theta}$ e, portanto, $\theta = 2, 1 \to u = \frac{1}{2,1}$. Analogamente, $\theta \to \infty$ implica $u = 0$. Isso nos dará os limites de integração "virados", indo do maior para o menor valor. Para invertê-los, temos que colocar um sinal de menos na frente da integral - e vamos utilizar o sinal que aparece no cálculo de du. Portanto:

$$\begin{aligned}
\int_{2.1}^{\infty} \theta^5 p(\theta|y) d\theta = & \\
&= \int_{1/2.1}^{0} \left(\frac{1}{u}\right)^5 p\left(\frac{1}{u}\right)\left(-\frac{1}{u^2}\right) du \\
&= \int_{0}^{1/2.1} u^{-7} p(1/u) du = \frac{1}{2.1} \int_{0}^{1/2.1} 2.1 u^{-7} p(1/u) du.
\end{aligned}$$

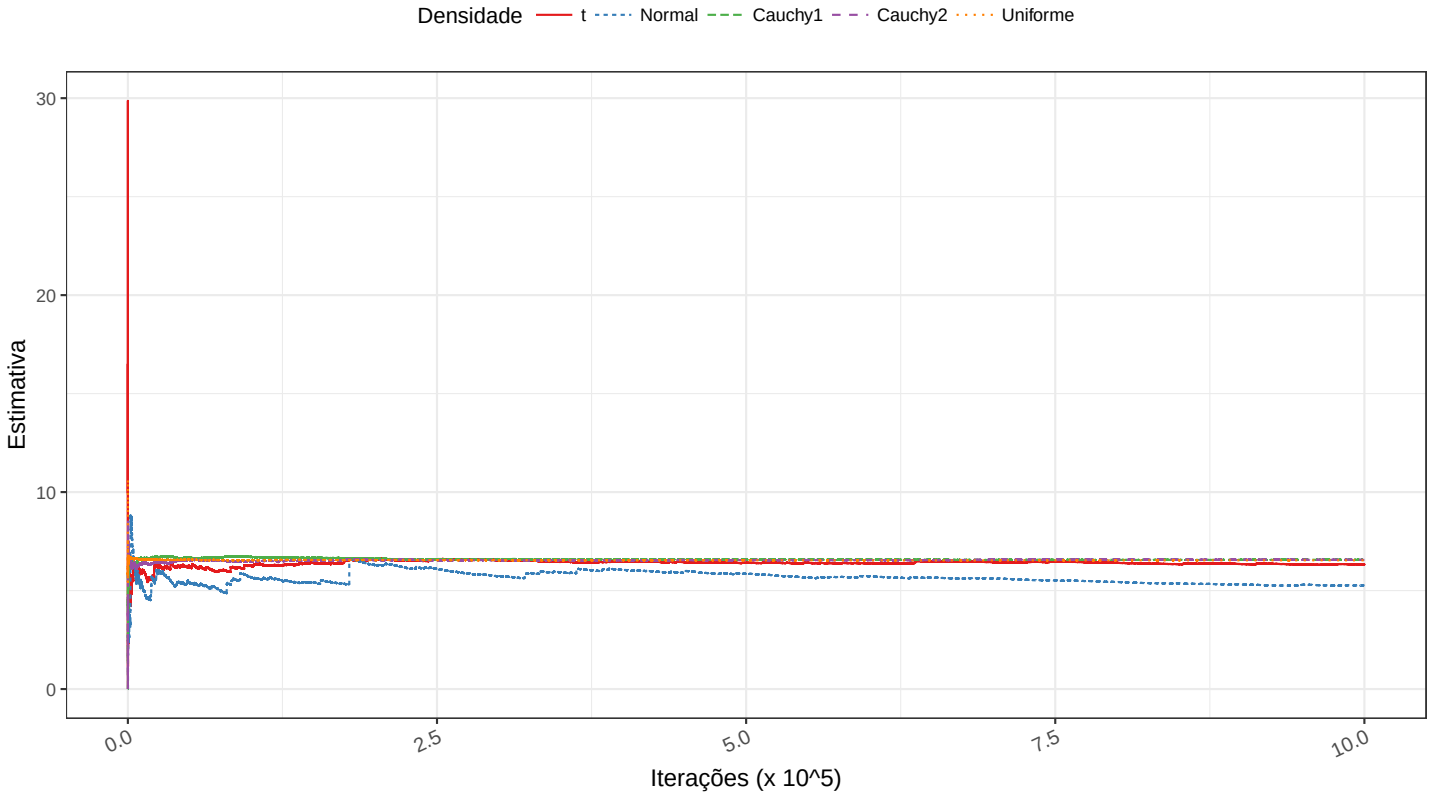
O último passo foi feito para que possamos enxergar a integral como escrita em termos de uma variável aleatória uniforme,

[34] O exemplo do livro é mais geral e é recomendado como exercício que se tente reproduzir ele.

de maneira que o amostrador por importância correspondente será

$$\hat{I} = \frac{1}{2.1 \cdot m} \sum_{j=1}^{m} u_j^{-7} f(1/u_j)$$

em que $u_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1/2.1)$. A figura (24) contém uma comparação para o cálculo de $\mathbb{E}[\theta^5]$ utilizando diferentes amostradores por importância. Note que usar a distribuição t (que é a distribuição original de θ) não é a opção mais eficiente, uma vez que a convergência é demorada e ocorre apenas por volta de $N = 12.500$. A distribuição Normal também se mostra uma escolha ruim: uma vez que suas caudas são mais leves que a distribuição t, o quociente entre as duas densidades acaba ficando muito alto, levando a problemas na variância das estimativas. O mesmo não acontece utilizando a Cauchy, pois ela tem caudas mais pesadas que a t.



O gráfico contém a evolução dos valores estimados usando amostragem por importância. No eixo X está o valor de N usado e o eixo Y contém a respectiva estimativa. Para a distribuição candidata Cauchy foram utilizados dois métodos: no primeiro, usou-se a função rcauchy() e dcauchy() e no segundo utilizou-se a equivalência da densidade Cauchy com a densidade t com um grau de liberdade. As densidades utilizadas como candidatas foram: distribuição $t(0, 12)$ (linha vermelha), $\mathcal{N}(0, 1)$ (linha azul), Cauchy$(0, 1)$ (gerada a partir da função rcauchy() - linha verde), Cauchy$(0, 1)$ gerada a partir de uma distribuição t com um grau de liberdade - linha roxa) e $\mathcal{U}(0, 1/2.1)$ (linha laranja).

Figura 24: Evolução da convergência de 100.000 valores aproximados de $\mathbb{P}(\theta) \geq 2.1$ quando $\theta \sim t(0, 12)$ usando amostragem por importância.

4.4 Introdução a Cadeias de Markov e MCMC

Até o momento, estudamos o método de Monte Carlo com o foco de obter valores i.i.d. de uma densidade de interesse f, tanto de maneira direta, como de maneira indireta (via amostragem por importância). Agora iremos mudar o foco para métodos que geram uma amostra *correlacionada* de valores a partir de uma Cadeia de Markov (do inglês *Markov Chain*). A

teoria de processos estocásticos (mais especificamente, as propriedades de cadeias de Markov) pode ser explorada para obter densidades candidatas mesmo quando um amostrador por importância não é facilmente encontrado. Mais especificamente, métodos de MCMC colocam poucos "requisitos" nas densidades candidatas f, mesmo quando não conseguimos encontrar um amostrador por importância de maneira fácil. Além disso, pode-se dividir um problema de alta dimensão em vários pequenos problemas de menor dimensão, tornando o processo mais eficiente (Robert and Casella, 2010a).

Métodos de Cadeias de Markov via Monte Carlo (MCMC) foram introduzidos na econometria por volta de 1990, após serem desenvolvidos na estatística. Um dos motivos para o sucesso da técnica nos trabalhos de econometria bayesiana é sua facilidade de uso em comparação com a amostragem por importância. Apesar de ser um método de redução de variância de MC, IS torna-se pouco eficaz pois é uma técnica que se ajusta pontualmente aos problemas: é necessário encontrar uma boa aproximação para a densidade a posteriori e um amostrador por importância de uma aplicação (seja por causa de um modelo ou até mesmo um determinado conjunto de dados) pode não servir para outras aplicações. Os métodos de MCMC, por outro lado, se mostram mais fácil de implementar sem a necessidade de despender tanto tempo procurando pelas densidades candidatas (Bauwens et al., 2003).

Por não imporem muitas restrições nas densidades candidatas, os métodos de MCMC têm ampla aplicabilidade. No entando, sua performance é bastante variável e depende da complexidade do problema que se quer resolver. Intuitivamente, o método consiste em produzir aproximações para integrais e outras quantidades de interesse a partir de uma Cadeiade Markov $\{\theta^{(t)}\}$ cuja distribuição limite é exatamente a densidade que temos interesse. Esta ideia de utilizar o comportamento assintótico de uma cadeia de Markov surgiu mais ou menos na mesma época do primeiro algoritmo de Monte Carlo, porém não havia recursos computacionais suficientes na época que permitissem sua disseminação (Robert, 2007).

Os métodos de MCMC mais conhecidos são o *amostrador de Gibbs* (do inglês *Gibbs Sampler*) e o algoritmo de *Metropolis-Hastings* (MH), sendo que este último se relaciona com amostragem por importância, como veremos no capítulo 5. Ambos métodos podem ser combinados, dando origem ao *Metropolis within Gibbs*, para o caso onde não é possível implementar um amostrador de Gibbs sozinho. Uma vez que são métodos de MCMC, eles são baseados na geração de amostras da posteriori que não são independentes entre si e serão necessários métodos para avaliar a convergência das amostras para uma determinada distribuição de interesse (Bauwens et al., 2003).

4.4.1 Processos estocásticos

Nossa intenção é aproximar uma esperança por uma média amostral:

$$\mathbb{E}[\theta|y] = \int \theta f(\theta|y) d\theta \approx \frac{1}{S} \sum_{i=1}^{S} \theta_i$$

Nem sempre o método de integração por Monte Carlo é viável para nossas aplicações, pois amostrar diretamente de f pode não fornecer valores suficientes no intervalo de interesse. Uma alternativa que surge é a amostragem por importância, que envolve manipular a esperança para então aproximá-la por uma média amostral ponderada:

$$\int \theta \frac{f(\theta|y)}{m(\theta)} m(\theta) d\theta \approx \frac{1}{S} \sum_{i=1}^{S} \omega_i \theta_i \qquad \omega_i \equiv \frac{f(\theta|y)}{\theta}$$

Porém mesmo este método nem sempre pode ser utilizado de forma eficiente. Agora iremos fazer uma introdução às cadeias de Markov, que um processo estocástico e formam a base dos métodos MCMC.

Definição 4.4.1. Processo estocástico

Um processo estocástico em $\mathbb{R}^k$ é uma sequência de vetores aleatórios $\{X_t\}_{t\geq 0}$ definida em um *espaço de probabilidade* $(S, \mathcal{B}, \mathbb{P})$ comum.

Relembrando: O espaço de probabilidade é a tripla formada pelo espaço amostral, S, pelo espaço de eventos $\mathcal{B}$ e pela medida de probabilidade $\mathbb{P}$ e contém toda a informação necessária para associar probabilidades aos eventos do experimento. Se necessário, revise o conceito na Seção (1.2.6).

Note que quando definimos nosso vetor aleatório estamos impondo uma estrutura, a partir do espaço de probabilidade. O fato dos vetores aleatórios da família $\{X_t\}_{t\geq 0}$ estarem no mesmo espaço de probabilidade implica que probabilidades do tipo $\mathbb{P}(X_t \geq 0 \ \forall \ t)$ ou $\mathbb{P}(X_t \leq -10)$ são bem definidas. Embora a estrutura do espaço de probabilidades nos permita calcular probabilidades de interesse, ainda assim ela tem limitações que não nos permitem fazer qualquer tipo de inferência. Portanto, iremos precisar colocar um pouco mais de estrutura no nosso processo.

Um processo estocástico útil é uma sequência i.i.d., pois para este processo a lei dos grandes números nos garante que as esperanças finitas podem ser aproximadas por médias amostrais, enquanto o teorema do limite central nos permite quantificar quão boa é essa aproximação. E justamente esses dois resultados que permitem utilizarmos a Integração de Monte Carlo. No entanto, a hipótese de que as variáveis aleatórias são independentes e identicamente distribuídas não representa bem o que observamos na maioria das aplicações do mundo real. Logo, para casos mais gerais, essas propriedades assintóticas úteis podem não mais funcionar, devido a ausência da característica i.i.d. nas observações e, sem restrições adicionais ao processo, não podemos falar muita coisaa respeito de estimação. Portanto, a primeira coisa a fazer é descrever propriedades que fazem com que o nosso processo estocástico se pareça minimamente com um processo i.i.d..

4.4.2 Estacionariedade e Ergodicidade

Definição 4.4.2. Estacionariedade Um processo estocástico $\{X_t\}_{t\geq 0}$ é chamado de *estacionário* se a distribuição F de qualquer subconjunto dos vetores aleatórios não é afetada por mudanças em t, ou seja,

$$F(X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+k}) = F(X_{t+m}, X_{t+m+1}, X_{t+m+2}, \ldots, X_{t+m+k}) \ \forall m, t \ \in \mathbb{N} \tag{171}$$

Claramente, qualquer processo estacionário é então identicamente distribuído, preservando a parte "i.d." dos processos i.i.d. (e um processo i.i.d., por sua vez, é sempre estacionário).

Um passeio aleatório $X_t = X_{t-1} + u_t$ com $X_0 = 0$ pode ser escrito como $X_t = \sum_{j=1}^{t} u_j$, sendo $\{u_j\}_{j=1}^{t}$ uma sequência i.i.d. (isso significa que estamos usando uma sequência i.i.d. para construir uma sequência dependente). Se $Var[u_t] = \sigma^2$, para $0 < \sigma^2 < \infty$, então $Var[X_t] = \sum_{j=1}^{t} Var[u_j] = t \cdot \sigma^2$, fazendo com que a variância de X_t, e, por consequência, $F(X_t)$, dependa de t, de forma que $F(X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+k}) \neq F(X_{t+m}, X_{t+m+1}, X_{t+m+2}, \ldots, X_{t+m+k})$, o que torna o processo não estacionário[35]. Já o processo estocástico $\{X_t\}_{t\geq 0}$ para o qual $X_1 \sim F$ e $X_{t+1} = X_t \ \forall \ t \geq 1$ é estacionário, pois todos os valores de X_t são iguais e vieram de F. Entretanto, esse processo é extremamente dependente. Note que $\bar{X}_T = {}^1/_T \sum X_t = X_1$, que é um vetor aleatório. Ou seja, quando $T \to \infty$, $\bar{X}_T$ não converge para a média de F ou qualquer outra constante, pois cada vez que observarmos uma realização de X_1, esse valor muda. Então, estacionariedade não garante que a lei dos grandes números

[35] Observe que é necessária a restrição da variância ser não nula, já que estamos lidando com variáveis aleatórias. O exemplo é a variável aleatória constante, que assume um único valor com probabilidade 1 é independente de qualquer outra variável que seja definida da mesma forma. Se definirmos então duas (ou mais) v.a. que sejam iguais a c com probabilidade 1, elas claramente serão identicamente distribuídas e independentes com variância igual a zero.

seja válida (de fato o que faz a lei dos grandes números deixar de funcionar é essa dependência extrema, que faz com que novas observações não impliquem em aumento de informação). Porém, processos dependentes podem também satisfazer[36] a LGN, quando forem *ergódicos*.

Definição 4.4.3. Ergodicidade
Um processo estocástico é ergódico se

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \xrightarrow{P} \mu_h \equiv \mathbb{E}[h(X_t)]$$

quando $T \to \infty$ para $\mathbb{E}[h(X_t)] < \infty$.

Em resumo: estacionariedade implica que as distribuições são iguais e ergodicidade é necessária para que a lei dos grandes números funcione.

4.4.3 Sequências estocásticas recursivas

Apesar de teoricamente úteis, processos i.i.d. são muito simples para modelar observações do mundo real como PIB (é uma série com tendência) ou crescimento econômico (uma série sem tendência). Porém, estas séries podem ser relativamente bem aproximadas por um processo AR(1) gaussiano[37] dado por:

$$X_{t+1} = a + b \cdot X_t + d \cdot u_{t+1}, \text{ com } \{u_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \tag{172}$$

e X_0 dado. Os valores a, b, d são chamados de *parâmetros* e X_t são chamados de *variáveis de estado*. O contradomínio de todas as variáveis de estado é chamado de espaço de estados do processo. Se o espaço de estados for discreto, então o processo estocástico é discreto, se for contínuo, diz-se que o processo estocástico é contínuo. A equação (**??**) é uma equação em diferença estocástica e $\{X_t\}_{t\geq 0}$ é uma sequência estocástica recursiva ou processo estocástico. Uma realização desse processo é uma série de tempo.

A dinâmica de $\{X_t\}$ em (172) depende dos parâmetros e com variações deles podemos modelar uma ampla gama de séries de tempo. Mesmo nos casos onde (172) não permita capturar alguns aspectos importantes da série de tempo em mãos, é possível generalizar o processo de várias formas. Por exemplo, podemos eliminar a hipótese de normalidade, aumentar as defasagens, etc. Especificamente, se temos um processo AR(p), podemos transformar ele em um VAR(1) com p variáveis, isto é, podemos transformar um modelo com várias defasagens em um com uma defasagem, sempre. Por exemplo, o AR(2)

$$X_{t+1} = a + b \cdot X_t + \gamma X_{t-1} + u_{t+1}, \text{ com } \{u_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

pode ser escrito usando $Y_t = X_{t-1}$ com o seguinte VAR(1):

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} = b\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} a & \gamma \\ 1 & 0 \end{bmatrix}\begin{bmatrix} X_t \\ Y_t \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_{t+1}$$

Isso vai nos motivar a nos manter estudando processos com uma defasagem.

[36] Na verdade é uma *versão* da LGN, adaptada. Para mais detalhes, consulte Meyn and Tweedie (2012).
[37] Note que não estamos requerendo que ele seja um processo estacionário.

4.4.4 Processo de Markov

Processos de Markov são importantes porque são gerais o bastante para incluir uma ampla gama de processos estocásticos, porém com estrutura e restrições suficientes para permitir o desenvolvimento de resultados do nosso interesse.

Definição 4.4.4. Processo de Markov
Diz-se que o processo estocástico é Markoviano de 1^a ordem se:

$$\mathbb{P}(X_{t+1} \in A | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t)$$

Note que isso implica que toda a informação contida nos estados antes de $t-1$ é irrelevante se tivermos X_{t-1} para calcular a probabilidade de X_t. A função $Q(x_t, A) = \mathbb{P}(X_{t+1} \in A | X_t = x_t)$ é chamada de *função de probabilidade de transição*[38] e é fundamental para o estudo dos processos de Markov. A distribuição de X_1 é chamada de distribuição inicial do processo. Se a distribuição inicial e $Q(\cdot)$ forem conhecidas, o processo está completamente espcificado e todas suas distribuições podem ser calculadas.

Um processo Markoviano de 1^a ordem geral assume a seguinte forma:

$$x_{t+1} = G(x_t, u_{t+1}) \tag{173}$$

com $F(X_0) = F_0$ e $\{u_t\}_{t \geq 1}$ i.i.d. com distribuição Ψ e G é uma função mensurável que mapeia X_t de seu espaço de estados E, e o choque u_{t+1} em um novo estado $X_{t+1} \in E$. A condição inicial X_0 e os $\{u_t\}$ são independentes. Note que se o processo tem a carinha dada em (173), então:

$$\begin{aligned} Q(x_t, A) &= \mathbb{P}\{\overbrace{G(x_t, u_{t+1})}^{X_{t+1}(x_t, u_{t+1})} \in A\}; \qquad x_t \in E,\ A \in \mathcal{B} \\ &= \Psi\{u_{t+1} : G(x_t, u_{t+1}) \in A\}. \end{aligned} \tag{174}$$

Note que usamos o fato de que a única coisa aleatória na primeira linha é o u_{t+1}. Logo, precisamos apenas conhecer a distribuição do u_t, isto é, a expressão (174) deixa claro que Q é definida por Ψ e G.

Exemplo 4.4.5. AR(1) gaussiano.
Suponha $u_t \sim \mathcal{N}(0, 1)$. Neste caso,

$$\begin{aligned} Q(x_t, A) &= \mathbb{P}(a + b \cdot x_t + d \cdot u_{t+1} \in A) \\ &= \Phi(u_{t+1} : a + b \cdot x_t + d \cdot u_{t+1} \in A) \end{aligned}$$

[38] Ela significa "*Dado que estamos agora em x_t, qual a probabilidade de cairmos na região A?*".

Perceba que

$$
\begin{aligned}
x_1 &= G(x_0, u_1) \\
x_2 &= G(x_1, u_2) = G(\overbrace{G(x_0, u_1)}^{x_1}, u_2) \\
&\vdots \\
x_{t+1} &= G(x_t, u_{t+1}) = G(G(G(G(\ldots(G(G(G(x_0, u_1)u_2)u_3)\ldots)u_{t-2})u_{t-1})u_t)u_{t+1}) \\
&\Rightarrow X_{t+1} = H_t(x_0, u_1, \ldots, u_{t+1})
\end{aligned}
$$

De forma que agora conseguimos escrever x_{t+1} como função do estado inicial x_0 e de todos os choques sofridos até o momento.

No caso do processo AR(1), $H(\cdot)$ tem fórmula fechada e define a representação média móvel de X_t:

$$x_t = a \cdot \sum_{k=0}^{t-1} b^k + \sum_{k=0}^{t-1} b^k \cdot d \cdot u_{t-k} + b^t x_0 \tag{175}$$

Dizer que $\{x_t\}_{t\geq 0}$ pertencem ao mesmo espaço de probabilidade $(S, \mathcal{B}, \mathbb{P})$ significa que um elemento ω de S é selecionado pela "natureza" em $t = 0$, determinando a condição inicial x_0 e também os choques $u_t(\omega)$. Com isso, todos os $\{x_t\}$ serão determinados por:

$$x_{t+1}(\omega) = H_t(x_0(\omega), u_1(\omega), u_2(\omega), \ldots, u_{t+1}(\omega))$$

Definição 4.4.6. Densidade de transição
Se $Q(x_t, \cdot)$ for absolutamente contínua para todo x_t, é possível definir a *densidade de transição* ou *núcleo de transição*:

$$q(x_t, \cdot) \equiv \text{densidade de } \; Q(x_t, \cdot) \; \forall \; x_t \in E$$

No caso do modelo AR(1) gaussiano, a densidade de transição que dá a probabilidade de sair hoje de x_t e ir para um estado x' seria[39]:

$$q(x_t, x') = \frac{1}{\sqrt{2\pi}d} \exp\left\{-\frac{1}{2}\frac{(x' - a - b \cdot x_t)^2}{d^2}\right\}$$

Definição 4.4.7. Distribuições marginal
A distribuição marginal de X_t é simplesmente a distribuição incondicional $F(X_t)$.

No caso do AR(1) gaussiano, suponha que $F(X_0) = \mathcal{N}(\mu_0, \sigma_0^2)$. Usando a representação MA dada em (**??**), temos:

$$F(X_t) = \mathcal{N}(\mu_t, \sigma_t^2), \quad \text{com} \quad \mu_t = a \sum_{k=0}^{t-1} b^k + b^t \mu_0 \quad \text{e} \quad \sigma_t^2 = \sum_{k=0}^{t-1} b^{2k} \cdot d^2 + b^{2t} \cdot \sigma_0^2$$

[39] Estamos usando x' ao invés de x_{t+1} pois este último é o estado que de fato ocorre em $t + 1$, enquanto que x' é um estado possível.

ou, da forma recursiva,

$$\mu_{t+1} = a + b\mu_t \qquad \text{e} \qquad \sigma^2_{t+1} = b^2\sigma^2_t + d^2 \tag{176}$$

Tire a esperança dos dois lados de (172) para conferir o resultado.

Para processos mais gerais é mais difícil acompanhar a evolução dos momentos e ainda mais difícil acompanhar a evolução das densidades. Porém, o seguinte resultado nos ajuda:

$$F(X_{t+1} \in B) = \int Q(x_t, B)F(dx_t) \qquad (B \in \mathcal{B}, t \geq 0)$$

ou seja, para calcular a probabilidade de estar em B no instante $t+1$, somamos a probabilidade desse evento acontecer dado que estamos em x_t, ponderado pela chance de x_t ocorrer hoje (uma vez que não sabemos exatamente onde estamos hoje). Note que ao ponderarmos pela chance de todos os x_t e integrar seu efeito fora, conseguimos uma marginal que não depende mais do período passado. Para o caso da densidade, temos:

$$f(x_{t+1}) = \int q(x_t, x_{t+1})f(x_t)dx_t$$

sendo q a densidade de transição correspondente a Q e f a densidade correspondente a F. Com isso podemos decompor a densidade conjunta.

Seja $\{X_t\}$ um processo de Markov com densidade de transição q e densidade inicial $f(X_0)$, então a função densidade de probabilidade conjunta de $X_0, \ldots, X_T$ é

$$f_T(X_0, \ldots, X_T) = f_0(X_0)\prod_{t=0}^{T-1} p(X_{t+1}|X_t, \ldots, X_0) = f_0(X_0)f(X_{t+1}|X_t) = f_0(X_0)\prod_{t=0}^{T-1} q(X_t, X_{t+1})$$

4.4.5 Estacionariedade de processos de Markov

De acordo com a noção de estacionaridade já introduzidaem (4.4.2), uma condição necessária para a estacionaridade de uma cadeia de Markov é que as f.d.p. marginais sejam constantes. Portanto, nem todo processo de Markov é estacionário.

Exemplo 4.4.8. Convergência de um processo AR(1)
Suponha um AR(1) gaussiano com $F(X_0) = \mathcal{N}(0, 1)$ mas com a mesma estrutura de (**??**):

$$X_{t+1} = a + b \cdot X_t + d \cdot u_{t+1} \qquad \text{com} \qquad \{u_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{e} \quad |b| < 1$$

Uma simulação desse processo deixa claro que a sequência de marginais $\{F_t\}_{t=0}^T$ não será constante, como exemplificado na Figura (25). Logo, como o processo tem marginais diferentes, ele não é estacionário e isso ocorre sempre que começarmos fora da distribuição estacionária[40]. Porém, como $|b| < 1$, a diferença entre as distribuições sucessivas diminui com o tempo. Analiticamente, tomando o limite de (176) quando $t \to \infty$, teremos:

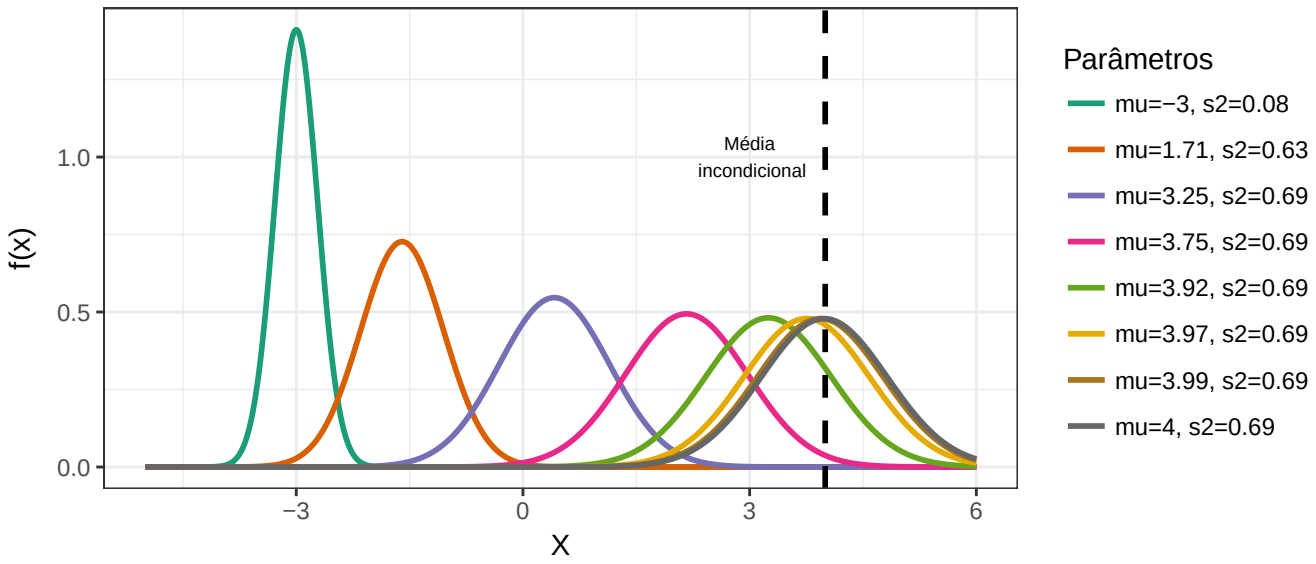
[40] Não é à toa que na função de verossimilhança de um processo AR, quando trabalhamos com séries temporais, usamos os momentos incondicionais para o valor "inicial". Isso nos garante que não vamos ter esse comportamento de marginais diferentes.

$$\mu_\infty = \frac{a}{1-b} \qquad \sigma^2_\infty = \frac{d^2}{1-b^2}$$

Denotando a distribuição associada com esses limites por F_∞, temos:

$$F_\infty = \mathcal{N}(\mu_\infty, \sigma^2_\infty) = \mathcal{N}\left(\frac{a}{1-b}, \frac{d^2}{1-b^2}\right)$$

Intuitivamente, percebe-se que a sequência de marginais ficará contida a partir de um certo momento, se estabilizando em F_∞ (demonstre!). Essa distribuição limite é o que chamamos de *distribuição estacionária* (também chamada de distribuição invariante ou incondicional).



O gráfico contém a evolução do processo $X_{t+1} = 0.8 + 0.8 \cdot X_t + 0.5 \cdot u_{t+1}$ com $\{u_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, valores iniciais $X_0 = 0$, $\mu_0 = -3$ e $\sigma^2_0 = 0.08$. Foram realizadas 50 iterações. No gráfico, estão exibidas as densidades obtidas a cada $1 + (i * (i-1))/2$ iterações, em que $i \in \{1, \ldots, 10\}$. É possível observar que os valores de μ_t e σ^2_t se aproximam rapidamente dos valores dos parâmetros da distribuição estacionária, $\mu_\infty = 4$ (indicado pela linha pontilhada preta) e $\sigma^2_\infty \approx 0.69$.

Figura 25: Evolução de um processo AR(1)

Definição 4.4.9. Distribuição estacionária

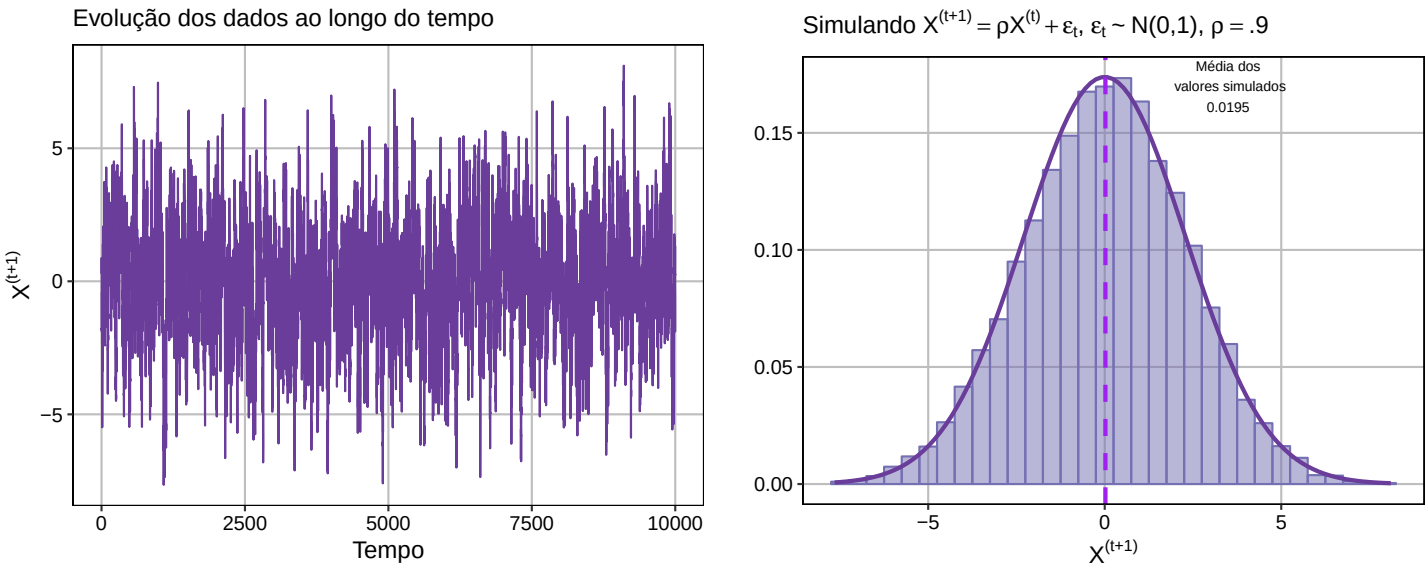
Uma distribuição estacionária, F_∞ é qualquer distribuição que satisfaz

$$F_\infty(X_{t+1} \in B) = \int Q(x_t, B) F_\infty(dX_t)$$
$$f_\infty(X_{t+1}) = \int q(x_t, x_{t+1}) f_\infty(X_t) dx_t$$

Comparando com a recursão $F(x_{t+1} \in B) = \int Q(x_t, B) F(dX_t)$, percebe-se que a definição da distribuição estacionária implica justamente que se $F(X_t) = F\infty$, então $F(X_{t+1}) = F\infty$. Logo, qualquer processo começando em F_∞ será igualmente distribuído. Na verdade, ele será estacionário, de acordo com a nossa definição.

Exemplo 4.4.10. (Adaptado de Robert and Casella (2010a))

Considere a cadeia de Markov definida por $X^{(t+1)} = \rho X^{(t)} + \epsilon_t$, com $\epsilon_t \sim \mathcal{N}(0,1)$. Simule $X^{(0)} \sim \mathcal{N}(0,1)$ e plote o histograma de uma amostra de $X^{(t)}$ para $t \leq 10^4$ e $\rho = 0.9$. Verifique a aderência da distribuição estacionária $\mathcal{N}\left(0, \frac{1}{(1-\rho^2)}\right)$.



O lado esquerdo da figura contém a evolução de $X^{(t)}$ ao longo do tempo, enquanto que o valo direito apresenta o histograma das frequências relativas dos valores junto com a densidade $\mathcal{N}\left(0, \frac{1}{(1-\rho^2)}\right)$ (linha contínua). A linha pontilhada na parte central do histograma representa a média dos valores amostrados (0.0195).

Figura 26: Evolução de $X^{(t+1)} = \rho X^{(t)} + \epsilon_t$, com $\epsilon_t \sim \mathcal{N}(0, 1)$ e $t \in \{1, \ldots, 10000\}$.

Os valores gerados da cadeia estão na Figura (25), sendo que no histograma (lado direito) foi plotada a distribuição $\mathcal{N}\left(0, \frac{1}{1-\rho^2}\right)$, por ser a distribuição estacionária da cadeia. Para verificar[41], vamos calcular a média e a variância, assumindo $X^{(t+1)} = X^{(t)}$:

$$\begin{aligned}
\mathbb{E}[X^{(t+1)}] &= \mathbb{E}[\rho X^{(t)} + \epsilon_t] \\
\mathbb{E}[X] &= \mathbb{E}[\rho X + \epsilon_t] \\
\mathbb{E}[X] &= \rho\mathbb{E}[X] + \mathbb{E}[\epsilon_t] \\
(1-\rho)\mathbb{E}[X] &= 0 \\
\mathbb{E}[X] &= 0
\end{aligned}$$

Se a média é 0, temos $Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (0)^2 = \mathbb{E}[X^2]$. Então:

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[(\rho X + \epsilon_t)^2] \\
\mathbb{E}[X^2] &= \mathbb{E}[(\rho X)^2 + 2 \cdot \rho X \epsilon_t + \epsilon_t^2] \\
\mathbb{E}[X^2] &= \mathbb{E}(\rho X)^2 + 2\mathbb{E}(\rho X)\mathbb{E}(\epsilon_t) + \mathbb{E}(\epsilon_t^2) \\
\mathbb{E}[X^2] &= \rho^2 \cdot \mathbb{E}[X^2] + 0 + 1 \\
(1-\rho^2)\mathbb{E}[X^2] &= 1 \\
\mathbb{E}[X^2] &= \frac{1}{(1-\rho^2)}
\end{aligned}$$

[41] Uma pessoa mais atenta pode questionar que o fato da média e a variância serem essas que foram calculadas, isso não garante que a distribuição é de fato estacionária. O cálculo da distribuição invariante pode ser visto em Greenberg (2008), página 86.

4.4.6 Componentes assintóticos de Processos de Markov

Quando um processo de Markov exibe propriedades como ergodicidade? Para ganhar intuição, considere o processo $X_{t+1} = aX_t + u_{t+1}$, com $\{u_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ e $|a| < 1$, estacionário e com f.d.p. marginais para X_t dadas por $F_\infty = \mathcal{N}(0, 1/1-a^2)$. A lei dos grandes números para esse processo é dada por:

$$\bar{X}_T \xrightarrow{P} \int X_t F_\infty(dX_t) = 0 \text{ quando } T \to \infty \tag{177}$$

Sabemos do básico de teoria assintótica que para (177) ocorrer, é suficiente que

$$\mathbb{E}[\bar{X}_T] \to 0 \qquad \text{e} \qquad Var[\bar{X}_T] \to 0 \quad \text{quando } T \to \infty$$

A primeira parte é óbvia, pois $\mathbb{E}[\bar{X}_T] = \frac{1}{T}\sum \mathbb{E}[X_t] = 0$. Para a variância, como a media é zero, teremos:

$$Var\left[\frac{1}{T}\sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right)^2\right] = \frac{1}{T^2}\sum_{n=1}^{T}\sum_{m=1}^{T}\mathbb{E}[X_n \cdot X_m]$$

Usando o fato que $\mathbb{E}[X_n \cdot X_m] = \mathbb{E}[X_m \cdot X_n]$, podemos encontrar um limite superior:

$$Var[\bar{X}_T] \leq \frac{2}{T^2}\sum_{1\leq n\leq m\leq T}\mathbb{E}[X_n \cdot X_m] = \frac{2}{T^2}\sum_{j=0}^{T-1}\sum_{t=1}^{T-j}\mathbb{E}[X_{t+j} \cdot X_t].$$

Logo, para que $\bar{X}_T \to 0$ quando $T \to \infty$, os termos da covariância $\mathbb{E}[X_{t+j} \cdot X_t]$ precisam ir para 0 suficientemente rápido.

Resumindo, se as correlações desaparecem de maneira rápida o suficiente, então podemos usar a LGN mesmo em processos dependentes e que não são identicamente distribuídos[42]! As condições para que essas correlações desapareçam não serão detalhadas aqui mas podem ser encontradas em Meyn and Tweedie (2012) e Robert and Casella (2010b).

4.4.7 Cadeias de Markov via Monte Carlo (MCMC)

MCMC é uma forma de obter amostras simuladas de alguma distribuição f sem utilizar diretamente esta distribuição. O princípio básico é usar uma cadeia de Markov ergódica com distribuição estacionária igual a f. Para um valor inicial arbitrário x_0, uma cadeia $\{X_t\}_{t=1}^{T_0}$ é gerada usando um núcleo de densidade de transição cuja densidade estacionária é f. Portanto, para T_0 suficientemente grande, X_{T_0} pode ser considerado uma realização de f.

Em comparação com métodos de Monte Carlo tradicionais como amostragem por importância, MCMC são subótimos, pois dependem de resultados assintóticos não só para a aproximação da integral de interesse, mas também para a aproximação da função densidade de probabilidade estacionária, f. Entretanto, esse tipo de algoritmo pode ser utilizado em situações muito mais gerais, onde métodos de integração de Monte Carlo não podem ser usados.

Na seção de integração de Monte Carlo apresentamos o problema de aproximar a integral

$$\int h(x)f(x)dx.$$

A nossa estratégia até então era enxergar a integral como uma esperança populacional para aproximá-la por uma média

[42] Na verdade o que foi feito aqui é apenas **um** caso onde isso funciona, mas irá funcionar razoavelmente bem nos demais, atendidas as condições de estacionariedade e ergodicidade.

amostral. Nos casos onde o suporte comum de $h(x)$ e $f(x)$ é pequeno, podemos escolher um amostrador por importância que vai atuar fazendo o papel de "pesos" para que as amostras geradas estejam dentro do intervalode interesse com uma probabilidade um pouco maior (método da amostragem por importância). Mesmo assim, nem sempre iremos conseguir estimativas boas, seja pela falta de um suporte comum seja pela dificuldade em escolher um bom amostrador por importância.

O que iremos discutir agora é como aproximar a integral fazendo a escolha de um núcleo de transição cuja distribuição estacionária é a distribuição alvo que temos interesse. Vimos nas seções anteriores as condições para as quais é possível que uma cadeia de Markov convirja para sua densidade estacionária, isto é, para um valor inicial x_0, uma cadeia $\{X_t\}$ é gerada utilizando um núcleo de transição K (ou P) com distribuição estacionária f (ou π), e sob determinadas condições, a convergência de $\{X_t\}$ para uma v.a. que foi amostrada diretamente de f está garantida pelos resultados anteriores.

Definição 4.4.11. Método de Cadeias de Markov via Monte Carlo
Um método de MCMC para a simulação da distribuição f é qualquer método que produza uma cadeia de Markov ergódica cuja distribuição estacionária é f.

Vimos que a característica marcante das cadeias de Markov é que seu estado atual depende exclusivamente do estado que a cadeia encontra-se no momento anterior, isto é, para uma cadeia $\{\theta^{(t)}\}_{t\geq 0}$, a distribuição de $\theta^{(t+1)}$ dado o valor atual e os valores passados da cadeia, depende apenas de $\theta^{(t)}$. Essa probabilidade é representada pelo núcleo de transição

$$q(\theta^{(t)}, A) = \mathbb{P}\left(\theta^{(t+1)} \in A | \theta^{(t)}\right)$$

em que A é um conjunto mensurável da σ-álgebra onde a posteriori $\pi(\theta|y)$ está definida. O que queremos fazer então é encontrar um Kernel $q(\theta, A)$ tal que a probabilidade de que a cadeia esteja no conjunto A dada uma condição inicial $\theta^(0)$ convirja para a densidade algo que temos interesse. Mais especificamente, queremos que

$$q^{(t)}(\theta^{(0)}, A) = \mathbb{P}\left(\theta^{(t)} \in A | \theta^{(0)}\right)$$

convirja, quando t for suficientemente grande, para a probabilidade que a distribuição alvo atribui a A ocorrer. Uma vez que o núcleo $q(\cdot, \cdot)$ está definido, pode-se obter retiradas da seguinte forma:

$$\begin{aligned}
\theta^{(1)} &\sim q(\theta^{(0)}, \cdot) \\
\theta^{(2)} &\sim q(\theta^{(1)}, \cdot) \\
&\vdots \\
\theta^{(t+1)} &\sim q(\theta^{(t)}, \cdot) \\
&\vdots
\end{aligned}$$

Como a cadeia foi construída de forma a ter distribuição invariante igual a $\pi(\theta|y)$, os valores a partir de um certo n_0, $\{\theta^{(n_0+t)}\}_{t=1}^{T}$ poderão ser considerados valores amostrados a partir da posteriori (para n_0 suficientemente grande). Já as observações iniciais, $\{\theta^{(t)}\}_{t=0}^{T}$ são conhecidas como *burn in* e descartadas. Se de fato n_0 é grande e se T também for grande, podemos construir uma sequência de valores simulados e combinar com o método de Monte Carlo para poder calcular momentos da distribuição a posteriori.

Apesar de serem métodos mais flexíveis que os métodos de integração de Monte Carlo, os métodos de MCMC requerem cautela na hora de sua implementação. Chib (2013) chama atenção para o fato de que não existe um único Kernel dada

uma distribuição alvo, sendo que, dentre os núcleos possíveis de serem usados, alguns produzirão amostras da densidade estacionária de maneira mais (ou menos) eficiente, em particular quando as amostras são altamente correlacionadas (fazendo com que não seja explorado todo o suporte da densidade que temos interesse).

4.4.8 O amostrador de Gibbs (*Gibbs sampler*)

No caso do MNRL nós gostaríamos de uma cadeia de Markov em que $f_\infty = f(\theta, y)$, para $\theta = (\beta, h)$. No entanto, só dispomos das distribuições condicionais $f(\beta|h, y)$ e $f(h|\beta, y)$, dadas em (141) e (144), respectivamente. Sabemos que se a densidade $f(\theta, y)$ é uma densidade estacionária, então ela irá satisfazer

$$f_A(\theta_{t+1}) = \int q(\theta_t, \theta_{t+1}) f_A(\theta_t) d\theta_t \tag{178}$$

em que foi omitida a condicional a y para simplificar a notação. Além disso, adotou-se o subscrito A em $f_A(\cdot)$ para denotar que é a nossa *densidade alvo*. Podemos pensar em um núcleo da seguinte forma:

$$q(\theta_t, \theta_{t+1}) = f(h_{t+1}|\beta_t) f(\beta_{t+1}|h_{t+1}). \tag{179}$$

Perceba que este núcleo de fato tem como distribuição estacionária a densidade conjunta a posteriori.

Demonstração. Utilizando (179) em (178), teremos:

$$\begin{aligned}
&\int\int f(h_{t+1}|\beta_t) \underbrace{f(\beta_{t+1}|h_{t+1})}_{\text{não depende de } \beta \text{ e } h} \underbrace{f(\beta_t, h_t)}_{\text{conjunta}} d\beta_t dh_t = \\
&= f(\beta_{t+1}|h_{t+1}) \int f(h_{t+1}|\beta_t) \underbrace{\int f(\beta_t, h_t) dh_t}_{\substack{f(\beta_t) \\ \text{marginal}}} d\beta_t \\
&= f(\beta_{t+1}|h_{t+1}) \underbrace{\int f(h_{t+1}|\beta_t) f(\beta_t) d\beta_t}_{(\bigstar)} \\
&= f(\beta_{t+1}|h_{t+1}) \cdot f(h_{t+1}) \qquad (\heartsuit) \\
&= f(\beta_{t+1}, h_{t+1}) \equiv f(\theta|y)
\end{aligned}$$

Alguns comentários sobre a demonstração:

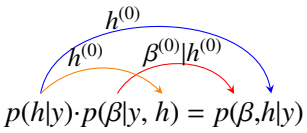
1. Perceba que o termo dentro da integral em ($\bigstar$) é similar ao que temos em (178), pois temos uma densidade condicional (que em (178) é o núcleo de transição $q(\cdot, \cdot)$) vezes uma densidade de uma variável aleatória (ou ponderada por essa densidade) e integramos em todo o domínio da v.a. O que sobra após esse procedimento é apenas a probabilidade marginalizada. Neste caso, estamos avaliando qual a probabilidade de estar em h_{t+1} condicional a estar em qualquer um dos β_t. Ao integrarmos fora o efeito do β_t ficamos simplesmente com a probabilidade de ocorrer h_{t+1}.
2. A respeito da equação ($\heartsuit$), note que ela pode ser vista como a densidade conjunta, pois da regra da probabilidade condicional temos $\mathbb{P}(A|B) = {}^{\mathbb{P}(A\cap B)}/_{\mathbb{P}(B)} \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B)$.

□

O amostrador de Gibbs é um instrumento poderoso para a simulação da f.d.p. a posteriori através de partições do vetor

de parâmetros θ, que pode ser particionado em vários blocos: $\theta = (\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(B)})$, sendo que $\theta_{(j)}$ para $j = 1, \ldots, B$ pode ser tanto um vetor como um escalar. No MNRL é conveniente fazer $B = 2$ e $\theta = (\beta', h)'$. Como vimos, se a priori usada for Normal-Gama independente, não é possível amostrar diretamente de $p(\theta|y)$, mas conseguimos amostrar de $p(\theta_{(1)}|\theta_{(2)}, Y) = p(\beta|h, Y)$ e de $p(\theta_{(2)}|\theta_{(1)}, Y) = p(h|\beta, Y$. No caso geral de B blocos, as distribuições condicionais $p(\theta_{(1)}|Y, \theta_{(2)}, \theta_{(3)}, \ldots, \theta_{(B)})$, $p(\theta_{(2)}|Y, \theta_{(1)}, \theta_{(3)}, \ldots, \theta_{(B)})$, ..., $p(\theta_{(B)}|Y, \theta_{(2)}, \theta_{(3)}, \ldots, \theta_{(B-1)})$ são chamadas de *distribuições posteriores condicionais totais*, pois elas definem a posteriori condicional para um bloco dado **todos** os outros blocos[43].

No caso do MNRL, temos $p(\beta|Y, h) = \mathcal{N}(\bar{\beta}, \bar{V})$ e $p(h|Y, \beta) = \mathcal{G}(\bar{s}^{-2}, \bar{\nu})$ e é fácil amostrar delas. Vimos também que se amostrarmos das posterioris condicionais totais em sequência (ver equações (179) e (178)), após a convergência da cadeira de Markov, teremos uma amostra $\{\theta^{(i)}\}_{i=1}^{S}$ que pode ser usada para aproximar $\mathbb{E}[g(\theta)|y]$ usando integração de Monte Carlo. Mais especificamente, o amostrador de Gibbs no MNRL com priori Normal-Gama independente funciona da seguinte forma. Podemos escrever $\theta = \{\beta, h\}$, ou seja, $B = 2$ e vamos supôr que temos uma observação da f.d.p. posterior **marginal** de h, $p(h|Y)$, denotada por $h^{(0)}$. Como $p(\beta, h|y) = p(\beta|Y, h)p(h|Y)$, segue que uma realização de β, chamada de $\beta^{(0)}$, de $p\left(\beta|Y, h^{(0)}\right)$ é uma realização de $p(\beta, h|Y)$, pois $p(\beta, h|Y) = p(\beta|Y, h)p(h|Y)$ e amostrar usando o lado esquerdo ou direito dessa expressão é indiferente. Portanto, agora temos $h^{(0)}$ e $\beta^{(0)}$, de forma que podemos usar $\beta^{(0)}$ em $p(h|\beta, Y)$ para amostrar $h^{(1)}$ e essa realização será uma realização de $p(\beta, h|Y)$ pelo mesmo argumento anterior. O primeiro passo deste procedimento para o MNRL pode ser representado da seguinte forma:



No entanto, nós *não* temos um valor inicial de $p(h|y)$, pois nós não conhecemos essa marginal e portanto nossos valores iniciais $\{\beta^{(0)}, h^{(0)}\}$ **não** são pares da posteriori conjunta! Nós resolvemos isso montando uma cadeia de Markov cuja distribuição estacionária é a distribuição conjunta, na linha do que foi feito no Exemplo (4.4.8). Assim, o que nós fazemos na prática é usar um valor $h^{(0)}$ que vem de $p(h|y, \beta)$ para encontrar $\beta^{(0)}$ e assim sucessivamente. Com um número suficientemente grande B de iterações, se o processo for ergódico, as observações a partir de $B + 1$, isto é, $\{\beta^{(B+1)}, h^{(B+1)}\}$ são observações da distribuição posterior conjunta.

No caso geral de B blocos, como descrito no Algoritmo (2), o amostrador de Gibbs irá produzir uma amostra com S observações de $\theta^{(s)}$, para $s = 1, \ldots, S$. Após descartar as S_0 realizações iniciais para eliminar o efeito de $\theta^{(0)}$, as $S_1 = S - S_0$ realizações restantes podem ser usadas para estimar características da distribuição conjunta de θ. Ou seja, usando a teoria de processos estocásticos e cadeias de Markov é possível mostrar que, sob hipóteses de regularidade fracas, o valor inicial $\theta^{(0)}$ não importa e o amostrador de Gibbs irá convergir para uma sequência amostral de $p(\theta|y)$. Logo, é comum escolher θ^0 e rodar o amostrador de Gibbs S vezes. Porém, as primeiras S_0 realizações são descartadas baseando-se na ideia de que até S_0 o amostrador ainda não convergiu para $p(\theta|y)$ e as realizações $\theta^{(S_0)}$ para $S_0 = 1, 2, \ldots, S_0$ não são realizações de $p(\theta|y)$. Isto significa que, assim como a integração de Monte Carlo, uma lei fraca dos grandes números garante que

$$\hat{g}_{S_1} = \frac{1}{S_1} \sum_{s=S_0+1}^{S} g(\theta^{(s)}) \to \mathbb{E}[g(\theta)|y] \qquad \text{quando} \quad S_1 \to \infty.$$

Entretanto, diferentemente do que ocorre com a integração de Monte Carlo tradicional, com o amostrador de Gibbs:

1. Precisamos garantir que $\theta^{(0)}$ não tenha mais efeito nas realizações θ^s para $s > S_0$ (não temos como saber qual o valor

[43] Este texto tem uma crítica de Del Negro e Primiceri ao trabalho de Primiceri(2005) por ter construído uma parte do seu amostrador de Gibbs sem considerar a condicional completa a todos os outros blocos.

Algoritmo 2: Amostrador de Gibbs para o caso geral

Entrada: Um valor inicial, $\theta^{(0)}$.
Saída: $\{\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_B^{(s)}\}_{s=1}^{S}$
início
  para $s = 1, \ldots, S$ **gere**
    1. Uma realização $\theta_{(1)}^{(s)}$ de $p\left(\theta_{(1)}|y, \theta_{(2)}^{(s-1)}, \theta_{(3)}^{(s-1)}, \ldots, \theta_{(B)}^{(s-1)}\right)$;
    2. Uma realização $\theta_{(2)}^{(s)}$ de $p\left(\theta_{(2)}|y, \theta_{(1)}^{(s)}, \theta_{(3)}^{(s-1)}, \ldots, \theta_{(B)}^{(s-1)}\right)$;
    3. Uma realização $\theta_{(3)}^{(s)}$ de $p\left((\theta_{(3)}|y, \theta_{(1)}^{(s)}, \theta_{(2)}^{(s)}, \theta_{(4)}^{(s-1)}, \ldots, \theta_{(B)}^{(s-1)}\right)$;
    $\vdots$
    B. Uma realização $\theta_{(B)}^{(s)}$ de $p\left((\theta_{(B)}|y, \theta_{(1)}^{(s)}, \theta_{(2)}^{(s)}, \ldots, \theta_{(B-1)}^{(s)}\right)$
  fim
fim

de S_0 necessário);

2. A sequência de realizações $\theta^{(s)}$ para $s = 1, \ldots, S$ não é i.i.d. pois $\theta^{(s)}$ depende de $\theta^{(s-1)}$, gerando uma amostra autocorrelacionada.

Na prática, isso significa que precisamos escolher S muito maior para o amostrador de Gibbs do que para a integração de Monte Carlo tradicional. Neste sentido, iremos utilizar ferramentas de *diagnóstico de convergência* para tentar verificar se a nossa cadeia convergiu para a distribuição estacionária.

4.4.9 Diagnóstico para Cadeias de Markov Monte Carlo (MCMC)

O fato da realização S, $\theta^{(S)}$, do amostrador de Gibbs depender de $\theta^{(s-1)}$ implica que a sequência $\{\theta^{(i)}\}_{i=1}^{S}$ é uma Cadeia de Markov. Precisaremos verificar tanto o erro de aproximação (proveniente do uso do método de Monte Carlo) como verificar a convergência da Cadeia de Markov. No segundo caso, já argumentamos que, como $\theta^{(s)}$ são correlacionados, precisamos usar uma outra versão do TLC. Isto é, supondo que o amostrador de Gibbs convergiu para uma sequência de $p(\theta|y)$, temos:

$$\sqrt{S_1}\left\{\hat{g}_{S_1} = \mathbb{E}[g(\theta)|Y]\right\} \to \mathcal{N}(0, \sigma_g^2), \qquad \text{se } S_1 \to \infty \tag{180}$$

Porém, σ_g^2 não tem a mesma fórmula simples como quando $\theta^{(s)}$ eram i.i.d.'s e não há nenhuma forma de estimar σ_g^2 que seja totalmente justificada. Geweke (1992), utilizando argumentos de série de tempo, sugere o seguinte estimador:

$$\hat{\sigma}_g^2 = \frac{S(0)}{S_1} \tag{181}$$

em que $S(0)$ é a densidade espectral de $\left\{\theta^{(s)}\right\}_{s=1}^{S}$ avaliada em 0. Com isso, o desvio padrão numérico pode ser aproximado por $\hat{\sigma}_g / \sqrt{S_1}$.

Geweke (1992) propõe que para verificar a convergência da cadeia a amostra (após descarte do *burn-in*) seja dividida em três pedaços:

$$S_1 = S_A + S_B + S_C.$$

Posto de outra forma, iremos dividir $\left\{\theta^{(s)}\right\}_{s=S_0+1}^{S}$ em 3 grupos A, B e C tais que

$$\left\{\theta^{(s)}\right\}_{s=S_0+1}^{S} = \left\{\left\{\theta^{(s)}\right\}_{s=S_0+1}^{S_A}, \left\{\theta^{(s)}\right\}_{s=S_0+S_A}^{S_B}, \left\{\theta^{(s)}\right\}_{s=S_0+S_A+S_B}^{S_C}\right\}.$$

Intuitivamente, esperamos que se um número suficientemente grande de valores for amostrado, a estimativa de $g(\theta)$ baseada na primeira parte da amostra deveria ser muito similar à estimativa calculada com os últimos valores amostrados. Se essas duas estimativas forem muito diferentes, isso pode indicar tanto que o tamanho amostral é pequeno (e a diferença estaria associada apenas por ter uma amostra finita) ou que o efeito do valor inicial $\theta^{(0)}$ ainda não desapareceu e é necessário aumentar o *burn-in*. Então, o diagnóstico será comparar resultados obtidos por S_A com aqueles obtidos usando S_C. Calcula-se $\hat{g}_{S_A}$ e $\hat{g}_{S_C}$ e, usando resultados assintóticos, temos:

$$\frac{\hat{g}_{S_A} - \hat{g}_{S_B}}{\sqrt{\frac{\hat{\sigma}_A^2}{S_A} + \frac{\hat{\sigma}_C^2}{S_C}}} \xrightarrow{\text{d}} N(0,1) \tag{182}$$

A ideia é que tanto $\hat{g}_{S_A}$ como $\hat{g}_{S_C}$ deveriam convergir para uma mesma média g^* (por isso quando tomamos a diferença entre elas este termo acaba "desaparecendo" do denominador) e se padronizarmos pelos desvios padrão, iremos cair em uma v.a. com distribuição normal padrão. Na prática, escolhe-se $S_A = 0.1S_1$, $S_B = 0.5S_1$ e $S_C = 0.4S_1$. Apesar de bem informativo, este diagnóstico pode indicar convergência quando na realidade ela não ocorreu (o que significa que a escolha de S_0 não é adequada). Isso mais frequentemente acontecerá em 2 casos[44]:

1. caso a posteriori seja multimodal; e,
2. quando $\theta^{(s)}$ for escolhido muito longe da região de maior probabilidade de $p(\theta|y)$.

Uma técnica comum para evitar esse tipo de problema é iniciar mais de uma sequência simultaneamente com valores iniciais bem distintos. Caso todas as cadeias gerem resultados similares, temos mais confiança na convergência da cadeia de Markov. Gelman and Rubin (1992) sugerem comparar o desvio padrão calculado dentro de cada sequência com o desvio padrão calculado entre sequências. Intuitivamente, após a convergência, esses valores devem ser próximos. Uma estimativa da variância de uma particular sequência é:

$$S_i^2 = \frac{1}{S_1 - 1} \sum_{s=S_0+1}^{S} \left[g\left(\theta^{(s_{1,i})}\right) - \hat{g}_{S_1}^{(i)}\right]^2 \tag{183}$$

e é chamada também de variância dentro da sequência (*wihin-sequence variance*). Usando isso, a variância média das sequências, i.e., uma média das variâncias de cada sequências, será dada por:

$$W = \frac{1}{m} \sum_{i=1}^{m} S_i^2 \tag{184}$$

Gelman (1996) mostra que a variância entre sequências (distintas) pode ser estimada por:

$$B = \frac{S_1}{m-1} \sum_{i=1}^{m} \left(\hat{g}_{S_1}^{(i)} - \hat{g}\right)^2 \tag{185}$$

[44] Para ver exemplos, leia a página 66-67 de Koop (2003).

com

$$\hat{g} = \frac{1}{m}\sum_{i=1}^{m} \hat{g}_{S_1}^{(i)} \tag{186}$$

Note que W é uma estimativa para $Var[g(\theta)|Y]$. É possível mostrar que a quantidade

$$\widehat{Var[g(\theta)|Y]} = \frac{S_1 - 1}{S_1} W + \frac{1}{S_1} B \tag{187}$$

também é uma estimativa para $Var[g(\theta)|Y]$. Porém, se a cadeia ainda não convergiu, W irá subestimar $Var[g(\theta)|Y]$ - pois ao tirarmos a média, mesmo que hajam cadeias muito diferentes, essa distância vai acabar sendo suavizada. B, por outro lado, será grande, pois $\theta^{(0)}$ foram escolhidos de forma muito dispersa intencionalmente. Com isso, um diagnóstico de convergência usual é:

$$\hat{R} = \frac{\widehat{Var[g(\theta)|Y]}}{W} \tag{188}$$

$\hat{R}$ será próximo de 1 se houver convergência e maior do que 1 se não houver. A quantia $\sqrt{\hat{R}}$ é chamada de *estimated potential scale reduction* e pode ser interpretado como uma quota para o quão distante as estimativas do desvio padrão de $g(\theta)$ podem ser, dadas uma convergência ruim. Gelman (1996) sugere que um valor de $\hat{R}$ maior que 1.2 indica não convergência.

4.4.10 Previsão

Como vimos no capítulo anterior, a densidade preditiva é dada por:

$$p(Y^*|Y) = \int\int p(Y^*|Y,\beta,h)p(\beta,h|Y)d\beta dh \tag{189}$$

Diferentemente do que ocorre com a priori conjugada natural, a priori Normal Gama independente não possui solução analítica para 189 e simulação de MC precisarão ser utilizadas.

Qualquer característica de $p(Y^*|Y)$ pode ser calculada por:

$$\mathbb{E}[g(Y^*)|Y] = \int g(Y^*)p(Y^*|Y)dY^* \approx \frac{1}{S_1}\sum g\left(Y^{(s)}\right) \tag{190}$$

- Se $g(Y^*) = Y^*$, calcula-se a previsão para Y;
- $g(Y^*) = Y^{*2}$ pode ser utilizado para, juntamente com $\mathbb{E}[g(\hat{Y}^*)|Y]$, calcular a variância das previsões.

Porém, o amostrador de Gibbs só nos dá realizações de $\theta = \{\beta, h\}$ e não de Y^*, mas podemos usar realizações de $\beta^{(s)}$ em $h^{(s)}$ em $p(Y^*|Y,\beta^{(s)},h^{(s)})$ (f. de verossimilhança do modelo) para amostrar Y^* (colocamos esses valores em $Y^{(s)}$ na equação 190).

5 Parte 5 - Modelo de Regressão Linear com Matriz de Covariância Geral para os Erros

Este capítulo é baseado no capítulo 6 de Koop (2003). Considere o MNRL:

$$Y = X\beta + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, h^{-1}\Omega). \tag{191}$$

Todos os modelos deste capítulo terão as seguintes suposições a respeito de (191):

1. ε tem distribuição multivariada Normal com média 0_N e matriz de covariância $h^{-1}\Omega$, em que Ω é uma matriz $N \times N$ positiva definida;
2. Todos os elementos de X são fixos ou, no caso de serem aleatórios, eles são independentes de todos os elementos do vetor ε e possuem uma distribuição de probabilidade $p(X|\lambda)$, em que λ é um vetor de parâmetros que não inclue β nem h.

Nos capítulos anteriores utilizamos a hipóteses de que $\varepsilon \sim \mathcal{N}(0, h^{-1}\mathbb{I}_n)$ porém agora iremos generalizar os erros para o caso da matriz de variância e covariância geral, como em (191).

Este arcabouço permite modelar:

- **Erros heterocedásticos:** Os elementos da diagonal principal de Ω não são todos iguais, isto é, denotando os elementos da diagonal de Ω por ω_{ii}, $i \in \{1, 2, \ldots, n\}$, temos que $\omega ii \neq \omega_{jj}$ para pelo menos um $i \neq j$:

$$\Omega = \begin{bmatrix} \omega_{11} & 0 & \ldots & 0 \\ 0 & \omega_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \omega_{nn} \end{bmatrix}.$$

- **Erros autocorrelacionados:** Os elementos fora da diagonal principal de Ω não são todos iguais a zero:

$$\Omega = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1,n-1} & \rho_{1,n} \\ \rho_{21} & 1 & \ldots & \rho_{2,n-1} & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-1,2} & \rho_{n-1,2} & \ldots & 1 & \rho_{n-1,n} \\ \rho_{n,1} & \rho_{n,2} & \ldots & \rho_{n,n-1} & 1 \end{bmatrix} \qquad \text{para algum } \rho_{ij} \neq 0.$$

  Note que, como os elementos da diagonal principal são todos iguais, os erros são homocedásticos. Obviamente, é possível combinar os dois casos acima para montar Ω tal que os erros sejam heterocedásticos e autocorrelacionados.

- **Dados em painel:** A matriz de variâncias e covariâncias Ω será bloco-diagonal;

entre outras possibilidades, incluindo o modelo que já vimos. Note que para o MNRL dos capítulos anteriores, temos

simplesmente

$$\Omega \equiv \mathbb{I}_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Caso Ω seja conhecida, podemos transformar o modelo de forma a chegarmos aos modelos já estudados. No entanto, em geral, não conhecemos essa matriz e iremos precisar estimar seus elementos. Especificamente, na abordagem bayesiana, iremos precisar de uma priori e uma forma de fazer atualização do nosso conhecimento para chegar à posteriori. Nosso problema está em obter uma posteriori conjunta para todos os parâmetros desconhecidos, então conhecer Ω é crucial para a determinação das distribuições posteriori marginais para β e h.

5.1 A função de verossimilhança

A verossimilhança para o modelo dado em (191) será dada por:

$$P(Y|\beta, h, \Omega) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}}|\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{h}{2}(Y - X\beta)'\Omega^{-1}(Y - X\beta)\right\}. \tag{192}$$

Como Ω é positiva semidefinida, Ω^{-1} existe e podemos encontrar P tal que $\Omega^{-1} = P^T P$ (que é a decomposição de Choleski da inversa). Pré multiplicando os dois lados de 191 por P, temos:

$$Y^* = X^*\beta + \varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}(0, h^{-1}\mathbb{I}), \tag{193}$$

com $Y^* = PY$, $X^* = PX$ e $\varepsilon^* = P\varepsilon$. Nós simplesmente pré-multiplicamos o modelo pelo padrão de heterocedasticidade e conseguimos deixar o termo de erro normalizado (esse procedimento é igual ao que fazemos em mínimos quadrados generalizados - MQG ou GLS).

A matriz Ω^{-1} é necessária pois queremos "dividir" o modelo e a decomposição $P^T P$ é como se tirássemos uma raiz quadrada (que seria o desvio padrão). Observe que se $Y \sim (X\beta, h^{-1}\Omega)$, então $Var[Y] = h^{-1}\Omega$. Assim, $Var[PY] = PVar[Y]P^T = Ph^{-1}\Omega P^T$. Agora note que se $\Omega^{-1} = P^T P$, então $\Omega \equiv (\Omega^{-1})^{-1} = (P^T P)^{-1}$, de forma que $Var[PY] = h^{-1}P[P^T P]^{-1}P^T = h^{-1}PP^{-1}(P^T)^{-1}P^T = h^{-1}\mathbb{I}$. Perceba que o fato de deixarmos o h^{-1} do lado de fora de Ω implica que estamos aceitando um desconhecimento da variância do modelo (pois já vimos como estimar este escalar). Dessa forma, a verossimilhança dada em (192) pode ser escrita como:

$$P(Y^*|\beta, h, \Omega) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{h}{2}(Y^* - X^*\beta)'(Y^* - X^*\beta)\right\} \tag{194}$$

Podemos utilizar as seguintes quantidades de MQG para reescrever (195):

$$\nu = N - k \tag{195}$$

$$\hat{\beta}(\Omega) = (X^{*'}X^*)^{-1}X^{*'}Y = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \tag{196}$$

$$s^2(\Omega) = \frac{\left(Y^* - X^*\hat{\beta}(\Omega)\right)'\left(Y^* - X^*\hat{\beta}(\Omega)\right)}{\nu} = \frac{\left(Y - X\hat{\beta}(\Omega)\right)'\Omega^{-1}\left(Y - X\hat{\beta}(\Omega)\right)}{\nu} \tag{197}$$

Então a verossimilhança será:

$$p(Y|\beta, h, \Omega) = \frac{1}{(2\pi)^{\frac{N}{2}}} \underbrace{h^{\frac{k}{2}} \exp\left\{-\frac{h}{2}(\beta - \hat{\beta}(\Omega))'X'\Omega^{-1}X(\beta - \hat{\beta}(\Omega))\right\}}_{\beta|h,\Omega \sim \mathcal{N}(\hat{\beta}(\Omega), h^{-1}(X'\Omega^{-1}X)^{-1})} \cdot \underbrace{h^{\frac{\nu}{2}} \exp\left\{-\frac{h\nu}{2s^{-1}(\Omega)}\right\}}_{h|\beta,\Omega \sim \mathcal{G}(\nu, s^2(\Omega))} \tag{198}$$

Demonstração. A demonstração é recomendada como exercício. □

5.2 A densidade a priori

Assumindo independência, isto é, assumindo $p(\beta, h, \Omega) = p(\beta)p(h)p(\Omega)$ teremos:

$$p(\beta) = f_{\mathcal{N}}(\beta \mid \underline{\beta}, \underline{V}) \tag{199}$$

$$p(h) = f_{\mathcal{G}}(h \mid \underline{\nu}, \underline{s}^{-2}). \tag{200}$$

A suposição de independência a priori entre os parâmetros não é um requerimento (necessário) do modelo, porém, como vimos nos capítulos anteriores, apesar da priori conjugada natural ajudar nas contas, ela não permite flexibilidade nas informações a priori e por isso a hipótese de independência acaba sendo mais conveniente. Dado que não haverá uma conjugada natural para o modelo com Ω (e vamos ter que fazer simulação de qualquer jeito), usaremos uma priori mais geral para todo mundo. Fica faltando uma priori para Ω, mas vamos pular essa etapa, por dois motivos: (1) podemos pensar na posteriori sem pensar ainda nessa priori e (2), sem assumir uma forma para Ω, não podemos ter uma priori para ela; isto é, a quantidade de parâmetros dessa matriz depende da estrutura dos erros que estamos assumindo para o modelo.

5.3 A densidade a posteriori

Como sabemos, a posteriori é proporcional à priori vezes a função de verossimilhança:

$$p(\beta, h, \Omega|Y) \propto p(\Omega) \times \exp\left\{-\frac{1}{2}\left[\overbrace{h(Y^* - X^*\beta)'(Y^* - X^*\beta)}^{\text{verossimilhança}} + \overbrace{(\beta - \underline{\beta})'\underline{V}^{-1}(\beta - \underline{\beta})}^{p(\beta)}\right]\right\} \overbrace{h^{\frac{N+\underline{\nu}-2}{2}} \exp\left\{-\frac{h\underline{\nu}}{2\underline{s}^{-2}}\right\}}^{p(h)} \tag{201}$$

A densidade a posteriori em (201) não tem a forma de nenhuma f.d.p. conhecida de forma que não é possível amostrar diretamente dela. No entanto, é possível amostrar das f.d.p.'s condicionais:

$$\beta \mid Y, h, \Omega \sim \mathcal{N}(\bar{\beta}, \bar{V}), \tag{202}$$

sendo

$$\bar{V} = (\underline{V}^{-1} + hX'\Omega^{-1}X)^{-1} \tag{203}$$

e

$$\bar{\beta} = \bar{V}\left(\underline{V}^{-1}\underline{\beta} + hX'\Omega^{-1}X\hat{\beta}(\Omega)\right). \tag{204}$$

Além disso, temos

$$h \mid y, \beta, \Omega \sim \mathcal{G}(\bar{s}^{-2}, \bar{\nu}) \tag{205}$$

em que

$$\bar{\nu} = N + \underline{\nu} \tag{206}$$

e

$$\bar{s}^2 = \frac{(Y - X\beta)'\Omega^{-1}(Y - X\beta) + \underline{\nu}\underline{s}^2}{\bar{\nu}}. \tag{207}$$

A posteriori para Ω dependerá de sua priori $p(\Omega)$ e, dados β e h, terá o seguinte núcleo:

$$p(\Omega \mid y, \beta, h) \propto p(\Omega) \cdot |\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{h}{2}(Y - X\beta)'\Omega^{-1}(Y - X\beta)\right\}. \tag{208}$$

Se (208) assumir a forma de uma f.d.p. conhecida, poderíamos usar o amostrador de Gibbs normalmente como para o MNCRL com priori Normal-Gama independente, apenas adicionando um bloco para Ω. Entretanto, dada a forma desconhecida em (208) e teremos que derivar simuladores específicos para cada especificação de Ω e $p(\Omega)$. Inicialmente, iremos supôr que ela tem uma heterocedasticidade de forma conhecida.

5.3.1 Heterocedasticidade de forma conhecida

Em nosso modelo, heterocedasticidade significa que:

$$\Omega = \begin{bmatrix} \omega_1 & 0 & 0 & \dots & 0 \\ 0 & \omega_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega_n \end{bmatrix} \tag{209}$$

ou seja, os erros ε_i possuem variâncias diferentes. Em vários estudos transversais (*cross-section*), sabemos qual é a estrutura da heterocedasticidade, de forma que podemos escrever[45]

$$\omega_i = g(z_i, \gamma), \tag{210}$$

na qual g é uma função positiva que depende dos parâmetros γ e de um vetor de observáveis z_i, que pode incluir colunas de X. Uma escolha comum para $g(\cdot)$ é[46]:

$$g(z_i; \gamma) = (1 + \gamma_1 \cdot z_{i1} + \gamma_2 \cdot z_{i2} + \cdots + \gamma_p \cdot z_{ip})^2 \tag{211}$$

Dessa maneira, $\Omega = \Omega(\gamma)$ (lê-se "Ω é função de γ") e precisamos simplesmente amostrar γ para conhecermos Ω (pois z_i são da base de dados). Porém, usando (211) em (208) não nos dá uma densidade para $p(\gamma|Y, \beta, h)$ e precisamos utilizar um outro simulador de MCMC para amostrar de $p(\gamma|Y, \beta, h)$. Este simulador é bem geral e é conhecido como *Algoritmo*

[45] Nas aulas, ao invés de γ, foi utilizada a notação α. Essa notação mudou aqui nas notas de aula para não confundir com a probabilidade de aceitação do algoritmo de Metropolis-Hastings.

[46] Note que o primeiro elemento é 1 pois "puxamos" o termo h^{-1} para fora de Ω, exigindo que não precisamos conhecer um termo α_0 de intercepto.

de Metropolis-Hastings. Nosso esquema será: usar um bloco do Gibbs para amostrar $\beta|h, Y, \Omega$, outro bloco para amostrar $h|\beta, Y, \Omega$ e um terceiro bloco que irá usar um Metropolis-Hastings para estimar γ (e com isso obter Ω). Esse procedimento é chamado de *Metropolis within Gibbs* (Metrópolis dentro do Gibbs). Faremos uma apresentação geral desse amostrador e depois o utilizaremos para o MNRL com erros heterocedásticos com forma funcional conhecida.

5.4 O algoritmo Metropolis-Hastings

O algoritmo[47] Metropolis-Hastings (M.H.) é útil em situações nas quais amostrar da posteriori é difícil ou impossível devido a sua forma funcional, mas é possível amostrar de uma aproximação dela (chamada de *densidade candidata*). M.H. irá, em geral, gerar uma cadeia de Markov e, portanto, é um algoritmo de MCMC, assim como o amostrador de Gibbs. Todos os diagnósticos de MCMC vistos podem e devem ser usados para avaliar o M.H..

O algoritmo tem a seguinte estrutura:

Algoritmo 3: Algoritmo de Metropolis-Hastings

Entrada: Um valor inicial, $\theta^{(0)}$.
Saída: $\{\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_B^{(s)}\}_{s=1}^{S}$
início
  para $s = 1, \ldots, S$ **gere**
    1. Uma realização θ^* da densidade candidata $q(\theta^{(s-1)}, \theta)$;
    2. Uma probabilidade de aceitação $\alpha(\theta^{(s-1)}, \theta^*)$;
    3. A atualização de $\theta^{(s)} = \theta^*$ com probabilidade $\alpha\left(\theta^{(s-1)}, \theta^*\right)$ ou mantenha $\theta^{(s)} = \theta^{(s-1)}$ com probabilidade $1 - \alpha(\theta^{(s-1)}; \theta^*)$;
  fim
  Calcule a média das S amostras, $g(\theta^{(1)}, g(\theta^{(2)}, \ldots, g(\theta^{(S)}$.
fim

Intuitivamente, o que fazemos é utilizar uma densidade candidata, $q(\cdot)$ para "amostrar" valores do parâmetro de interesse. Como ela não é exatamente igual à posteriori $p(\theta|y)$, nem todos os valores amostrados de $q(\cdot)$ serão mantidos: a probabilidade de amostrar um valor e manter ele na nossa seleção é igual à α. A figura (27) representa um caso mais simples para deixar claro que ao amostrarmos da candidata teremos valores de $q(\cdot)$ que não estamos interessados mas são sorteados por estarem em regiões onde $q(\cdot)$ é grande e outros valores que gostaríamos de amostrar, porém estão na cauda da densidade candidata (ao mesmo tempo que se encontram numa região de alta densidade a posteriori):

A figura (28) mostra o papel de α no processo: nas regiões laranjas da figura o valor de α deve ser baixo pois estamos amostrando nessas regiões mais do que deveríamos. Por outro lado, nas regiões azuis, estamos amostrando menos do que deveríamos (lembre-se que as amostras estão vindo da densidade candidata q, representada pela linha suave laranja). Dessa maneira, no ponto θ^{s-1} amostramos menos da posteriori e no ponto θ^* amostramos mais da posteriori, logo α em θ^* deve ser pequeno, para "compensar".

Assim como o amostrador de Gibbs, o M.H. também exige a escolha de um valor inicial $\theta^{(0)}$, e suas amostras dependendem de $\theta^{(S-1)}$. Logo, para garantir que $\theta^{(0)}$ não tem mais efeito e que a cadeia de Markov tenha convergido para a distribuição invariante $p(\theta|Y)$, é necessário descartar as primeiras S_0 realizações. Com base em $S_1 = S - S_0$ realizações de θ, podemos estimar $\mathbb{E}[g(\theta)|Y]$ através de $\frac{1}{S_1} \sum_{s=S_0+1}^{S} g\left(\theta^{(s)}\right)$. Entretanto, ainda precisamos falar mais a respeito da probabilidade de

[47] Seção 5.5 de Koop (2003)

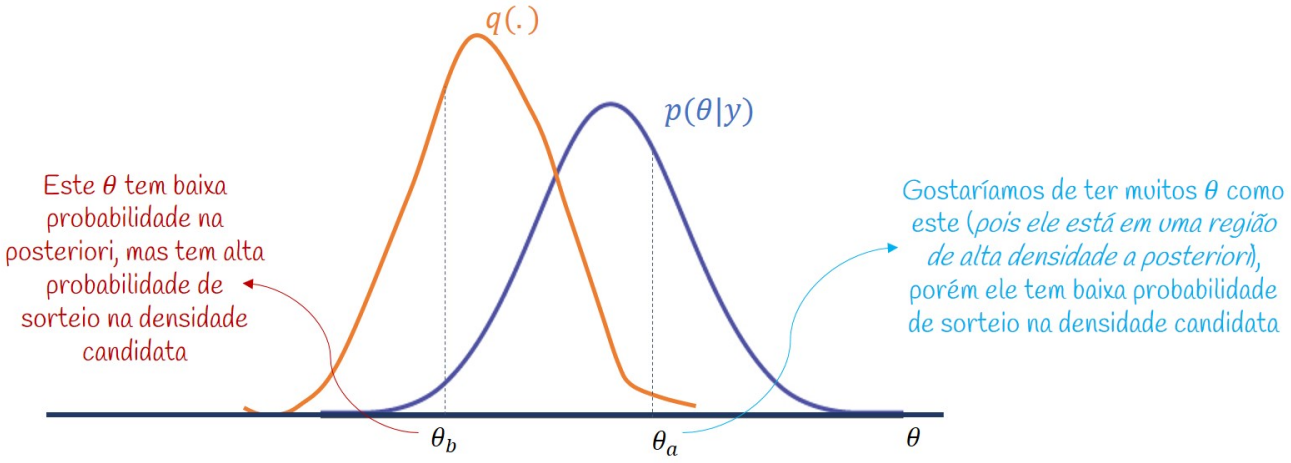


Figura 27: Ideia do algoritmo de M.H..

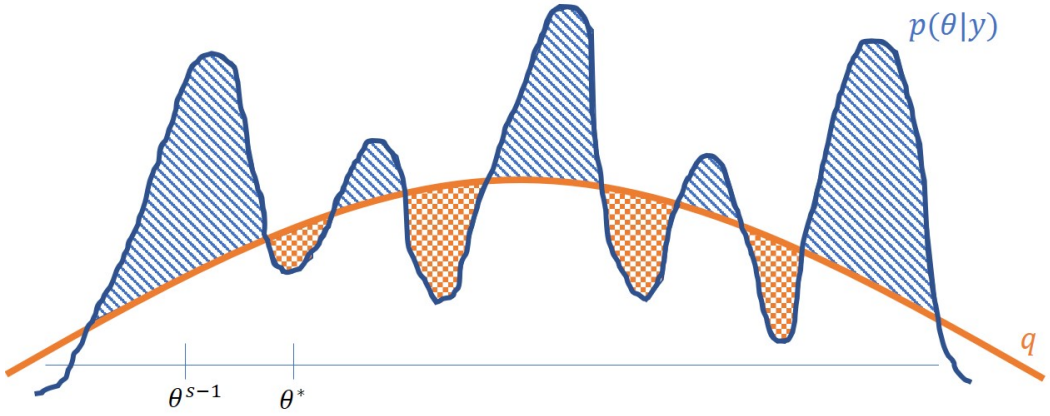


Figura 28: Intuição do funcionamento da probabilidade de aceitação α do algoritmo de M.H..

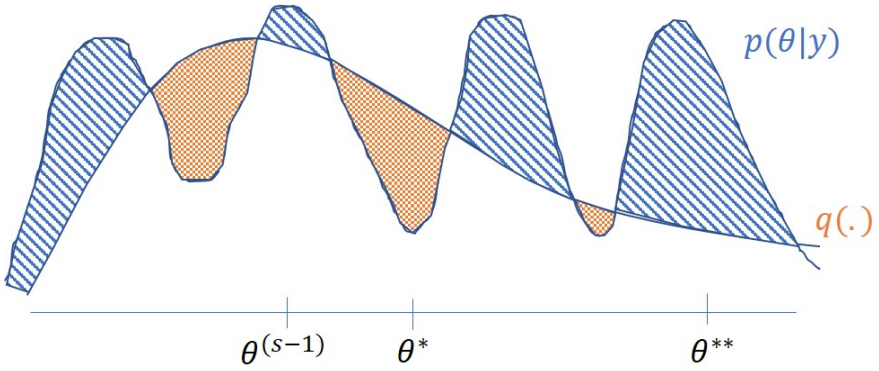
aceitação $\alpha\left(\theta^{(s-1)}; \theta^*\right)$.

Como $q(\cdot) \neq p(\theta|y)$[48], é necessário algum tipo de correção para que realizações de $q(\cdot)$ sejam consideradas realizações de $p(\theta|y)$ e é justamente a probabilidade de aceitação α que irá realizar essa correção. A probabilidade de aceitação do M.H. será alta se $\theta^{(s-1)}$ está em região de baixa importância em relação a θ^*, mas será baixa se $\theta^{(s-1)}$ estiver em uma região de muita massa de probabilidade a posteriori. Ao não sair de regiões de grande massa de probabilidade, o algoritmo dá mais peso a estas de regiões do que $q(\cdot)$ daria, enquanto ao sair com alta probabilidade de regiões de baixa importância para $p(\theta|y)$ o algoritmo dá menos peso a essas regiões do que $q(\cdot)$ daria.

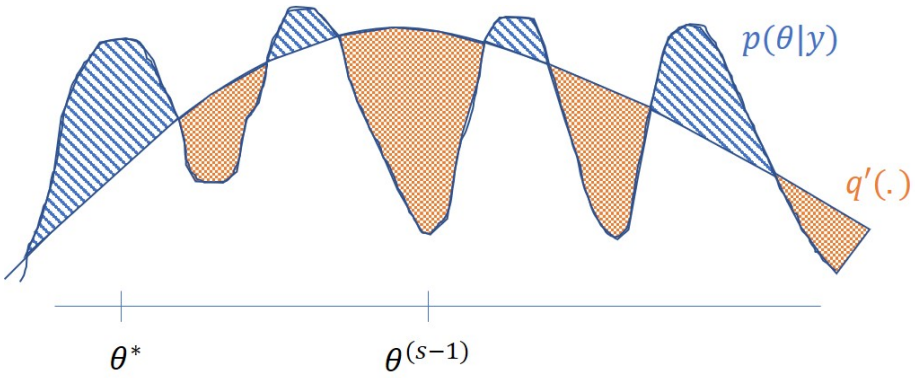
Apesar do algoritmo possuir uma tendência a sair das regiões de baixa probabilidade a posteriori, não queremos que ele nunca vá a estas regiões, pois ele deve explorar todo o suporte de $p(\theta|y)$. O ponto é que realizações nessas regiões devem ocorrer proporcionalmente menos do que nas realizações de outras regiões de suporte de $p(\theta|y)$.

Na Figura (29a), o ponto atual $\theta^{(s-1)}$ está em um local onde a posteriori tem mais massa de probabilidade do que a densidade candidata. Já a região onde θ^* foi amostrado é uma região de baixa importância à posteriori relativa à densidade candidata (áreas em laranja). Isso significa que o algoritmo, provavelmente, irá visitar essas áreas com maior frequência do que gostaríamos, portanto, a probabilidade de aceitação desse θ^*, dado que estamos em $\theta^{(s-1)}$, deve ser baixa. Se por outro lado nós calculamos θ^{**}, que está em uma região com quase a mesma importância a posteriori que $\theta^{(s-1)}$, mas na qual a candidata tem baixa massa de probabilidade, a probabilidade de aceitação de θ^{**} deve ser mais alta, pois $q(\cdot)$ não irá visitar

[48]Lembre-se que nós não conhecemos a posteriori $p(\theta|y)$.



(a) θ^* está em região de menor importância que $\theta^{(s-1)}$ e θ^{**} em região de maior importância que $\theta^{(s-1)}$



(b) θ^* em região de maior importância que $\theta^{(s-1)}$

Figura 29: Determinação do α no algoritmo de M.H. e mudança da $q(\cdot)$ conforme mudança do $\theta^{(s-1)}$

essa região com tanta frequência quanto irá visitar a região onde $\theta^{(s-1)}$ se encontra. Analogamente, se estivermos na situação da Figura (29b), deveremos ter uma alta probabilidade de aceitação pois estamos em uma região no ponto $\theta^{(s-1)}$ na qual a densidade a posteriori tem pouca massa em comparação com a candidata, o que significa que iremos sobreamostrar aquela região usando q e não precisamos ficar naquela região. Note que na figura (29a) houve uma mudança no formato da função candidata. No painel (29a) a média de $q(\cdot)$ estava centrada mais à esquerda (onde está o valor $\theta^{(s-1)}$, enquanto que no painel (29a) a densidade $q'(\cdot)$ mudou para uma distribuição um pouco mais simétrica (e ainda centrada em $\theta^{(s-1)}$. Neste caso, a densidade candidata depende do valor $\theta^{(s-1)}$, e portanto irá mudar de forma a cada passo. Por exemplo, considere um modelo AR: $\theta_t = \rho\theta_{t-1} + \varepsilon_t$; onde $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Sabemos que $\mathbb{E}[\theta_t] = \rho\theta_{t-1}$, de forma que no momento que temos um θ_{t-1} diferente, esta média muda. A família da $q(\cdot)$ não irá mudar, o que muda é apenas os seus parâmetros (por exemplo, variância, média, etc).

A probabilidade de aceitação do M.H. é dada por:

$$\alpha\left(\theta^{(s-1)}, \theta^*\right) = \min\left[\frac{p\left(\theta = \theta^*|y\right) \cdot q\left(\theta^*, \theta = \theta^{(s-1)}\right)}{p\left(\theta = \theta^{(s-1)}|y\right) \cdot q\left(\theta^{(s-1)}, \theta = \theta^*\right)}, 1\right] \tag{212}$$

Note que para calcular (212) não é necessário conhecer a constante de integração de $p(\theta|y)$, pois ela se cancelará com a razão $\frac{p(\theta=\theta^*|y)}{p(\theta=\theta^{(s-1)}|y)}$.

Qual a relevância da escolha de $q(\cdot)$? Bom, se escolhermos uma $q(\cdot)$ de maneira que não tenha suporte em comum com $p(\theta|y)$, então nunca iremos amostrar valores que poderiam ter vindo da posteriori, pois neste caso as amostras provenientes da densidade candidata não poderiam se "parecer" com valores da densidade posterior.Além disso, mesmo que o suporte delas

seja comum, não significa que iremos conseguir convergência rapidamente. Na figura (30) temos uma densidade candidata que tem suporte comum com a posteriori, porém nos locais onde a posteriori tem uma densidade maior, a candidata tem apenas a cauda e vice-versa, de maneira que seriam necessárias muitas amostras da candidata para termos uma sequência de valores apropriada para inferência a posteriori. Logo, a escolha de $q(\cdot)$ determina a velocidade de convergência do algoritmo. Dependendo das características da densidade candidata, teremos diferentes versões do algoritmo de M.H., como iremos ver nas próximas subseções.

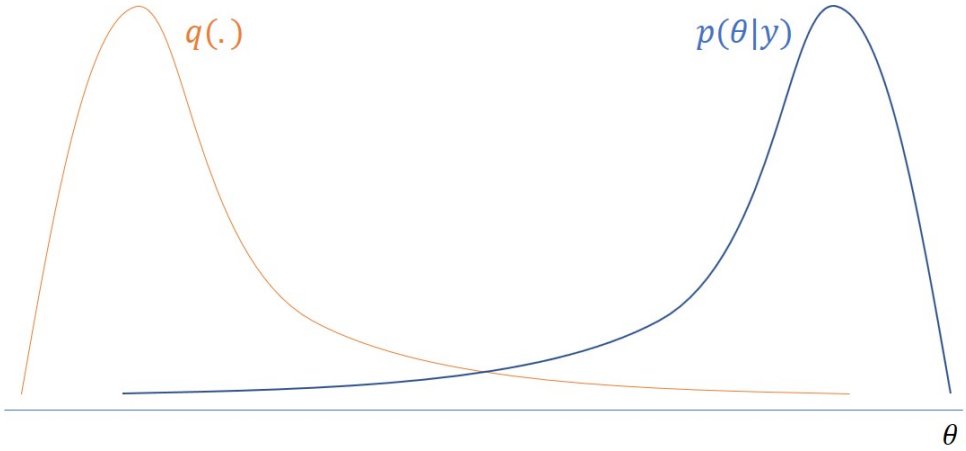


Figura 30: Exemplo de densidade candidata com pouco suporte comum com $p(\theta|y)$

5.4.1 Metropolis-Hastings Cadeia Independente

Quando existe uma aproximação conveniente (e boa) para $p(\theta|y)$, será possível usar uma distribuição candidata que não depende de $\theta^{(s-1)}$ (isto significa que $q(\cdot)$ não fica mudando a cada passo, ela fica fixa). Um exemplo está ilustrado na figura (27). Nesse caso, a probabilidade de aceitação simplifica para:

$$\alpha\left(\theta^{(s-1)}, \theta^*\right) = \min\left[\frac{p\left(\theta = \theta^*|y\right) \cdot q^*\left(\theta = \theta^{(s-1)}\right)}{p\left(\theta = \theta^{(s-1)}|y\right) \cdot q^*\left(\theta = \theta^*\right)}, 1\right] \tag{213}$$

Uma prática comum para encontrar $q^*(\theta)$ é usar uma distribuição t com parâmetros $\hat{\theta}_{ML}$, $\widehat{Var(\hat{\theta}_{ML})}$ e graus de liberdade ν, escolhido de forma a garantir uma probabilidade de aceitação próxima de 0.5. A ideia que está por trás dessa escolha é proveniente dos resultados vistos anteriormente: quando a amostra vai para ∞, o peso da priori vai a zero e toda nossa posteriori fica em cima das estimativas de máxima verossimilhança. Só que como os resultados de convergência nos levam para uma distribuição normal, que tem caudas leves, opta-se por utilizar uma distribuição com caudas mais pesadas, no caso a t. Para mais detalhes, ver Greenberg (2008).

5.4.2 Metropolis-Hastings Passeio Aleatório

O algoritmo M.H.P.A. é útil na situação oposta do algoritmo cadeia independente: quando não é possível conseguir uma boa aproximação para $p(\theta|y)$. Enquanto o M.H. cadeia independente escolhe $q(\cdot)$ para aproximar $p(\theta|y)$, o M.H.P.A. apenas escolhe uma família de densidades que vai explorar livremente o suporte da distribuição a posteriori para aproximá-la. Ou seja, os momentos da distribuição candidata irão mudar, fazendo com que as distribuições candidatas naveguem pelo suporte de $p(\theta|y)$.

Formalmente, no M.H.P.A. a densidade candidata $q(\theta^{(s-1)};\theta)$ é definida por:

$$\theta^* = \theta^{(s-1)} + z \tag{214}$$

em que z é chamada de variável aleatória incremental e irá determinar a família da densidade candidata. Usualmente, $z \sim \mathcal{N}(0,\Sigma)$ é escolhida e, nesse caso, $\theta^{(s-1)}$ determinará a média da densidade candidata. Esta situação é similar à que está ilustrada na Figura (31), porém com uma distribuição que não é a distribuição normal *porque a Aisha não consegue desenhar bem a distribuição normal*.

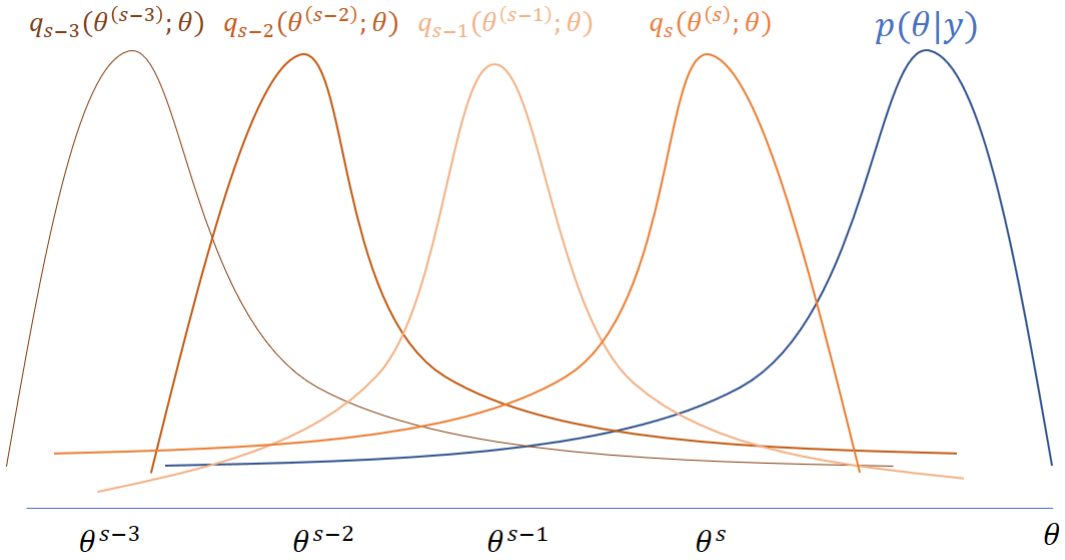


Figura 31: Exemplo de M.H. passeio aleatório

A matriz Σ deve ser escolhida de forma a gerar uma probabilidade de aceitação não muito baixa[49]. Se α é pequeno, isso significa que novas realizações são raramente aceitas e demorará muito para que a cadeia explore todo o suporte de $p(\theta|y)$, exigindo um valor de S muito grande. Esse caso (α pequeno) é evidência de que Σ é muito grande e amostras estão sendo obtidas muito longe da região de importância de $p(\theta|y)$. Por outro lado, se α é muito grande, isso significa que Σ é muito pequeno e as realizações θ^* estão muito próximas umas das outras, o que .

A distribuição candidata no caso do M.H.P.A. deve ser simétrica (para mais detalhes, veja Robert and Casella (2010b)). Por exemplo, no caso da densidade candidata normal, temos:

$$q(\theta^{(s-1)};\theta) = \frac{1}{(2\pi)^{\frac{K}{2}}}|\Sigma|^{-\frac{1}{2}}\left\{-\frac{1}{2}(\theta-\theta^{(s-1)})'\Sigma^{-1}(\theta-\theta^{(s-1)})\right\}. \tag{215}$$

Assim, $q(\theta^{(s-1)},\theta=\theta^*) = q(\theta^*,\theta=\theta^{(s-1)})$ (verifique!) e, com isso, a probabilidade de aceitação dada em 212 simplifica para:

$$\alpha\left(\theta^{(s-1)},\theta^*\right) = \min\left[\frac{p(\theta=\theta^*|y)}{p(\theta=\theta^{(s-1)}|y)},1\right], \tag{216}$$

o que deixa claro que a cadeia de Markov com passeio aleatório tenderá a se mover na direção de maior probabilidade a posteriori.

[49] Regra de bolso é que seja enttre 0.25 e 0.40. Será obviamente menor do que na cadeia independente.

5.4.3 Metropolis dentro do Gibbs

É possível mostrar que o uso do M.H. para amostrar de uma distribuição posterior condicional de um algoritmo de Gibbs é perfeitamente válida. Isto é, se temos fórmula analítica para $p(\theta_{(1)}|y, \theta_2, \theta_{(3)})$ e para $p(\theta_{(2)}|y, \theta_{(1)}, \theta_{(3)})$, mas não para $p(\theta_{(3)}|y, \theta_{(1)}, \theta_{(2)})$, podemos usar as fórmulas analíticas para $\theta_{(1)}^{(s)}$ e $\theta_{(2)}^{(s)}$, e um M.H. para amostrarmos $\theta_{(3)}^{(s)}$. As realizações $\{\theta^{(i)}\}_{i=S_0+1}^{S}$ obtidas dessa forma são realizações válidas de $p(\theta|y)$.

Portanto, voltando ao MNRL com heterocedasticidade de forma conhecida, temos:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, h^{-1}\Omega)$$

em que Ω é diagonal com elementos dados por:

$$\omega_i \equiv g(z_i; \gamma) = (1 + \gamma_1 \cdot z_{i1} + \gamma_2 \cdot z_{i2} + \cdots + \gamma_p \cdot z_{ip})^2. \tag{217}$$

Usando a priori normal gama independente para β e h como já estabelecidas, respectivamente, em (199) e (200), mais uma priori não informativa e imprópria para γ dada por:

$$p(\gamma) \propto 1, \tag{218}$$

pPodemos desenvolver o seguinte algoritmo Metropolis dentro de Gibbs com 3 blocos de parâmetros:

- Bloco β: $\beta|y, h, \gamma \sim \mathcal{N}(\bar{\beta}, \bar{V})$, sendo que $\bar{V}$ e $\bar{\beta}$ estão definidos em (203) e (204);
- Bloco h: $h|y, \beta, \gamma \sim \mathcal{G}(\bar{s}^2, \bar{\nu})$, com $\bar{s}^2$ e $\bar{\nu}$ definidos como em (206) e (207);
- Bloco γ: $\gamma|y, \beta, h \sim p(\gamma|y, \beta, h)$

Usando (217) e (218) em (208), temos:

$$p(\gamma|y, \beta, h) \propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{h}{2}(Y - X\beta)'\Omega^{-1}(Y - X\beta)\right\} \tag{219}$$

Como (219) não sugere nenhuma distribuição conhecida, podemos usar o M.H. passeio aleatório com $z \sim \mathcal{N}(0, \Omega)$. Para selecionar Ω, podemos começar com $\Omega^{(0)} = c \cdot \mathbb{I}$ sendo que c é escolhido de forma a termos $\alpha \approx 0.2$. Então, podemos usar essas realizações iniciais para estimar $\Omega = c \cdot \widehat{Var(\theta)}$. Com $\Omega = c \cdot \widehat{Var(\theta)}$ rodamos a cadeia final que nos dará $\left\{\theta^{(i)}\right\}_{i=S_0+1}^{S}$. A constante c serve para "inflar" a variância e não termos nossos valores amostrados todos próximos uns dos outros.

6 Parte 6 - Modelos para Vetores Autoregresssivos (VAR)

Modelos VAR são uma generalização dos modelos AR univariados que, por sua vez, são um caso especial do modelo de regressão linear, no qual $X_t = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ (a matriz de variáveis explicativas contém apenas a constante e defasagens da variável dependente, isto é, as variáveis explicativas são simplesmente as variáveis explicadas defasadas). Isso sugere que podemos utilizar as várias técnicas já discutidas para MNRL para realizar inferência Bayesiana em modelos VAR.

De acordo com Del Negro and Schorfheide (2013), o modelo VAR, que foi inicialmente proposto por Sims (1980), é uma das principais ferramentas empíricas para macroeconomia. A ideia de Sims era ter uma ferramenta alternativa aos modelos macroeconométricos de larga escala desenvolvidos na década de 60 (decorrentes do trabalho de Tinbergen), uma vez que estes últimos impunham restrições grandes aos dados e acabavam se tornando inconsistentes com a ideia de que os agentes tomam decisões ótimas a cada instante de tempo. Desde então os modelos VAR tem sido utilizados para previsões macroeconômicas, análise de políticas para investigar as fontes das flucutações no ciclo econômico e também como referência para a avaliação e desenvolvimento da teoria macroeconomica dinâmica.

Modelos VAR são modelos lineares para séries de tempo desenvolvidos para capturar a dinâmica conjunta das séries e, no caso de um $VAR(1)$, podemos escrever o modelo como:

$$Y_t = \alpha_0 + \beta Y_{t-1} + \varepsilon_t \tag{220}$$

em que

- Y_t é um vetor $M \times 1$ de séries temporais;
- $\varepsilon_t \sim N(0, \Sigma)$ (iid) é um vetor $M \times 1$ de erros;
- α_0 é um vetor $M \times 1$ de interceptos;
- β é uma matriz $M \times M$ de coeficientes.

No caso mais geral do $VAR(p)$, temos:

$$Y_t = \alpha_0 + \sum_{j=1}^{p} \beta_j Y_{t-j} + \varepsilon_t \tag{221}$$

sendo que β_j são matrizes $M \times M$ de coeficientes (existe uma matriz β para cada defasagem). Variáveis exógenas, tendências determinísticas, e/ou componentes sazonais também podem ser incorporados a (221), mas não o faremos aqui para manter a notação mais simples.

O modelo VAR pode ser escrito em forma matricial usando um vetor Y de dimensão $M \cdot T \times 1$, o qual empilha todas as T observações das M séries (e vai ter uma distribuição normal multivariada), ou usando uma matriz Y de dimensão $T \times M$, que empilha as observações y_t'. Definindo $x_t = (1, Y_{t-1}', \ldots, Y_{t-p}')$ (x_t é $1 \times K$), $X = [x_1', \ldots, x_T']$ (isto é, X_t é uma matriz $T \times K$) e se $K = 1 + M \cdot p$ é o número de coeficientes em cada equação do VAR, então X é $T \times K$. Além disso, se $A = (\alpha_0, \beta_1, \ldots, \beta_p)'$, então $\alpha = vec(A)$ é um vetor $K \cdot M \times 1$ que empilha todos os coeficientes do VAR. Com essas definições podemos escrever o VAR da equação (220) como:

$$Y_{T \times M} = X_{T \times K} \cdot A_{K \times M} + E_{T \times M} \qquad \text{(modelo matricial)} \tag{222}$$

ou como

$$y_{M\cdot K\times 1} = (\mathbb{I}_M \otimes X_{T\times K})\alpha_{K\cdot M\times 1} + \varepsilon_{M\cdot T\times 1} \qquad \text{(modelo empilhado)}, \tag{223}$$

em que $\otimes$ é o produto de Kronecker, $\varepsilon \sim \mathcal{N}(0, \Sigma\otimes\mathbb{I}_T)$ é um vetor $M\cdot T\times 1$ análogo ao y e $E = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T)' \sim MN(0, \Sigma, \mathbb{I}_T)$ (é uma distribuição matriz normal

6.1 Função de Verossimilhança

Observe que a densidade de (223) é uma normal multivariada pois y é vetor, isto é, $y_t \sim \mathcal{N}(X_t\beta, \Sigma)$). Formalmente, temos,

$$p(y|\alpha, \Sigma) = (2\pi)^{\frac{TM}{2}} |\Sigma \otimes \mathbb{I}_T|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[y - (\mathbb{I}_M \otimes X)\alpha\right]' (\Sigma \otimes \mathbb{I}_T)^{-1} \left[y - (\mathbb{I}_M \otimes X)\alpha\right]\right\} \tag{224}$$

A equação (224) é idêntica à expressão da verossimilhança da (222) (a prova pode ser vista em Bauwens et al. (2003)) e será proporcional ao produto de um núcleo Wishart com o núcleo de uma normal:

$$\begin{aligned} p(y|\alpha, \Sigma) &\propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}[(\alpha - \hat{\alpha})'(\Sigma^{-1} \otimes X'X)(\alpha - \hat{\alpha})]\right\} \times \\ &\times \exp\left\{-\frac{1}{2} tr\left[(\Sigma^{-\frac{1}{2}} \otimes \mathbb{I}_T)y - \left(\Sigma^{-\frac{1}{2}} \otimes X\right)\hat{\alpha}\right]' \left[(\Sigma^{-\frac{1}{2}} \otimes \mathbb{I}_T)y - \left(\Sigma^{-\frac{1}{2}} \otimes X\right)\hat{\alpha}\right]\right\} \\ &\propto |\Sigma|^{-\frac{K}{2}} \exp\left\{-\frac{1}{2}[(\alpha - \hat{\alpha})'(\Sigma^{-1} \otimes X'X)(\alpha - \hat{\alpha})]\right\} \times |\Sigma|^{-\frac{T-K}{2}} \exp\left\{-\frac{1}{2} tr[\Sigma^{-1}S]\right\} \end{aligned} \tag{225}$$

em que $S = [(y - (\mathbb{I}_M \otimes X)\hat{\alpha})'(y - (\mathbb{I}_M \otimes X)\hat{\alpha})]$ é o inverso da matriz de escala da distribuição Wishart (ver apêndice B de Koop (2003)). Ou seja, a função de verossimilhança para o BVAR $y = (\mathbb{I}_m \otimes X)\alpha + \varepsilon$ é proporcional ao produto de uma normal para $\alpha|\Sigma, y$ e uma Wishart para $\Sigma^{-1}|y$ com

$$\alpha|\Sigma, y \sim \mathcal{N}(\hat{\alpha}, \Sigma \otimes (X'X)^{-1}) \tag{226}$$

e

$$\Sigma^{-1}|y \sim \mathcal{W}(S^{-1}, T - K - M - 1), \tag{227}$$

em que S foi definido anteriormente. É um modelo com muitos parâmetros: se tivermos 3 variáveis e 2 defasagens, então o modelo terá 21 coeficientes a serem estimados. Caso se tenham 5 variáveis e 4 defasagens, esse número aumenta para 105 coeficientes.

Demonstração. Desconsiderando as constantes em (224), podemos trabalhar com o símbolo de proporcionalidade e expandir os termos dentro da exponencial para obter:

$$
\begin{aligned}
p(y|\alpha,\Sigma) &\propto |\Sigma\otimes\mathbb{I}_T|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[y-(\mathbb{I}_M\otimes X)\alpha\right]'(\Sigma\otimes\mathbb{I}_T)^{-1}\left[y-(\mathbb{I}_M\otimes X)\alpha\right]\right\}\\
&\propto |\Sigma\otimes\mathbb{I}_T|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\Sigma^{-1/2}\otimes\mathbb{I}_T)\left[y-(\mathbb{I}_M\otimes X)\alpha\right]\right]'(\Sigma^{-1/2}\otimes\mathbb{I}_T)\left[y-(\mathbb{I}_M\otimes X)\alpha\right]'\right\}\\
&\propto |\Sigma\otimes\mathbb{I}_T|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\Sigma^{-1/2}\otimes\mathbb{I}_T)y-(\Sigma^{-1/2}\otimes\mathbb{I}_T)(\mathbb{I}_M\otimes X)\alpha\right]'\left[(\Sigma^{-1/2}\otimes\mathbb{I}_T)y-(\Sigma^{-1/2}\otimes\mathbb{I}_T)(\mathbb{I}_M\otimes X)\alpha\right]\right\} \quad (228)
\end{aligned}
$$

em que usamos a regra de que a inversa do produto de kroneker é o produto de kroneker das inversas, isto é, $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$. Além disso, também usamos $\Sigma^{-1}\otimes\mathbb{I}_T = (\Sigma^{-1/2}\otimes\mathbb{I}_T)(\Sigma^{-1/2}\otimes\mathbb{I}_T)$ (isto decorre da decomposição de Cholesky, mas como Σ é simétrica, não precisa o transposto). Nós usamos a decomposição de Cholesky para que a nossa variável fique como "padronizada". Isto fica claro no último passo de (228), pois a matriz de variâncias e covariâncias não aparece mais.

Definindo $\hat{\alpha} = (\Sigma^{-1}\otimes X'X)^{-1}(\Sigma^{-1}\otimes X)'y$, podemos escrever (some e subtraia $\hat{a}$ no lado esquerdo):

$$
(\Sigma^{-1/2}\otimes\mathbb{I}_T)y-(\Sigma^{-1/2}\otimes X)\alpha = (\Sigma^{-1/2}\otimes\mathbb{I}_T)y-(\Sigma^{-1/2}\otimes X)\hat{\alpha}+(\Sigma^{-1/2}\otimes X)(\hat{\alpha}-\alpha). \quad (229)
$$

Usando a equação (229) em (228) e lembrando que $(A\otimes B)(C\otimes D) = AC\otimes BD$, temos[50]:

$$
p(y|\alpha,\Sigma)\propto|\Sigma\otimes\mathbb{I}_T|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\Sigma^{-\frac{1}{2}}\otimes\mathbb{I}_T)y-\left(\Sigma^{-\frac{1}{2}}\otimes X\right)\hat{\alpha}\right]'\left[(\Sigma^{-\frac{1}{2}}\otimes\mathbb{I}_T)y-\left(\Sigma^{-\frac{1}{2}}\otimes X\right)\hat{\alpha}\right]\right\} \quad (230)
$$

$$
\cdot\ \exp\left\{-\frac{1}{2}[(\alpha-\hat{\alpha})'(\Sigma^{-1}\otimes X'X)(\alpha-\hat{\alpha})]\right\}. \quad (231)
$$

A primeira parte (Equação 230) vamos fazer virar uma Wishart (pois aparece uma soma de quadrados dos resíduos) e a segunda parte (Equação 231) vai virar uma distribuição normal[51] e chegamos na Equação (225):

$$
\begin{aligned}
p(y|\alpha,\Sigma) &\propto |\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}[(\alpha-\hat{\alpha})'(\Sigma^{-1}\otimes X'X)(\alpha-\hat{\alpha})]\right\}\times\\
&\times\exp\left\{-\frac{1}{2}tr\left[(\Sigma^{-\frac{1}{2}}\otimes\mathbb{I}_T)y-\left(\Sigma^{-\frac{1}{2}}\otimes X\right)\hat{\alpha}\right]'\left[(\Sigma^{-\frac{1}{2}}\otimes\mathbb{I}_T)y-\left(\Sigma^{-\frac{1}{2}}\otimes X\right)\hat{\alpha}\right]\right\}\\
&\propto |\Sigma|^{-\frac{K}{2}}\exp\left\{-\frac{1}{2}[(\alpha-\hat{\alpha})'(\Sigma^{-1}\otimes X'X)(\alpha-\hat{\alpha})]\right\}\times|\Sigma|^{-\frac{T-K}{2}}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}S]\right\}
\end{aligned}
$$

em que S foi definido anteriormente. □

Já a verossimilhança para a equação (222) será dada por:

$$
p(Y|A,\Sigma) = (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\,\mathrm{Tr}\left[\Sigma^{-1}(Y-XA)'(Y-XA)\right]\right\} \quad (232)
$$

[50] Estamos usando $(\Sigma^{-1/2}\otimes\mathbb{I}_T)(\mathbb{I}_M\otimes X) = \Sigma^{-1/2}\mathbb{I}_M\otimes\mathbb{I}_T X = \Sigma^{-1/2}X)$

[51] Perceba que o determinante em (230) fica somente o determinante de Σ repetido T vezes

que é equivalente a

$$p(Y|A,\Sigma) = \overbrace{(2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T-K}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}S\right]\right\}}^{\propto \mathcal{W}(S^{-1},T-K-M-1)} \tag{233}$$

$$\underbrace{\cdot\,|\Sigma|^{-\frac{K}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}(A-\hat{A})'X'X(A-\hat{A})\right]\right\}}_{\propto \mathcal{N}_{T\times M}(\hat{A},\Sigma,(X'X)^{-1})} \tag{234}$$

em que $S = (Y - X\hat{A})'(Y - X\hat{A})$. Assim, a função de verossimilhança de (221) é também proporcional ao produto de uma densidade matriz normal para $A|\Sigma$ por uma densidade Wishart para Σ^{-1}.

Demonstração. Usando (232) podemos expandir a forma quadrática para obter:

$$\begin{aligned} p(Y|A,\Sigma) &= (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}(Y-XA)'(Y-XA)\right]\right\} \\ &= (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}(Y'Y-2A'X'Y+A'X'XA)\right]\right\}. \end{aligned} \tag{235}$$

Agora, defina $\hat{A} = (X'X)^{-1}X'Y$. Então, some e subtraia $\hat{A}'X'X\hat{A}$ de (235) de forma que[52]

$$p(Y|A,\Sigma) = (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}(Y'Y-2A'X'Y+A'X'XA+\hat{A}'X'X\hat{A}-\hat{A}'X'X\hat{A})\right]\right\}. \tag{236}$$

Observe que $X'Y = \underbrace{X'X(X'X)^{-1}}_{\text{Identidade}}X'Y = X'X\hat{A}$. Aplicando isso em (236), chegamos em

$$p(Y|A,\Sigma) = (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}\left(Y'Y-\hat{A}'X'X\hat{A}+A'X'XA-2A'X'X\hat{A}+\hat{A}'X'X\hat{A}\right)\right]\right\}. \tag{237}$$

Note que $A'X'XA - 2A'X'X\hat{A} + \hat{A}'X'X\hat{A} = (A-\hat{A})'X'X(\hat{A}-A)$ e reescreva $-\hat{A}'X'X\hat{A}$ como $-2\hat{A}'X'X\hat{A} + \hat{A}'X'X\hat{A}$ em (237):

$$p(Y|A,\Sigma) = (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}\left(Y'Y-2\hat{A}'X'X\hat{A}+\hat{A}'X'X\hat{A}+(A-\hat{A})'X'X(\hat{A}-A)\right)\right]\right\}. \tag{238}$$

Uma vez que $\hat{A}'X' = \underbrace{\left[(X'X)^{-1}X'Y\right]'}_{\hat{A}'}X' = Y'X(X'X)^{-1}X'$, podemos escrever o termo $2\hat{A}'X'X\hat{A}$ da seguinte forma:

$$2Y'X\underbrace{(X'X)^{-1}X'X}_{\mathbb{I}}\hat{A} = 2Y'X\hat{A}. \tag{239}$$

Assim,

$$Y'Y - 2\hat{A}'X'X\hat{A} + \hat{A}'X'X\hat{A} = Y'Y - 2Y'X\hat{A} + \hat{A}'X'X\hat{A} = (Y-X\hat{A})'(Y-X\hat{A}) := S \tag{240}$$

e podemos escrever (238) da seguinte forma:

$$p(Y|A,\Sigma) = (2\pi)^{\frac{TM}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\Sigma^{-1}\left(S+(A-\hat{A})'X'X(\hat{A}-A)\right)\right]\right\}, \tag{241}$$

[52] Os grifos coloridos são para auxiliar a identificar onde os termos estão ao longo da demonstração.

que pode ser fatorado como em (233)-(234). □

6.2 A priori

Modelos VAR são altamente parametrizados. Note que α contém $K \cdot M$ parâmetros e em um VAR(4) com 5 variáveis dependentes teremos 105 elementos em α. Logo, sem o uso de informação a priori, é muito difícil obter estimativas precisas para esses modelos. Frente a isso, diversos tipos f.d.p. a priori foram desenvolvidas para modelos VAR.

6.2.1 A priori conjugada natural

As Equações (226) e (227) sugerem a seguinte estrutura para a priori conjugada natural

$$\alpha|\Sigma \sim \mathcal{N}(\underline{\alpha}, \Sigma \otimes \underline{V}) \tag{242}$$

e

$$\Sigma^{-1} \sim \mathcal{W}(\underline{S}^{-1}, \underline{\nu}), \tag{243}$$

sendo que $\underline{\alpha}$, $\underline{V}$, $\underline{\nu}$, $\underline{S}^{-1}$ são hiperparâmetros. O produto de kroneker dentro da priori para alpha é de certa forma restritiva uma vez que $\underline{V}$ impõe uma certa dependência entre V e Σ, e, assim, estamos impondo uma estrutura de variância e covariância que restringe nossos parâmetros. Se Σ é $M \times M$, então ela tem $M(M-1)/2$ elementos livres. Analogamente, $\underline{V}$ tem $T(T-1)/2$ elementos. Portanto, o produto de kroneker tem $[M(M-1) + T(T-1)]/2$ elementos (que é igual a $[M^2 - M + T^2 - T]/2$). Por outro lado, se tivéssemos uma matriz só, ela seria de tamanho $T \cdot M \times T \cdot M$ e com isso teria $TM(TM-1)/2$ elementos, que é muito mais do que se usamos o Kroneker. Isso quer dizer que ao utilizar a priori conjugada natural nós ganhamos em resultados analíticos mas perdemos em flexibilidade pois o número de elementos reduz da matriz de variâncias e covariâncias em α reduz consideravelmente.

A posteriori baseada na priori conjugada natural é dada por

$$\alpha|\Sigma, y \sim \mathcal{N}(\bar{\alpha}, \Sigma \otimes \bar{V}) \tag{244}$$

e

$$\Sigma^{-1}|y \sim \mathcal{W}(\bar{S}^{-1}, \bar{\nu}) \tag{245}$$

sendo que

- $\bar{V} = [\underline{V}^{-1} + X'X]^{-1}$,
- $\bar{A} = \bar{V}[\underline{V}^{-1}\underline{A} + X'X\hat{A}]$, isto é, $\bar{A}$ é a média ponderada da verossimilhança com a priori. Isso significa que a média a posteriori é uma média entre a média a priori e a media da posteriori, sendo que, assim como no modelo de regressão, terá o peso dos dados e o peso da priori,
- $\bar{\alpha} = \text{vec}(\bar{A})$,
- $\bar{\nu} = T + \underline{\nu}$.

Além disso, $\bar{S} = S + \underline{S} + \hat{A}'X'X\hat{A} + \underline{A}'\underline{V}^{-1}\underline{A} - \bar{A}'(\underline{V}^{-1} + X'X)\bar{A}$. Esses termos em X mostram como que o aumento das defasagens e das equações há um aumento de $\bar{S}$, que determina a média da Wishart e vai parar na variância dos $\bar{\alpha}$ (pois é como uma soma de termos ao quadrado).

Já as equações (233) e (234) sugerem o uso de uma priori conjugada natural da seguinte forma:

$$A|\Sigma \sim \mathcal{N}(\underline{A}, \Sigma, \underline{V}) \tag{246}$$

$$\Sigma^{-1} \sim \mathcal{W}(\underline{S}^{-1}, \underline{\nu}) \tag{247}$$

onde $\underline{A}$, $\underline{V}$, $\underline{\nu}$ e $\underline{S}^{-1}$ são hiperparâmetros. Cálculos similares aos do MNCRL com a priori conjugada nos permitem encontrar a seguinte f.d.p. a posteriori

$$A|\Sigma, Y \sim \mathcal{N}(\bar{A}, \Sigma, \bar{V}) \tag{248}$$

$$\Sigma^{-1}|Y \sim \mathcal{W}(\bar{S}^{-1}, \bar{\nu}) \tag{249}$$

sendo $\bar{V}$, $\bar{A}$, $\bar{\nu}$ e $\bar{S} = S + \underline{S} + \hat{A}'X'X\hat{A} + \underline{A}'\underline{V}^{-1}\underline{A} - \bar{A}'(\underline{V} + X'X)\bar{A}$ foram definidos anteriormente.

Assim como no capítulo 3 de Koop (2003), a f.d.p. a posteriori marginal de α (e A) tem fórmula analítica e é dada por uma distribuição t multivariada (matricial) com média $\bar{\alpha}$ ($\bar{A}$) e variância dada por:

$$Var[\alpha|y] = \frac{1}{\bar{\nu} - M - 1}\bar{S} \otimes \bar{V}.$$

Com estes resultados, podemos fazer inferência analítica no modelo VAR.

A distribuição preditiva para Y_{t+1} também possui fórmula analítica dada por $f(Y_{t+1}|Y_t) \sim t(X\bar{A}, [1 + X_{t+1}\bar{V}X'_{t+1}]\bar{S}, \bar{\nu} - 2)$, que é uma distribuição t matricial com média $X_{T+1}\bar{A}$ e variância dada por $\left[1 + X_{T+1}\bar{V}X'_{T+1}\right]\frac{\bar{S}}{\bar{\nu}-2}$. Porém, para previsões mais passos a frente será necessário usar simulações ou se basear no *método direto de previsões*. Essa fórmula nos diz que precisamos do X no futuro (em $t + 1$). Como X são as defasagens de Y, nós temos Y_t e portanto temosX_{t+1}. No entando, se quisermos calcular previsão mais passos à frente, nós não vamos ter X_{t+4} pois não conhecemos Y_{t+3} e teríamos que integrar fora esses valores de Y que não dispomos. Precisaremos de métodos de integração para poder fazer isso e já não teremos em fórmula fechada. Atualmente isso é bem simples, inclusive podemos fazer isso via integração de Monte Carlo. Como temos a preditiva um passo à frente em fórmula fechada, podemos simular vários deles e a partir disso conseguir fazer dois passos à frente e depois três, etc. O trabalho de Kadiyala and Karlsson (1997) explica como fazer essas previsões.

Os mesmos problemas da priori conjugada natural que tínhamos anteriormente no MNRL ocorrerão aqui (não é possível ser não informativo a respeito de Σ^{-1} e informativo a respeito de A). Adicionalmente, para o caso de VAR, a priori dada em (246) e (247) impede a utilização de diferentes variáveis explicativas em X para cada uma das equações devido à estrutura $(I_M \otimes X)$ de (223). Mais ainda, a estrutura $\Sigma \otimes \underline{V}$ da priori para α implica que a priori para os coeficientes da equação i devem ser proporcionais aos da equação j, $\forall\ i, j$. A variância a priori em relação a algum coeficiente i é proporcional à variância dos demais coeficientes, o que não permite que possamos modificar os impactos de cada um dos coeficientes de maneira individualizada. Poderíamos pensar em uma priori tipo normal wishart independente, porém isso iria requerer métodos de Monte Carlo sendo que na época que o modelo VAR foi desenvolvido, ainda não haviam tantos recursos computacionais.

6.3 A priori de Minnesota

A priori de minnesota é baseada na substituição da matriz desconhecida Σ por uma estimativa $\hat{\Sigma}$ dela. Ao substituir uma v.a. por um valor fixo, as f.d.p. posterioris e preditivas não irão refletir corretamente o efeito da incerteza em relação aos parâmetros. Porém, permitirá manter resultados analíticos para a distribuição posterior; além de sanar os problemas da priori

conjugada natural. Os trabalhos iniciais sobre o uso de distribuições a priori em VARs para encolher as estimativas e, com isso, melhorar sua precisão, foram realizadas na Universidade de Minnesota e no FED de Minneapolis. Portanto, esse tipo de priori ficou conhecido como a *priori de Minnesota*. A priori de Minnesota é baseada em uma aproximação que flexibiliza a definição da priori, ao mesmo tempo que mantém as expressões analíticas para a distribuição a posteriori.

A aproximação usada na priori de Minnesota é a substituição da matriz desconhecida Σ por uma estimativa $\widehat{\Sigma}$ dela. Ao substituir uma variável aleatória por uma constante ao invés de integrá-la como na abordagem bayesiana tradicional, as densidades preditivas não irão refletir corretamente o efeito da incerteza a respeito dos parâmetros em Σ, pois substituímos um valor que desconhecemos por uma constante (ao invés de atribuirmos uma densidade de probabilidade para caracterizar esse desconhecimento). Logo, a incerteza apresentada pela a posteriori de α será inferior à real incerteza sobre esse parâmetro. Entretanto, essa simplificação permitirá sanar alguns problemas da priori conjugada natural e ainda manter resultados analíticos para a f.d.p. a posteriori.

Se Σ é substituída por $\widehat{\Sigma}$, só precisamos pensar a respeito de uma priori para α (ou A). Da priori conjugada natural, temos[53]:

$$\alpha|\widehat{\Sigma} \sim \mathcal{N}(\underline{\alpha}_{\text{MN}}, \underline{V}_{\text{MN}}). \tag{250}$$

A escolha de $\underline{\alpha}_{\text{MN}}$ (média da distribuição a priori) é baseada na observação empírica de que séries macroeconômicas são bem caracterizadas por passeios aleatórios (Nelson and Plosser, 1982), ou seja:

$$Y_t = Y_{t-1} + \varepsilon_t \tag{251}$$

A equação (251) implica que $\alpha_0 = 0, \beta_1 = \mathbb{I}_m, \beta_2 = \beta_3 = \ldots = \beta_p = 0_M$ em (220). Colocando esses valores em forma matricial, obtemos

$$\underline{A}_{M\times N} = (0_{M\times 1}, \mathbb{I}_M, 0_{M\times M}, \ldots, 0_{M\times M})$$

e

$$\underline{\alpha}_{\text{MN}} = \text{vec}(\underline{A}_{\text{MN}}). \tag{252}$$

Dessa forma, conseguimos exprimir nossa crença de que as séries macroeconômicas incluídas no VAR seguem passeios aleatórios[54]. Note que se $\underline{\alpha}_{MN}$ tiver 105 elementos, a matriz $\underline{V}_{MN}$ terá tamanho 105×105, o que gera o problema de determinar seus elementos. A matriz $\underline{V}_{MN}$ em (250) precisa então descrever o grau de crença em relação à média da f.d.p. a priori estabelecida em (252).

Baseado na ideia de que séries macroeconômicas seguem passeios aleatórios e de que modelos parcimoniosos conseguem fazer boas previsões, a priori de Minnesota sugere encolhimento mais forte na direção de 0 para coeficientes de defasagem mais longas, o que reflete também a ideia de que observações mais distantes têm menor influência. Ainda com base no passeio aleatório, defasagens da variável j na equação da variável i também sofrem influência mais forte na direção de 0. Além disso, a priori de Minnesota usa $\underline{V}_{\text{MN}}$ diagonal, de forma que as covariâncias dos elementos de α são 0.

Definindo $\underline{V}_i$ como o bloco de $\underline{V}_{\text{MN}}$ relativo à equação i e $\underline{V}_{i_{jj}}$ como seus elementos diagonais, temos:

[53] O subscrito *MN* indica *Minnesota*.

[54] Observe que é preciso modificar essas crenças se incluirmos variáveis em 1^a diferença no VAR, nesse caso $Y_t = \varepsilon_t$ é um ruído branco e então a matriz A só terá valores iguais a 0.

$$\underline{V}_{i_{jj}} = \begin{cases} \frac{a_1}{r^2} & \text{para coeficientes da própria variável e } r = 1, \ldots, p \text{ depende do coeficiente ser relativo à defasagem de ordem } r \\ \frac{a_2}{r^2}\frac{\sigma_{ii}}{\sigma_{jj}} & \text{para coeficientes de defasagem } r \text{ da variável } j \text{ na equação de } i \\ a_3\sigma_{ii} & \text{para coeficientes de variáveis exógenas.} \end{cases} \tag{253}$$

Sendo que:

- σ_{ii}^2 foram estimados por Doan et al. (1984) e Litterman (1986) através de $AR(1)$ para cada equação individual do VAR.
- a_1, a_2 e a_3 são hiperparâmetros da priori de Minnesota.

Note que a_1 é comum para todas as variáveis. Então para a primeira defasagem, a variância será a_1. Para a segunda, vai ser $a_1/4$, isso significa que a variância à priori é menor - nossa certeza sobre a média é maior. Como as defasagens mais longas tem média à priori igual a zero, estamos colocando mais certeza nesse zero.

Estimar σ_{ii}^2 usando os dados significa que estamos usando informação dos dados para obter informação a priori, o que é controverso na a abordagem bayesiana, mas vem ganhando adeptos com a difusão do "método de Bayes empírico"[55]. Note que, através de r^2 no denominador dos elementos de $\underline{V}_{i_{jj}}$, faz-se com que a certeza em relação a a priori seja reforçada à medida que as defasagens aumentam. Em outras palavras, quanto maior for a defasagem referente ao coeficiente que estamos analisando, maior é a certeza a priori de que este coeficiente é 0.

Os hiperparâmetros a_1, a_2 e a_3 são escolhidos, normalmente, de forma a colocar menos importância nas defasagens de outras variáveis e muito pouca importância para variáveis exógenas. Note que transformamos um problema que tinha aproximadamente 10.000 parâmetros para um com 105 (pelo fato de $\underline{V}$ ser diagonal) e então passamos para apenas 8 e por fim para apenas a_1, a_2 e a_3, para o caso que usamos de exemplo inicialmente.

A grande vantagem da priori de Minnesota é a possibilidade de gerar inferência a posteriori em fórmula analítica e permitir maior flexibilidade para definir o conhecimento a priori. Usando a priori de Minnesota é possível mostrar que a f.d.p. a posteriori tem distribuição Normal dada por:

$$\alpha | Y, \hat{\Sigma} \sim \mathcal{N}(\bar{\alpha}_{MN}, \bar{V}_{MN}) \tag{254}$$

sendo que

$$\bar{V}_{MN} = \left[\underline{V}_{MN}^{-1} + \underbrace{\left(\hat{\Sigma}^{-1} \otimes (X'X) \right)}_{\text{precisão da verosimilhança}} \right]^{-1}$$

e

$$\bar{\alpha}_{MN} = \bar{V}_{MN} \left[\underline{V}_{MN}^{-1} \underline{\alpha}_{MN} + \left(\hat{\Sigma}^{-1} \otimes X \right)' Y \right]$$

Entretanto, esta abordagem não é completamente bayesiana, pois Σ é desconhecida e a estratégia de fazer $\Sigma = \hat{\Sigma}$ ignora completamente esta fonte de incerteza.

[55] *Empirical Bayes*, em inglês.

7 Parte 7 - Modelos em Espaço de Estados

Modelos em espaço de estados nos permitem tratar diversos problemas da análise de séries temporais de forma mais flexível e unificada, pois, usando essa abordagem, podemos escrever muitos modelos (AR, VAR, DSGE) e dar um mesmo tipo de tratamento a eles. Em modelos de espaço de estados, a dinâmica, a dinâmica é determinada por uma série de variáveis aleatórias não observáveis, $\alpha_1, \alpha_2, \ldots, \alpha_T$, chamados de estados. A relação entre α_t e Y_t é especificada então pelo modelo em espaço de estados.

O principal objetivo da análise de espaço de estados é utilizar as informações trazidas por $\{y_t\}_{t=1}^T$ para realizar inferência a respeito das variáveis não-observáveis (os estados), $\{\alpha_t\}_{t=1}^T$. Além disso, previsões, extração de sinal e estimação de parâmetros também são possíveis.

7.1 O modelo de nível local

O modelo de nível local é um modelo univariado em estado de espaço simples que nos permitirá uma compreensão melhor desta classe geral de modelos na forma de espaço de estados. Mais especificamente, este modelo é dado por:

$$y_t = \alpha_t + \varepsilon_t, \tag{255}$$

em que cada elemento de (255) é um escalar, sendo $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, h^{-1})$ (essa equação é de um modelo simples que varia no tempo). As variáveis não observáveis α_t seguem um passeio aleatório:

$$\alpha_{t+1} = \alpha_t + u_t \tag{256}$$

onde $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \eta h^{-1})$ e ε_t e u_s são independentes para todo t e s. Note que a variância de u_t é proporcional à variância de ε_t, mas isso não é uma necessidade do sistema, apenas simplifica a nossa vida na hora de fazer estimação. Ainda com relação a u_t, o importante é que esse choque aconteça no fim de y_t e antes de y_{t+1}, isto é, nós temos α_t e temos u_t, chegamos em uma priori para α_{t+1} e só então vamos fazer a inferência de y_{t+1}. Por exemplo, considere o mercado acionário: a bolsa fecha à tarde e volta a abrir no outro dia de manhã. Tanto faz se o choque ocorreu às 20h ou às 8h. No fundo, nossa análise depende apenas de como definimos que vai ser o nosso tempo: podemos considerar que o período t vai até imediatamente antes de ocorrer $t + 1$ ou podemos dizer que t encerra no momento que y_t é observado e a partir disso temos $t + 1$.

Será necessário também definir uma priori para α_1, pois nós começamos com $y_1 = \alpha_1 + \varepsilon_1$ e desconhecemos quem é $\alpha_1 = \alpha_0 + u_0$. Neste caso, dizemos que α_1 é a condição inicial desse processo. A equação (255) é conhecida como equação de medida, enquanto que a equação (256) é conhecida como equação de transição dos estados. Note que se $\eta = 0$, então $\alpha_{t+1} = \alpha_t \; \forall \; t$ e:

$$y_t = \alpha + \varepsilon_t,$$

implicando que que y_t flutua ao redor de uma média fixa (constante) α. Essa situação é bastante similar com um modelo de regressão linear que não tem variáveis explicativas, isto é, um modelo de regressão que tem apenas a constante. Porém, se $\eta \neq 0$, y_t será composta por uma tendência estocástica definida por α_t e por um componente idiossincrático[56] ε_t que gerará flutuações ao redor de α_t.

[56] Idiossincrático é uma expressão usada para fazer referência, muitas vezes, a situações incomuns ou até impróprias. A palavra idiossincrático não é encontrada no dicionário da língua portuguesa, porém, é a forma mais usual da palavra idiossincrásico (Fonte: https://www.significados.com.br/idiossincratico, último acesso em 16/06/2018).

Substituindo (256) em (256) de maneira iterativa obtemos:

$$\alpha_t = \alpha_1 + \sum_{j=1}^{t-1} u_j. \tag{257}$$

Podemos ver como α_t define uma tendência, pois acumula todos os choques u_j para $j = 1, \ldots, t-1$. Logo, tomando α_1 como dado[57], temos:

$$Var[\alpha_t] = (t-1)\eta h^{-1}. \tag{258}$$

Portanto, a variância de α_t muda gradualmente de acordo com t (i.e. o processo é não estacionário), o que é consistente com a ideia de uma tendência[58]. A equação (255) decompõe a variável aleatória observável y_t em um componente de tendência, α_t, e um componente idiossincrático, ε_t. Observe que o η pode ser usado em um teste de raiz unitária: se ele for zero, temos um processo estacionário, se ele for diferente de zero então temos um processo que é não estacionário.

Em geral, modelos de espaço de estados decompõe as séries em observáveis em várias partes como por exemplo, tendência, ciclos, componentes sazonais, erro, etc. O modelo de nível local é usado para medir a importância da tendência em relação ao componente idiossincrático. Por isso as variâncias de u_t e ε_t foram definidas de forma proporcional. Basicamente queremos saber se a variância do choque é relevante em relação à variância da tendência. Se $\eta \to 0$, então a série é estacionária e não apresenta tendência. Se $\eta \neq 0$, a média de Y_t varia no tempo e este modelo pode ser visto como um modelo com parâmetros variando no tempo (a média varia ao longo do tempo). Portanto, se α_t são parâmetros, será importante definir uma f.d.p. a priori para eles. A equação (256) pode ser interpretada como uma priori hierárquica para o vetor $\alpha = (\alpha_1, \ldots, \alpha_T)$.

7.1.1 A priori hierárquica

A f.d.p. à priori hierárquica adiciona mais níveis ao modelo, permitindo a análise e inferência de modelos flexíveis e altamente parametrizados. Note que (255) possui um parâmetro α_t para cada observação y_t. Portanto, temos T observações e $T + 1$ parâmetros (lembre-se que precisamos estimar h). Porém, a priori para (256), apesar de permitir que todos os α_t sejam diferentes uns dos outros, impõe uma estrutura a eles. Portanto, temos a priori para todos os $T - 1$ parâmetros α utilizando apenas um hiperparâmetro, η.

Impôr restrições como em (256) é fundamental para permitir a estimação precisa de modelos altamente parametrizados. De forma geral, modelos hierárquicos adicionam mais camadas ao dar a priori para os hiperparâmetros e essa priori dependerá de outros hiperparâmetros. Por exemplo, no caso geral, temos:

$$y \sim p(y|\theta) \qquad \text{(f. verossimilhança)}, \tag{259}$$

$$\theta \sim p(\theta|\alpha_0) \quad \text{e} \tag{260}$$

$$\alpha_0 \sim p(\alpha_0, \alpha_{00}), \tag{261}$$

[57] Isso significa que ele ainda não tem uma priori, pois se tivesse ela teria uma variância.
[58] Note que isso deixa explícito que os α_t são variáveis aleatórias diferentes entre si, uma vez que sua variância não é a mesma.

em que α_{00} é especificado pelo pesquisador. A fdp $\alpha_0 \sim p(\alpha_0, \alpha_{00})$ seria o equivalente à nossa equação de transição. Escrevendo em termos da notação que vínhamos utilizando, teremos:

$$y \sim p(y|\underline{\alpha}, h) \tag{262}$$

$$\alpha \sim p(\alpha|\mu_\alpha, \sigma) \quad \text{e} \tag{263}$$

$$\mu_\alpha, \sigma \sim p(\mu_\alpha, \sigma|\eta). \tag{264}$$

7.1.2 A função de verossimilhança e a priori

Usando a notação matricial, temos:

$$Y_{T\times 1} = \mathbb{I}_{T\times T}\alpha_{T\times 1} + \varepsilon_{T\times 1} \tag{265}$$

Observe que a expressão (265) tem a mesma forma funcional do modelo de regressão linear com T parâmetros, com a matriz identidade fazendo o papel de X. Se $\varepsilon \sim \mathcal{N}(0_T, h^{-1}\mathbb{I}_T)$, então caímos no MNCRL para $X = \mathbb{I}_T$. Assim, já conhecemos o formato da função de verossimilhança. Perceba que nossa priori dos α é um passeio aleatório, de forma que a distribuição de u_t passa a reger o processo.

Podemos, por exemplo, usar uma priori conjugada natural para realizar inferência analítica. Para isso, é conveniente reescrever o modelo em 1^a diferença, usando a matriz D de tamanho $(T-1)\times T$:

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & -1 & 1. \end{bmatrix} \tag{266}$$

Logo,

$$D\alpha = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \vdots \\ \alpha_T - \alpha_{T-1} \end{bmatrix}$$

e a equação de transição dos estados (256) simplifica para

$$D\alpha = u$$

A hipótese $u \sim \mathcal{N}(0_{T-1}, h^{-1}\eta\mathbb{I}_{T-1\times T-1})$ define uma priori para $D\alpha|h$. Para completar o modelo, precisamos de uma priori para h e para a condição inicial α_1. Reescrevendo[59] (265), temos:

$$Y = W\theta + \varepsilon, \tag{267}$$

[59] Isso é feito para que a gente fique com um modelo só em termos dos u.

em que

$$\theta = \begin{bmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \\ \vdots \\ \alpha_T - \alpha_{T-1} \end{bmatrix}$$

e

$$W = \begin{bmatrix} 1 & 0'_{T-1} \\ \iota_{T-1} & C. \end{bmatrix}$$

ι_T é um vetor T de variáveis aleatórias, 0_T é um vetor de zeros e C é uma matriz triangular inferior contendo apenas 1 na diagonal principal e abaixo dela (é a inversa de D). Assim, a priori conjugada natural é dada por:

$$\theta, h \sim \mathcal{NG}(\underline{\theta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}) \tag{268}$$

Os termos $\underline{\theta}$ e $\underline{V}$ precisam representar as características da equação de transição dos estados dada em (256):

$$\underline{\theta} = \begin{bmatrix} \underline{\theta} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{269}$$

e

$$\underline{V} = \begin{bmatrix} \underline{V}_{11} & 0'_{T-1} \\ 0_{T-1} & \eta I_{T-1}. \end{bmatrix} \tag{270}$$

Observe que $\underline{\theta}$ é a média da condição inicial e $\underline{V}_{11}$ é sua variância. Note ainda que os demais elementos de $\underline{\theta}$ são zero pois a média de $\alpha_t - \alpha_{t-1}$ é a média de u_t, que é zero. Ou seja, a priori implica que $(\alpha_{T+1} - \alpha_T) \sim \mathcal{N}(0, \eta h^{-1})$ e η é o único hiperparâmetro. Um econometrista clássico interpretaria (256) como parte da função de verossimilhança, o que mostra como muitas vezes há certo grau de arbitrariedade na definição da função de verossimilhança e da priori.

A expressão (267) em momento algum fala da equação de transição dos estados, isto é, não estamos falando de u_t diretamente. Isto é, a equação $Y = W\theta + \varepsilon$ diz respeito apenas à equação de medida, e é apenas um modelo de regressão com parâmetros variando no tempo e sem a equação de movimento. Quem diz que $\alpha_2 - \alpha_1 = u_1$ somos nós, que estamos impondo isso na priori (além disso dizemos que este u_t não tem média). Isso implica que o $\underline{\theta}$ terá todas as médias, exceto a primeira, iguais a zero. A matriz $\underline{V}$ implica, no primeiro elemento, a variância do α_1 e no elemento 22 a variância dos u_t (na verdade essa não é bem a variância, pois ela ainda vai ter um h multiplicado ali de alguma forma). Note que se a equação de transição dos estados fosse a de um modelo um auto-regressivo, teríamos $\mathbb{E}[\alpha_2 - \alpha_1] = \mu + (\rho - 1)\mathbb{E}[\alpha_1]$.

Por fim, note que aqui temos uma sutileza no tratamento clássico e bayesiano. Para a abordagem clássica, os α_t são chamados de estados (não podem ser considerados parâmetros pois não são fixos) e portanto não maximizamos a verossimilhança em relação a α_t. A verossimilhança clássica não tem maximização em relação a estados nenhum (o filtro de Kalman, que veremos adiante, integra os estados para fora da verossimilhança). Na hora que simulamos os dados, iremos simular apenas os u_t e acumular os choques, enquanto que na abordagem bayesiana iremos amostrar todos os parâmetros pois eles são variáveis aleatórias.

7.1.3 A função distribuição de probabilidade a posteriori

Resultados para o MNCRL podem ser usados para mostrar que a f.d.p. a posteriori do modelo de nível local é dado por $\theta, h \sim \mathcal{NG}(\bar{\theta}, \bar{V}, \bar{s}^2, \bar{\nu})$ com

$$\bar{\theta} = \bar{V}(\underline{V}^{-1}\underline{\theta} + W'Y), \tag{271}$$

$$\bar{V} = (\underline{V}^{-1} + W'W)^{-1}, \tag{272}$$

$$\bar{\nu} = \underline{\nu} + T, \tag{273}$$

e

$$\bar{\nu}\bar{s}^2 = \underline{\nu}\underline{s}^2 + (Y - W\bar{\theta})'(Y - W\bar{\theta}) + (\bar{\theta} - \underline{\theta})'\underline{V}^{-1}(\bar{\theta} - \underline{\theta}). \tag{274}$$

Propriedades da f.d.p. normal[60] nos dizem que se $p(\theta|h, y)$ tem distribuição normal, então $p(\alpha|h, y)$ pertencerá à mesma família de distribuição. Com isso, sabemos que $(\alpha, h) \sim \mathcal{NG}(\bar{\alpha}, \bar{V}_\alpha, \bar{s}^{-2}, \bar{\nu})$ com

$$\bar{\alpha} = W\bar{\theta} \tag{275}$$

e

$$\bar{V}_\alpha = W\bar{V}W'. \tag{276}$$

Demonstração. A demonstração é sugerida como exercício. □

A interpretação do modelo de nível local como um MNCRL mostra que, apesar de termos o mesmo número de coeficientes e observações, a informação a priori nos permite realizar inferência mesmo em modelos altamente parametrizados.

7.2 Método de Bayes Empírico

Até agora vínhamos escolhendo os hiperparâmetros do modelo nos baseando na informação a priori ou usando a priori não informativa ou usando uma priori imprópria. Ambas alternativas tem problemas: a priori própria, por seu caráter subjetivo da escolha de hiperparâmetros, pode ser criticada por outros pesquisadores que tem diferentes prioris. Já a priori imprópria inviabiliza a comparação de modelos uma vez que a verossimilhança marginal resultante pode não estar definida.

Com isso, algumas pessoas usam o chamado "método de Bayes empírico"[61], de forma a tentar contornar os dois problemas acima citados. O método de Bayes empírico estima os hiperparâmetros usando os dados, ao invés de fazer a sua escolha de maneira subjetiva ou tornando-os não informativos. É importante notar que este método pode ser alvo de críticas por utilizar duas vezes os dados: uma na escolha dos hiperparâmetros e outra no cálculo da posteriori. O processo de estimação dos hiperparâmetros utiliza como ferramenta mais comum a verossimilhança marginal, onde o valor que a maximiza será utilizado como estimativa a priori do(s) hiperparâmetro(s) a serem utilizados no método de Bayes empírico.

[60] Combinações lineares de distribuições normais também seguem distribuição normal.
[61] *Empirical Bayes methods.*

O resultado já visto para o MNCRL implica que a verossimilhança marginal do modelo de nível local é dado por:

$$p(y|\eta) = c\left(\frac{|\bar{V}|}{|\underline{V}|}\right)^{\frac{1}{2}}\left(\bar{\nu}\bar{s}^2\right)^{-\frac{\bar{\nu}}{2}}, \tag{277}$$

com

$$c = \frac{\Gamma\left(\frac{\bar{\nu}}{2}\right)\left(\underline{\nu}\underline{s}^2\right)^{-\frac{\underline{\nu}}{2}}}{\Gamma\left(\frac{\underline{\nu}}{2}\right).\pi^{\frac{T}{2}}} \tag{278}$$

A notação de (277) deixa clara a dependência em relação a y. O valor de $\hat{\eta}$ que maximiza $p(y|\eta)$ é normalmente o valor a ser utilizado como priori do método de Bayes empírico. Porém, um método mais formal de realizar inferência a respeito de η é usar as leis de probabilidade e fazer inferência bayesiana tradicional usando a verossimilhança marginal, isto é:

$$p(\eta|y) \propto p(y|\eta)\cdot p(\eta) = c\cdot\left(\frac{|\bar{V}|}{|\underline{V}|}\right)^{\frac{1}{2}}\left(\bar{\nu}\bar{s}^2\right)^{-\frac{\bar{\nu}}{2}}\cdot p(\eta) \tag{279}$$

e integração de Monte Carlo pode ser usada para aproximar a posteriori $p(\eta|y)$.

7.3 O modelo geral de espaço de estados linear

Um modelo geral de espaço de estados pode ser escrito de forma mais geral como[62]:

$$y_t = W_t\cdot\delta + Z_t\cdot\beta_t + \varepsilon_t \tag{280}$$

e

$$\beta_{t+1} = \Pi_t\cdot\beta_t + u_t, \tag{281}$$

em que y_t é um vetor $M\times 1$ de observações, $\varepsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma)$ é um vetor $M\times 1$ de erros, W_t é uma matriz $M\times p_0$ conhecida, δ é um vetor $p_0\times 1$ de parâmetros, Z_t é uma matriz $M\times k$ conhecida e β_t é um vetor de parâmetros (estados) que evolui de acordo com (281) e Π_t é uma matriz de parâmetros. Adicionalmente, faremos a hipótese de que o vetor $k\times 1$ $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,Q)$ é independente de $\varepsilon_t\ \forall\ s,t$. O termo $Z_t\cdot B_t$ inclui no modelo uma dinâmica que permite flexibilidade maior ao modelo.

Observe que a equação (281) indica que o B_{t+1} segue um processo $AR(1)$. Se $\delta = 0$ e $Z_t = 1$, caímos no modelo de nível local. Também podemos colocar as matrizes Σ e Q variando no tempo. Se não houver o termo $Z_t\cdot\beta_t$ temos o modelo de regressão com matriz de variância e covariância geral. O que complica é que temos esse termo e que os parâmetros dependem um dos outros ao longo do tempo, no sentido de que se não tivermos β_t não podemos amostrar β_{t+1}.

Uma das grandes vantagens da abordagem bayesiana é o seu caráter modular visto, por exemplo, no amostrador de Gibbs. Com isso, modelos complicados podem ser estimados combinando resultados mais simples. Modelos em espaço de estados são um bom exemplo de como explorar essa característica modular. Portanto, ao invés de derivar a função de verossimilhança, descrever a priori e então encontrar a função densidade de probabilidade a posteriori, iremos direto para os métodos bayesianos de estimação usando os resultados que já vimos.

A principal complicação em (280)-(281) é o fato de β_t e β_{t+1} serem dependentes, o que impossibilita a simulação deles de forma individual e exige a amostragem de uma distribuição normal T dimensional. Note como essa situação é diferente de

[62]Esta é a forma linear, o caso não linear é um pouco mais complicado.

$Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, onde β_1 e β_2 são totalmente independentes. A estrutura autoregressiva imposta nos parâmetros do modelo de (eq de transição) nos dá uma lei de movimento que irá permitir estimar um parâmetro para cada período do tempo (se eles fossem independentes não teríamos graus de liberdade suficientes). A distribuição normal T dimensional na verdade pode ser $k \times T$ dimensional. Note que nós conhecemos quem é Z_t. Estamos interessados em amostrar da distribuição

$$p(\beta_1, \beta_2, \ldots, \beta_T | Y, \delta, \Sigma, Q).$$

que pode ser escrita da seguinte maneira

$$p(B_T | \beta_{T-1}, Y, \delta, \Sigma, Q) \cdot p(B_{T-1} | \beta_{T-2}, Y, \delta, \Sigma, Q) \ldots .$$

Note que mesmo escrevendo dessa forma, ainda temos a dependência dos β_T com relação ao anterior. Porém, Carter and Kohn (1994), Frühwirth-Schnatter (1994), DeJong and Shephard (1995) e Durbin and Koopman (2002) desenvolveram métodos eficientes para amostrar dessa Normal T dimensional usando algoritmos bayesianos no filtro de Kalman. Portanto, com o auxílio de um desses algoritmos será possível construir um amostrador de Gibbs usando resultados já vistos.

Supondo que $\beta_t, t = 1, \ldots, T$ são conhecidos, podemos reescrever a equação de medida $Y_t^* = W_t \delta + \varepsilon_t$, como $Y_t^* = y_t^* - Z_t \beta_t$. Note que (280) tem a mesma estrutura de um modelo VAR ou SUR. Isso sugere que podemos usar um algoritmo de Gibbs para amostrar de $p(\delta | y^T, \Sigma, \beta^T)$ e de $p(\Sigma^{-1} | y^T, \delta, \beta^T)$ usando a piori NOrmal-Wishart independentes ou até mesmo usando uma priori conjugada natural. Essa abordagem é conhecida como "aumento do dos dados"(data augmentation) pois estamos "aumentando"y_t com β_t.

Usando a priori normal para $\delta | \Sigma^{-1}, Q^{-1}, \beta^T$ e Wishart para $\Sigma^{-1} | \delta, \beta^T, Q^{-1}$, temos

$$p(\delta, \Sigma^{-1}, Q^{-1}, \beta_1, \beta_2, \ldots, \beta_T) = p(\delta) p(\Sigma^{-1}) p(Q^{-1} p(\beta, \ldots, \beta_T | Q^{-1})$$

em que

$$p(\delta) = \mathcal{N}(\underline{\delta}, \underline{V}) \tag{282}$$
$$p(\Sigma^{-1}) = \mathcal{W}_M(\underline{S}^{-1}, \underline{\nu}) \tag{283}$$
$$p(Q^{-1}) = \mathcal{W}_k(\underline{Q}^{-1}, \underline{\nu}_q) \tag{284}$$

e, como visto, a equação de transição de estados 281 pode ser vista como a priori para β

$$p(\beta_1, \beta_2, \beta_3, \ldots, \beta_T \mid Q^{-1}) = \prod_{t=1}^{T} p(\beta_{t+1} | \beta_t, Q) \qquad \text{com} \qquad p(\beta_1 \mid \beta_0, Q) = \mathcal{N}(0, Q). \tag{285}$$

Logo, a nossa priori para os β's é uma priori hierárquica, pois o parâmetro depende de um hiperparâmetro que tem uma priori. Sendo assim, é possível construir um algoritmo MCMC que amostra sequencialmente de $p(\delta | Y^T, \Sigma, \beta^T)$, $p(\Sigma^{-1} | Y^T, \delta, \beta^T)$, $p(Q^{-1} | Y^T, \delta, \beta^T)$ e de $p(\beta_1, \beta_2, \beta_3, \ldots, \beta_T | Y^T, \delta, \Sigma, Q)$. As três primeiras posterioris condicionais[63] são baseadas nos resultados para *VAR*:

[63] É condicional TOTAL porém não colocamos o Q. Note que se nós conhecemos β, então para δ e Σ não importa o valor de Q. Isto é, dado que sabemos β, então sabemos "implicitamente"o valor de Q.

$$\delta|y^T, \Sigma, \beta^T \sim \mathcal{N}(\bar{\delta}, \bar{V})$$

com

$$\bar{V} = \left(\underline{V}^{-1} + \sum_{t=1}^{T} W_t' \Sigma^{-1} W_t \right)^{-1}$$

e

$$\bar{\delta} = \bar{V} \left(\underline{V}^{-1} \underline{\delta} + \sum_{t=1}^{T} W_t' \Sigma^{-1} \left(y_t - Z_t \beta_t \right) \right).$$

Além disso,

$$\Sigma^{-1}|y^T, \delta, \beta^T \sim \mathcal{W}(\bar{S}^{-1}, \bar{\nu})$$

com

$$\bar{\nu} = T + \underline{\nu},$$

$$\bar{S} = \underline{S} + \sum_{t=1}^{T} (y_t - W_t \delta - Z_t \beta_t)(y_t - W_t \delta - Z_t \beta_t)',$$

e

$$Q^{-1}|y^T, \delta, \beta^T \sim \mathcal{W}(\bar{Q}^{-1}, \bar{\nu}_q)$$

com

$$\bar{\nu}_q = T + \underline{\nu}_q e \bar{Q} = \underline{Q} + \sum_{t=1}^{T} (\beta_{t+1} - \Pi_t \beta_t)(\beta_{t+1} - \Pi_t \beta_t)'.$$

Para completar o algoritmo, precisamos amostrar de $p(\beta_1, \beta_2, \beta_3, \ldots, \beta_T | Y^T, \delta, \Sigma, Q)$. Esse tipo de algoritmo é baseado no filtro de Kalman e no suavizador de Kalman. Intuiticamente, o algoritmo utiliza o filtro de Kalman para calcular

$$\prod_{t=1}^{T} p(\beta_{t+1} | Y^T, \delta, \Sigma, Q),$$

que são as densidades filtradas de $\beta_t \; \forall \; t$. Note que em T o filtro de Kalman nos dá

$$p(\beta_T | Y^T, \delta, \Sigma, Q), \tag{286}$$

que é a distribuição a posteriori para β_T, enquanto que para β_{T-1} temos apenas

$$p(\beta_{T-1}|Y^{T-1}, \delta, \Sigma, Q),$$

que **não** é a posteriori para β_{T-1} pois não leva em consideração a última observação Y_T.

Portanto, é preciso construir as posterioris para β_{t-1}, para $t = 2, \ldots, T$. O que vamos ver não é como conseguir a densidade, mas sim como amostrar dela. O suavizador de Kalman nos ajuda nesta tarefa. Mais especificamente, é preciso construir

$$p(\beta_t|Y^T, \beta_{t+1}, \Sigma, Q, \delta). \tag{287}$$

Note que com o conhecimento de (286) e (287), podemos amostrar $\beta_T^{(i)}$ de (286) e usar esta realização em (287) para amostrar $\beta_{T-1}^{(i)}$ e assim por diante, o que nos dá uma amostra $\{\beta_t^{(i)}\}_{t=1}^T$ da posteriori $p(\beta_1, \beta_2, \ldots, \beta_T|Y^T, \delta, \Sigma, Q)$.

Esse procedimento às vezes é chamado de *forward filter - backward sampler* (ou *smoother*). Essa ideia se baseia na equação de transição, que podemos escrever como $\beta_t = \beta_{t+1} - \varepsilon_t$. Ou seja, o Gibbs amostra sequencialmente de $p(\delta|Y^T, \Sigma, \beta^T)$, $p(\Sigma^{-1}|Y^T, \delta, \beta^T)$, $p(Q^{-1}|Y^T, \delta, \beta^T)$ e de $p(\beta_1, \beta_2, \beta_3, \ldots, \beta_T|Y^T, \delta, \Sigma, Q)$. Com isso, a inferência a posteriori baseada na amostra de Gibbs após convergência é feita exatamente como fizemos anteriormente. Esse algoritmo está descrito no capítulo 8 de Koop (2003).

Recapitulando, o modelo de espaço de estados gaussiano e linear tem 3 partes.

O modelo de espaço de estados Gaussiano e linear tem 3 partes:

- **Equação de transição ou de estado**: por causa do choque gaussiano ζ_t, ela tem distribuição normal com média condicional dada por $T_t\alpha_t$ e variância condicional dada por $R_t \cdot Q_t$. Essa notação mais geral permite que os valores de α variem com o tempo, além de permitir generalizar mais ainda o modelo, fazendo com que a matriz de covariâncias também varie, dando mais flexibilidade para as covariâncias dos estados. Por exemplo, é possível que o k ésimo elemento de alpha dependa do primeiro mais o segundo elemento de zeta, então temos uma matriz de variância e covariância singular. Podemos escrever a equação de transição dos estados como

$$\alpha_{t+1} = T_t\alpha_t + R_t\zeta_t, \qquad \zeta_t \overset{\text{i.i.d.}}{\sim} (0, Q_t). \tag{288}$$

- **Equação de medida ou de observação:** para esse modelo se manter gaussiano e linear, todo mundo, exceto os termos de erro, pode até variar no tempo, contanto que seja conhecido (não podem ser estocásticas). Por exemplo, vimos que Z_t em um VAR variando no tempo pode ser a matriz com as variáveis independentes X e neste caso a cada tempo teremos um valor diferente, mas mesmo assim é algo observável. A equação de medida pode ser escrita como

$$y_t = Z_t\alpha_t + \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} (0, H_t). \tag{289}$$

- **Distribuição inicial dos estados:** A distribuição inicial dos estados é como uma priori inicial, que combinada com a equação de transição pode ser usada para criar as prioris a cada instante de tempo. A distribuição inicial dos estados pode ser expressa da seguinte forma:

$$\alpha_1 \sim \mathbb{N}(a_1, P_1). \tag{290}$$

Note que

- ζ_t e ε_s são independentes para todo t, s, e independente de α_1;

- as observações y_t podem ser multivariadas;
- o vetor de estados α_t é não observável;
- $\theta = \{T_t, Z_t, R_t, Q_t, H_t\}$ determinam a estrutura do modelo.

Já vimos que podemos escrever vários modelos usando essa notação, como por exemplo VAR, modelo de nível local, DSGE, etc. Mas por que utilizar um modelo linear e gaussiano? Isso nos permite utilizar as propriedades da distribuição normal.

Observe que a equação de transição dos estados (288) define um VAR(1) e que as matrizes do sistema contém os parâmetros desconhecidos. A estimação clássica envolve dois passos: estimação de parâmetros desconhecidos e medir os estados (previsão, filtragem e suavização). Isso ignora completamente a incerteza com relação aos parâmetros, pois ao obter uma estimativa pontual (por exemplo, via estimador de MV), não incorporamos o fato disso ser uma estimativa. Iremos tratar, a partir das estimativas, essas quantidades como fixas para medir os estados não observáveis. Em outras palavras, iremos fazer a segunda etapa supondo que as estimativas são o valor populacional real. Por outro lado, a abordagem bayesiana permite fazer isso de maneira conjunta, então podemos incorporar a incerteza sobre os parâmetros ao mesmo tempo que tratamos dos estados.

7.3.1 O problema de filtragem

A filtragem busca obter as distribuições a posteriori, a cada período, usando as observações disponíveis até aquele momento. O objetivo da filtragem é obter as distribuições a posteriori $p(\alpha_t|y_t)\ \forall\ t$ em tempo real usando as observações y_t. Para $t = 1$ regra de Bayes, $P(A|B) = P(B|A)P(A)/P(B)$, nos dá:

$$p(\alpha_1|y_1) = \frac{p(y_1|\alpha_1)p(\alpha_1)}{p(y_1)}. \tag{291}$$

Temos na equação (293) as informações dos dados (verossimilhança) junto com a informação inicial dos estados dividido pela constante de integração. Já sabemos que se a priori é normal e a verossimilhança é normal, então temos uma posteriori conjugada natural, de forma que a posteriori será uma normal. A equação (289) e a equação (290) nos fornecem o numerador de (293). Nós queremos aprender algo sobre os estados, então temos em (289) o que os dados condicionais ao estado são e temos o palpite inicial (priori) pasra estados, de maneira que temos as duas peças chave para construir a posteriori.

Pensando no nosso modelo dado em (280)-(281), a equação equivalente à (293) seria $p(\beta_1|y_1) = \frac{p(y_1|\delta,\Sigma^{-1},Q)p(\beta_1)}{p(y_1)}$, Agora precisamos criar a mesma coisa para $\beta_2 \cdot p(\beta_2|y_2) = \frac{p(y_2|\beta_1,\delta,\Sigma^{-1},Q)\cdot p(\beta_2)}{p(y_2)}$. Só que nós não temos uma priori para β_2. Iremos então transformar a nossa posteriori do passo anterior $p(\beta_1|y_1)$ em uma priori $p(\beta_2|\beta_1, y_1)$. Mas como fazemos isso?

Olhando a equação (292), podemos calcular $p(\beta_2|y_1, Q) = \int p(\beta_2|\beta_1, Q)p(\beta_1|y_1)d\beta_1$, pois $p(\beta_2|\beta_1, Q)$ é justamente a equação de transição dos estados (e por isso ela é chamada de priori do modelo): $\beta_{t+1} = \beta_t + u_t, u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, Q)$. Já o termo $p(\beta_1|y_1)$ nós calculamos no passo anterior. Esse procedimento de calcular a integral de uma densidade multiplicada por uma densidade univariada em relação a outra variável aleatória é nosso procedimento conhecido de *marginalização*.

Mas o que significa $p(\beta_2|y_1)$? Nós temos a informação do período 1, representada por y_1, e estamos fazendo uma *previsão* para a variável aleatória do próximo período, β_2. Então a previsão da filtragem nos dá a priori. Quando chegamos no período 2 e temos disponível uma nova informação, y_2, nós iremos corrigir nossa previsão e obter $p(\beta_2|y_2)$. Reforçando, observe como sempre usamos a equação de transição para "ligar" as variáveis aleatórias (pois $\beta_i \neq \beta_j$, i.e., elas são variáveis aleatórias diferentes), ela dá a dinâmica deste processo de conseguir prioris a partir de previsões.

Note que $\beta_1|y_1$ deixa de ser uma posteriori no momento que recebemos uma nova informação.Essa distinção entre uma variável filtrada e suavizada é importante: a primeira incorpora todas as informações disponíveis até o período da variável aleatória, enquanto que o suavizador nos dá estimativas considerando todo o período dos dados. Isso implica que no último período (última observação da amostra) a informação filtrada e suavizada serão iguais (e somente no último período).

(289) e (290) nos dão $p(y_1|\alpha_1)$ e $p(\alpha_1)$, mas a aprendizagem em tempo real exige agora uma a priori para $\alpha_2|y_1$, isto é, como temos uma fórmula analítica, temos uma posteriori para $p(\alpha_1|y_1)$. Só que agora precisamos de uma priori para $\alpha_2|y_1$, sendo que α_2 é outra variável aleatória. Usamos então a equação de movimento (equação de transição dos estados) descrita em (288): ela é a informação de como os estados estão evoluindo. Enquanto que um clássico diria que isso entra na verossimilhança, o bayesiano vai olhar aquilo como uma priori. Colocando tudo junto, temos que

$$p(\alpha_2|y_1) = \int p(\alpha_2|\alpha_1)p(\alpha_1|y_1)\, d\alpha_1, \tag{292}$$

sendo $p(\alpha_2|\alpha_1)$ dada por (288). (292) com a regra de Bayes nos dá:

$$p(\alpha_2|y_2) = \frac{p(y_2|\alpha_2)p(\alpha_2|y_1)}{p(y_2|y_1)}. \tag{293}$$

Até aqui, o procedimento é genérico e chamado de *filtro bayesiano*.

Para construir uma priori para α_2 a partir do que já sabemos de $\alpha_1|y_1$, iremos usar a equação (292). y_1 é como se fosse uma amostra de treinamento para α_2. Então, continuando a aprendizagem a respeito de α_t à medida que novas informações y_t são obtidas e usando $Y_t = \{y_1, \ldots, y_t\}$, temos a fórmula geral para a densidade a posteriori dos estados:

$$p(\alpha_t|Y_t) = \frac{p(y_t|\alpha_t)p(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})} \propto p(y_t|\alpha_t)p(\alpha_t|Y_{t-1}), \tag{294}$$

e fórmula para obter a a priori para $t+1$ com a informação de t:

$$p(\alpha_{t+1}|Y_t) = \int p(\alpha_{t+1}|\alpha_t)p(\alpha_t|Y_t)\, d\alpha_t. \tag{295}$$

Note que de posse de $\alpha_2|y_1$, plugamos na equação da posteriori $\alpha_2|y_2$, junto com a verossimilhança da nova observação y_2. Utilizando essa mesma estrutura, podemos obter uma priori para α_3 e assim por diante, de forma que vamos aprendendo à medida que novas informações y_t são obtidas. É importante que o modelo seja linear para que a integral em (292) tenha fórmula analítica.

A constante de integração da a posteriori em (294) é dada por[64]:

$$p(y_t|Y_{t-1}) = \int p(y_t|\alpha_t)p(\alpha_t|Y_{t-1})d\alpha_t. \tag{296}$$

Na abordagem clássica, (296) nos dá a contribuição da observação t para a verossimilhança da amostra. A equação (296) nos dá uma função de verossimilhança em função apenas dos parâmetros removendo o efeito dos estados (ela é a verossimilhança do estatístico clássico, que não enxerga α_t como parâmetro). A contribuição de Kalman foi resolver (294), (295) e (296) em fórmula fechada quando o sistema é linear e gaussiano.

[64]Na prática nós tiramos o log e por isso aparece uma soma.

7.3.2 Filtro de Kalman

Como já sabemos, se a verossimilhança e a a priori em (294) são Gaussianas, a posteriori $p(\alpha_t|y_t)$ também será. Kalman (1960) mostra que, como o modelo é linear e Gaussiano, (292) pode ser calculada analiticamente e resultará em uma distribuição Normal. O filtro de Kalman nada mais é um algoritmo que resolve o problema de filtragem analíticamente para modelos lineares e gaussianos através do cálculo da média e variância das distribuições, pois elas determinam completamente a distribuição Normal.

O que o Kalman propôs foi utilizar a conjugada natural para o problema de filtragem. Com isso nós vamos ter os mesmos resultados já vistos no modelo de regressão linear normal com priori conjugada natural.A ideia é que só precisaremos calcular as médias e variâncias a cada instante para saber quem são $p(\beta_1|y_1), p(\beta_2|y_2), \ldots, p(\beta_T|y_T)$. O filtro de Kalman nos dá densidades, mas pelo fato delas sempre serem gaussianas, basta nós sabermos quem é a média e quem é a variância.

Modelo em espaço de estados: $\alpha_{t+1} = T_t\alpha_t + R_t\zeta_t, \quad y_t = Z_t\alpha_t + \varepsilon_t, \quad \alpha_1 \sim \mathcal{N}(a_1, P_1)$. Defina $a_{t+1} = \mathbb{E}(\alpha_{t+1}|Y_t), \quad P_{t+1} = \text{var}(\alpha_{t+1}|Y_t)$.

O erro de previsão é:

$$\begin{aligned} v_t &= y_t - \mathbb{E}(y_t|Y_{t-1}) \\ &= y_t - \mathbb{E}(Z_t\alpha_t + \varepsilon_t|Y_{t-1}) \\ &= y_t - Z_t\mathbb{E}(\alpha_t|Y_{t-1}) \\ &= y_t - Z_t a_t; \end{aligned}$$

Segue que $v_t = Z_t(\alpha_t - a_t) + \varepsilon_t$, logo $\mathbb{E}(v_t) = \mathbb{E}[\mathbb{E}(v_t|Y_{t-1})] = 0$. Sua variância é $F_t = \text{var}(v_t|Y_{t-1}) = Z_t P_t Z_t' + H_t$. A demonstração tradicional do filtro de Kalman usa um lema da teoria de regressão Normal multivariada.

Lema 7.3.1. *Suponha que x, y e z são vetores cuja distribuição conjunta é Normal com $\mathbb{E}(z) = 0$ e $Cov(y, z) = \Sigma_{yz} = 0$. Logo*

$$\begin{aligned} \mathbb{E}(x|y, z) &= \mathbb{E}(x|y) + \Sigma_{xz}\Sigma_{zz}^{-1}z, \\ var(x|y, z) &= var(x|y) - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}'. \end{aligned}$$

Como usamos a priori conjugada natural Gaussiana, podemos usar o lema para atualizar $p(\alpha_t|Y_{t-1})$, pois só precisamos de $\mathbb{E}(\alpha_t|Y_t)$ e $\text{var}(\alpha_t|Y_t)$ para definir $p(\alpha_t|Y_t)$. Para isso, usamos $x = \alpha_t$, $y = Y_{t-1}$ e $z = v_t$, pois $\mathbb{E}(v_t) = 0$ e $Cov(Y_{t-1}, v_t) = \mathbb{E}[y_j\mathbb{E}(v_t|Y_{t-1})'] = \mathbb{E}[y_j \cdot 0] = 0 \; \forall j \leq t-1$.

Note que $Y_t = \{Y_{t-1}, y_t\} = \{Y_{t-1}, v_t\}$. Além disso:

$$\mathbb{E}[\alpha_t v_t'|Y_{t-1}] = \mathbb{E}[\alpha_t(\overbrace{Z_t(\alpha_t - a_t) + \epsilon_t}^{v_t})'|Y_{t-1}] = \mathbb{E}[\alpha_t(\alpha_t - a_t)'Z_t'|Y_{t-1}] = \overbrace{P_t Z_t'}^{\Sigma_{\alpha_t v_t}}.$$

Seja $K_t = P_t Z_t' F_t^{-1}$ o ganho de Kalman, usando o Lema:

$$\begin{aligned}\mathbb{E}[\alpha_t|Y_t] &= \mathbb{E}(\alpha_t|Y_{t-1}, v_t) = \mathbb{E}(\alpha_t|Y_{t-1}) + \Sigma_{\alpha_t v_t}\Sigma_{v_t}^{-1} v_t \\ &= a_t + P_t Z_t' F_t^{-1} v_t \\ &= a_t + K_t v_t;\end{aligned}$$

O Lema nos dá também a variância de $p(\alpha_t|Y_t)$:

$$\begin{aligned}\text{var}(\alpha_t|Y_t) &= \text{var}(\alpha_t|Y_{t-1}, v_t) = \text{var}(\alpha_t|Y_{t-1}) - \Sigma_{\alpha_t v_t}\Sigma_{v_t}^{-1}\Sigma_{\alpha_t v_t}' \\ &= P_t - P_t Z_t' F_t^{-1} Z_t P_t' \\ &= P_t - K_t Z_t P_t';\end{aligned}$$

Usando $p(\alpha_t|Y_t)$ e $p(\alpha_{t+1}|\alpha_t)$ geramos uma a priori $p(\alpha_{t+1}|Y_t)$:

$$a_{t+1} = \mathbb{E}[\alpha_{t+1}|Y_t] = \mathbb{E}[T_t\alpha_t + R_t\zeta_t|Y_t] = T_t\mathbb{E}[\alpha_t|Y_t] = T_t a_t + T_t K_t v_t,$$

de forma similar para a variância:

$$\begin{aligned}P_{t+1} &= \text{var}(\alpha_{t+1}|Y_t) = \text{var}(\overbrace{T_t\alpha_t + R_t\zeta_t}^{\alpha_{t+1}}|Y_t), \\ &= \text{var}(T_t\alpha_t|Y_t) + \text{var}(R_t\zeta_t|Y_t) = T_t P_t T_t' - T_t K_t F_t K_t' T_t' + R_t Q_t R_t'.\end{aligned}$$

O *ganho de Kalman*

$$K_t = P_t Z_t' F_t^{-1}$$

é a matriz de peso ideal para a nova informação trazida por v_t. K_t é chamado de ganho de Kalman pois tínhamos uma previsão a_1 e após ter o erro v_1, podemos calcular o quanto do erro de previsão será incorporado na nossa informação a posteriori dos estados. Se a variância da equação de medida é muito grande (H) e a variância da equação de estados é pequena, temos que o ganho de Kalman vai dar pouca relevância ao erro de previsão (pois ele é afetado pelo erro que temos nas observações). Por outro lado se H é pequeno, nossa informação dos dados é muito rica e será incorporado na posteriori. Kalman consegue mostrar que isso é a quantidade ótima de cada um dos termos para este processo.

Se os α_t fossem constantes ao longo do tempo, isto é, $\alpha_t = \alpha \; \forall \; t$, então a posteriori de hoje é a priori de amanhã. Como isso não ocorre, os α' são variáveis aleatórias diferentes, mas note que elas se ligam através da equação de transição, de forma que podemos criar a priori para o novo estado utilizando a posteriori que era para o estado anterior. Em outras palavras, o problema de filtragem é diferente do problema de estimação recursiva: quando fazemos uso de uma posteriori como priori para acoplar com novas informações, estamos tratando de uma mesma variável aleatória. Mas isso não acontece no problema de filtragem, pois a cada instante de tempo nós temos uma nova v.a. a ser estimada, só que a posteriori no passo anterior é referente a outra v.a..

7.3.3 Suavizador de Kalman

O filtro de Kalman nos dá a sequência de densidades $p(\beta_1|y_1), p(\beta_2|y_2, y_1), p(\beta_3|y_3, y_2, y_1) \cdots p(\beta_T|\overbrace{y_T, y_{T-1}, \ldots, y_2, y_1}^{\underline{y}})$. No entanto, podemos querer calcular as densidades suavizada, isto é, queremos calcular alguma coisa utilizando a informação

de toda a amostra, $p(\beta_1|y_T, y_{T-1}, \ldots, y_1)$. Pense, por exemplo, que estamos interessados em saber se em 2010 houve uma recessão. Nós não sabemos isso diretamente (não existe uma variável mensurável para isso), mas nós podemos utilizar, por exemplo o PIB. Vai fazer diferença se nós utilizarmos a amostra apenas até 2010 ou utilizarmos até 2018 para tentar entender o comportamento do PIB.

O filtro calcula a média e a variância dado Y_t, enquanto o suavizador calcula a média e a variância dado todo o conjunto de observações $Y_T = \{y_t\}_{t=1}^T$. Note que o filtro já nos dá $p(\alpha_T|Y_t)$, mas para em $t = T - 1$ temos apenas $p(\alpha_{T-1}|Y_{T-1})$. Mais especificamente, o filtro nos dá os valores previstos com base até o momento da observação disponível. Então quando terminamos o algoritmo de filtragem, onde todos os períodos estão observados, somente a última previsão será feita baseada em toda a amostra. O que o suavizador irá fazer é fazer o processo inverso: tendo os dados completos, iremos "voltar" e atualizar as previsões que haviam sido feitas.

O objetivo do suavizador é melhorar as estimativas filtradas através da incorporação de informações de períodos futuros. Para melhorar $p(\alpha_{T-1}|Y_{T-1})$, precisamos apenas incorporar a informação adicional trazida por y_T, mas em $t = T - 2$, precisamos incorporar também y_{T-1} e assim por diante. Para isso, é possível explorar a propriedade Markoviana dos estados:

$$p(\alpha_t|\alpha_{t+1}, Y_T) = p(\alpha_t|\alpha_{t+1}, Y_T),$$

e atualizá-los de trás para a frente[65]. Isso vai nos permitir amostrar da densidade condicional total, pois $p(\beta_1, \ldots, \beta_T|\underline{y}, \delta, \Sigma^{-1}, Q^{-1}) = p(\beta_T|y_T, \delta, \Sigma^{-1}, Q^{-1}) \cdot (\beta_{T-1}|\beta_T, y_{T-1}, \delta, \Sigma^{-1}, Q^{-1}) \ldots p(\beta_1|\beta_2, y_1, \delta, \Sigma^{-1}, Q^{-1})$. Note que a cada densidade nós não precisamos mais de y_T inteiro.

Dado α_{t+1}, não precisamos de $y_{t+1}, \ldots, y_t$ para atualizar os estados filtrados:

$$\begin{aligned}
p(\alpha_t|\alpha_{t+1}, Y_T) &= p(\alpha_t|\alpha_{t+1}, Y_t), \quad \text{(independência condicional)} \\
&= \frac{p(\alpha_{t+1}|\overbrace{\alpha_t}^{\substack{\text{Se temos} \\ \alpha_t \text{ não} \\ \text{precisamos de} \\ \text{de } y_t}}, Y_t)p(\alpha_t|Y_t)}{p(\alpha_{t+1}|Y_t)}, \text{(regra de Bayes)} \\
&= \frac{\overbrace{p(\alpha_{t+1}|\alpha_t)}^{\substack{\text{Equação de} \\ \text{transição} \\ \text{dos estados}}} \overbrace{p(\alpha_t|Y_t)}^{\substack{\text{Densidade} \\ \text{filtrada}}}}{\underbrace{p(\alpha_{t+1}|Y_t)}_{\substack{\text{Densidade} \\ \text{prevista} \\ \text{dada pelo} \\ \text{filtro de Kalman}}}}, \text{(independ. cond.)}
\end{aligned} \tag{297}$$

Com isso, a conjunta para α_t e α_{t+1} fica:

$$\begin{aligned}
p(\alpha_t, \alpha_{t+1}|Y_T) &= p(\alpha_t|\alpha_{t+1}, Y_T)\overbrace{p(\alpha_{t+1}|Y_T)}^{\substack{\text{Densidade} \\ \text{suavizada} \\ \text{em } t+1}}, \text{(rel. conjunta e a condicional)} \\
&= \frac{p(\alpha_{t+1}|\alpha_t)p(\alpha_t|Y_t)p(\alpha_{t+1}|Y_T)}{p(\alpha_{t+1}|Y_t)}, \text{(usando 297)}
\end{aligned} \tag{298}$$

[65] Isso decorre da independência condicional: se soubermos α_{t+1}, não precisamos saber $y_{t+1}, \ldots, y_T$.

Usando (298), podemos escrever:

$$\begin{aligned} p(\alpha_t|Y_T) &= \int p(\alpha_t, \alpha_{t+1}|Y_T)d\alpha_{t+1} \\ &= \int \frac{p(\alpha_{t+1}|\alpha_t)p(\alpha_t|Y_t)p(\alpha_{t+1}|Y_T)}{p(\alpha_{t+1}|Y_t)}d\alpha_{t+1} \\ &= p(\alpha_t|Y_t)\int \frac{p(\alpha_{t+1}|\alpha_t)p(\alpha_{t+1}|Y_T)}{p(\alpha_{t+1}|Y_t)}d\alpha_{t+1} \end{aligned} \tag{299}$$

Se o modelo for linear e Gaussiano, a integral em (299) tem solução analítica. Nesse caso, podemos usar novamente o lema da teoria de regressão Normal multivariada para obter os resultados desejados. Note que a estrutura do modelo e o filtro de Kalman nos dão:

$$p(\alpha_{T-1}, \alpha_T|Y_{T-1}) = \overbrace{p(\alpha_T|\alpha_{T-1})}^{\text{eq. transição}}\overbrace{p(\alpha_{T-1}|Y_{T-1})}^{\text{densidade filtrada}}. \tag{300}$$

No caso clássico, estaremos interessados em calcular as densidades suavizadas (as marginais)

$$p(\beta_1|y_T)p(\beta_2|y_T)p(\beta_3|y_T)\dots p(\beta_T|y_T).$$

No nosso caso, estamos interessados em amostrar os betas de

$$p(\beta_1, \dots, \beta_T|y_T, Q, \Sigma^{-1}, \delta).$$

Nós não podemos amostrar esses betas de forma independente (usando as marginais) pois eles estão ligados pela equação de transição.

A distribuição conjunta (301) é Gaussiana e nos permite calcular $\mathbb{E}[\alpha_{T-1}|\alpha_T, Y_{T-1}]$ e var$[\alpha_{T-1}|\alpha_T, Y_{T-1}]$ usando o Lema, determinando completamente $p(\alpha_{T-1}|\alpha_T, Y_{T-1})$. Como vimos, a propriedade Markoviana garante que $p(\alpha_{T-1}|\alpha_T, Y_{T-1}) = p(\alpha_{T-1}|\alpha_T, Y_T)$, permitindo calcular:

$$p(\alpha_{T-1}, \alpha_T|Y_T) = p(\alpha_{T-1}|\alpha_T, Y_T)p(\alpha_T|Y_T). \tag{301}$$

Como (301) é Normal, é fácil calcular a marginal $p(\alpha_{T-1}|Y_T)$. Com $p(\alpha_{T-1}|Y_T)$, atualizamos $p(\alpha_{T-2}|Y_{T-2})$ usando:

$$p(\alpha_{T-2}, \alpha_{T-1}|Y_T) = p(\alpha_{T-2}|\alpha_{T-1}, Y_{T-2})p(\alpha_{T-1}|Y_T). \tag{302}$$

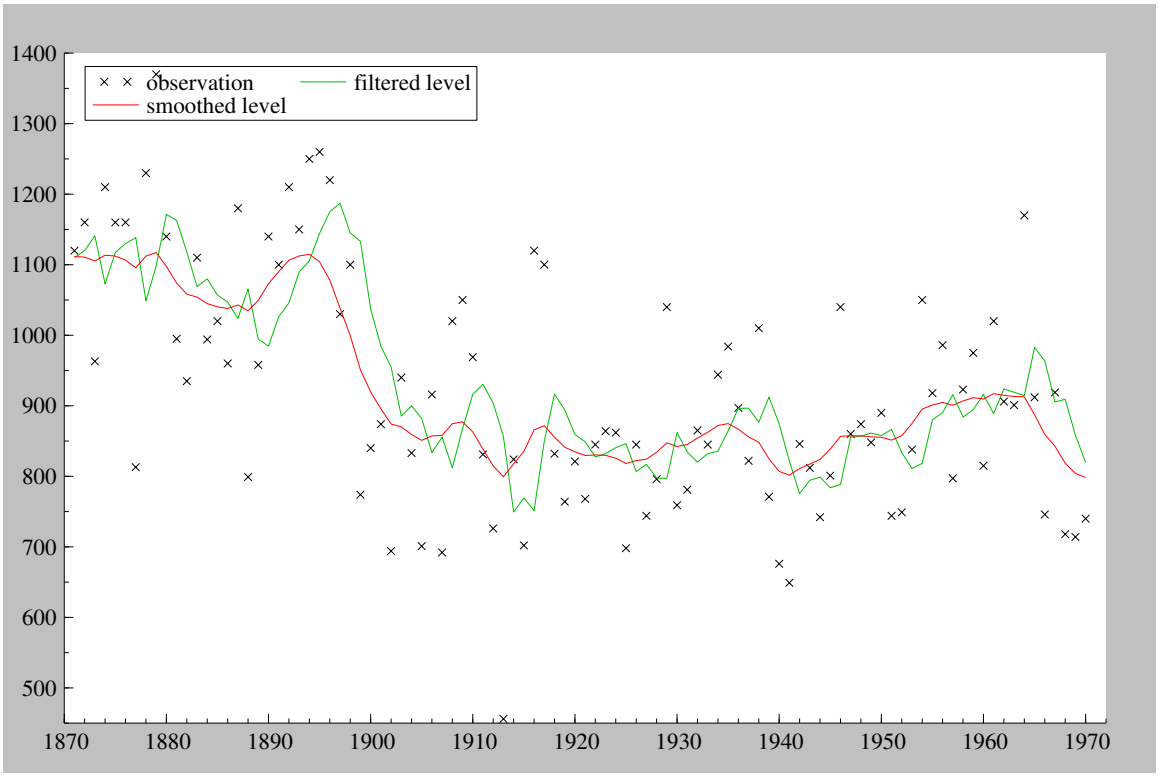


Figura 32: Filtro e Suavizador de Kalman no modelo de nível local (adaptado de Durbin and Koopman (2002)).

7.3.4 O Algoritmo de Carter e Kohn

O algoritmo de Carter and Kohn (1994) explora o suavizador para poder amostrar $\underline{\alpha}_T = \{\alpha_t\}_{t=1}^T$ da densidade a posteriori $p(\underline{\alpha}_T|Y_t)$. Note que:

$$\begin{aligned} p(\underline{\alpha}_T|Y_T) &= p(\alpha_T|Y_T)p(\underline{\alpha}_{T-1}|\alpha_T, Y_T), \\ &= p(\alpha_T|Y_T)p(\alpha_{T-1}|\alpha_T, Y_T)p(\underline{\alpha}_{T-2}|\alpha_T, \alpha_{T-1}, Y_T), \\ &= p(\alpha_T|Y_T)p(\alpha_{T-1}|\alpha_T, Y_{T-1})p(\underline{\alpha}_{T-2}|\alpha_{T-1}, Y_{T-1}), \\ &= p(\alpha_T|Y_T)\prod_{t=1}^{T-1} p(\alpha_t|\alpha_{t+1}, Y_t). \end{aligned} \tag{303}$$

O uso da barra embaixo em $\underline{\alpha}_T$ significa que estamos tratando de todos os α_t, $t = 1, \ldots, T$. A densidade $p(\alpha_T|Y_T)$ pode ser obtida pelo filtro de Kalman (é a própria densidade filtrada), enquanto o suavizador obtém $p(\alpha_t|\alpha_{t+1}, Y_t)$ recursivamente. Se neste último termo não tivesse α_{t+1}, nós teríamos a densidade filtrada. Porém α_{t+1} nos dá a informação do futuro para melhorar nossa estimativa.

Carter and Kohn (1994) sugerem então amostrar $\alpha_T^{(i)}$ de $p(\alpha_T|Y_T)$, $\alpha_{T-1}^{(i)}$ de $p(\alpha_{T-1}|\alpha_T^i, Y_t)$, $\alpha_{T-2}^{(i)}$ de $p(\alpha_{T-2}|\alpha_{T-1}^i, Y_t)$, e assim sucessivamente. A sequência $\underline{\alpha}_T^i = \{\alpha_t^i\}_{t=1}^T$ obtida dessa forma é uma realização da densidade a posteriori dos estados, $p(\underline{\alpha}_T|Y_T)$. Na prática, iremos amostrar δ de uma normal, Σ^{-1} e Q de uma wishart e então vamos fazer o filtro da Kalman e suavizador de Kalman. Usamos os betas obtidos para amostrar outro delta, σ, etc e assim por diante, no nosso amostrador de Gibbs.

Entretanto, quando R_t em (288) não tem posto completo, o algoritmo de Carter and Kohn (1994) não é válido, pois isso implicaria que a matriz de covariância de $p(\alpha_t|\alpha_{t+1}, Y_t)$ seria singular e não teríamos como amostrar dela. Nesses casos, é

preciso usar DeJong and Shephard (1995), pois este algoritmo amostra primeiro ζ_t^i de sua a posteriori e, com base base nessa realização, contrói uma realização a posteriori de α_t^i.

8 Parte 8 - Estimação Bayesiana de Modelos DSGE

Para entender de onde surgiram os modelos DSGE precisamos voltar um pouco no tempo. Até a década de 70 tínhamos os modelos derivados a partir do IS-LM, por exemplo, no modelo de Tinbergen. Esse modelo foi amplamente utilizado pois era uma ferramenta empírica para análise de dados econômicos. Só que com a crise do Petróleo os modelos deixaram de funcionar para explicar o que estava ocorrendo na economia.

Lucas criticou a forma como os modelos eram estimados: haviam modelos para equações agregadas, utilizando dados agregados, porém as estimativas estavam sendo feitas de forma independente, isto é, os parâmetros das equações afetavam a política econômica que por sua vez iria alterar as relações do modelo econométrico.

> "*Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.*" (Lucas, 1976)

Para Lucas, os modelos com parâmetros constantes no tempo, que não podem variar com as alterações de políticas econômicas, não eram adequados para modelar a economia. A partir disso, modelos para parâmetros que não são mutáveis frente às políticas econômicas como por exemplo, preferências e produção surgiram, levando a uma microfundamentação da macroeconomia. O trabalho de Hansen e Sargent incorporou essas mudanças aos modelos econométricos.

Então começaram a surgir modelos econométricos baseados em premissas teóricas, que amarravam mais as equações utilizadas frente a suposições que teriam que fazer mais sentido. A crítica de Lucas sugeria que para prever os resultados de políticas econômicas, é necessário modelar o comportamento dos agentes econômicos e não somente analisar o comportamento deles no passado. Juntamente com as contribuições metodológica de Sims (1972) e Hansen and Sargent (1980), a crítica de Lucas gerou uma enorme mudança de paradigma na pesquisa macroeconômica.

A mudança foi em direção a modelos que seguiam algum tipo de "disciplina" teórica, caracterizada pela imposição de restrições entre equações de modelos empíricos e introdução de comportamento *forward looking*. A contribuição de Kydland and Prescott (1982) levou esta mudança a um nível ainda maior. Nesta abordagem, a teoria não é somente usada para restringir o modelo estatístico, mas agora ela é a base sobre a qual a pesquisa empírica é feita. Kydland and Prescott (1982) buscaram desenvolver um modelo que conseguisse: 1) incorporar a análise do agregado (equilíbrio geral); 2) incorporar choques estocásticos; 3) incorporar mudanças de parâmetros. Este primeiro modelo era o modelo de ciclo real de negócios, que pode ser considerado como o pai de todos os modelos DSGE.

8.1 Modelos DSGE

Os modelos DSGE são caracterizados por regras de decisão derivadas dos problemas dos agentes com base na otimização (maximização de utilidade/lucro). Além disso, os agentes estão sujeitos a choques exógenos que alteram a produtividade total dos fatores; desvios de taxas de juros, etc. Dadas as distribuições dos choques, o DSGE gera uma densidade de probabilidade conjunta para variáveis endógenas como produto, inflação, etc. Se tivermos observações para as variáveis endógenas, essa pdf conjunta pode ser usada como a verossimilhança em estimações clássicas ou Bayesianas para os parâmetros estruturais do modelo DSGE.

8.2 Exemplo: modelo de ciclos reais de negócio

Kydland and Prescott (1982) usaram um modelo de equilíbrio geral sem imperfeições e microfundamentado para reproduzir ciclos econômicos americanos. A grande mudança trazida por esse modelo é a conclusão de que a economia, mesmo sem imperfeições apresentará ciclos (que até então acreditava-se que eram decorrentes apenas de imperfeições de mercado). Os ciclos acabam vindo da tecnologia, que sofre choques aleatórios com certa persistência. É um fenômeno exógeno trazendo o componente cíclico. Essa ideia veio do resíduo de Solow, que encontrou-se em evidências empíricas tem uma autocorrelação.

8.2.1 O problema de maximização intertemporal

O modelo básico contém um consumidor representativo que maximiza a utilidade do consumo e do lazer, sujeito a sua restrição orçamentaria.Existe também uma firma representativa que combina capital e trabalho para produzir um bem homogêneo, dado o nível de tecnologia existente. O capital se deprecia ao longo do tempo e investimentos aumentam o estoque de capital. O nível tecnológico, embora estocástico nunca pode ser negativo, por isso tiramos o log de Z_t. Dizemos que ele evolui de acordo com um modelo autoregressivo de ordem 1.

$$\max_{c_t, l_t} \quad E_0 \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t^{1-\sigma}}{1-\sigma} - L^{\lambda} \right), \tag{304}$$

s.a.

$$K_{t+1} = Z_t K_t^{\alpha} L_t^{(1-\alpha)} - C_t + (1-\delta) K_t \tag{305}$$

$$\log Z_{t+1} = \rho \log Z_t + \epsilon_t, \quad \epsilon_t \sim i.i.d. N(0, \sigma_{\epsilon}^2) \tag{306}$$

8.2.2 Condições de primeira ordem

A solução para este problema é caracterizada por 2 equações diferenciais parciais estocásticas, mais a lei de movimento das variáveis exógenas (capital e tecnologia) dadas por (305) and (306).

Condição intertemporal (equação de Euler):

$$\begin{aligned} c_t^{-\sigma} &= E_t \beta (1 + r_{t+1}) c_{t+1}^{-\sigma} \\ r_t &= \alpha z_t \left(\frac{k_t}{l_t} \right)^{\alpha - 1} - \delta. \end{aligned} \tag{307}$$

Condição intratemporal:

$$\begin{aligned} \lambda l_t^{\lambda - 1} &= w_t c_t^{-\sigma} \\ w_t &= (1-\alpha) z_t \left(\frac{k_t}{l_t} \right) \end{aligned} \tag{308}$$

A solução do modelo são as soluções políticas que dão a resposta ótima dos agentes frente ao problema de maximização e darão o nível de consumo e de trabalho ótimo para dado estoque de capital e nível de tecnologia. Teoricamente, poderíamos achar $L(K_t, Z_t)$ e $C(K_t, Z_t)$ resolvendo (307) e (308). Porém, a equação de Euler não tem fórmula fechada (uma solução analítica geral), isto é, como (307) é uma equação diferencial estocástica hiperbólica, não possui fórmula fechada e precisará ser aproximada. Então algumas aproximações (lineares e não lineares) surgem, como as descritas em Miranda and Fackler (2004) ou DeJong and Dave (2011). De maneira geral, elas terão a mesma estrutura:

$$\alpha_t = T(\alpha_{t-1}, \epsilon_t; \theta), \tag{309}$$

onde α_t é um vetor de variáveis de estado, ϵ_t é o vetor de choques e θ é o vetor de parâmetros estruturais do modelo.

As variáveis de estado são o L e o C; o choque é o ε e θ é o vetor de parâmetros estruturais do modelo. Se a aproximação é linear, então a aproximação T é linear em α e se T não é linear em α então a aproximação é obviamente não linear. A aproximação mais comumente utilizada é a linear, que irá envolver linearizar o modelo ou pelo menos as condições de primeira ordem. Utilizando a solução das condições de primeira ordem linearizadas no modelo original chega-se em uma aproximação. Isto é, no procedimento linear, a solução do modelo linear aproximado será usada como aproximação à solução não-linear desconhecida. Blanchard and Kahn (1980) introduziram um método para resolver sistemas de equações em diferença lineares que pode ser usado para solucionar a aproximação linear de modelos DSGE. Sims (2000), Klein (2000) e Uhlig et al. (1998) apresentam métodos alternativos. Ver DeJong and Dave (2011) para mais detalhes a respeito das diferenças entre estes métodos.

No final das contas resolve-se outro modelo e se utiliza a solução dele como solução do problema original (resolve-se o modelo sem os choques estocásticos, usando a noção de estado estacionário). Algumas formas diferentes são introduzidas, porém todos eles levam à mesma aproximação linear. A linearização é a expansão de Taylor na vizinhança do estado estacionário (que será o modelo na ausência de choques).

8.2.3 Função de verossimilhança da aproximação linear

A expansão de Taylor na vizinhança do estado estacionário nos dá:

$$\alpha_t = T\alpha_{t-1} + R\epsilon_t, \tag{310}$$

sendo que T e R dependem de θ e $\alpha = \{\widehat{k}_{t+1}, \widehat{z}_t\}$. Todas as outras variáveis podem ser escritas como:

$$y_t = Z(\theta)\alpha_t + u_t, \tag{311}$$

Note que precisamos que o número de choques exógenos seja no máximo igual ao número de variáveis observáveis para não ter problema de singularidade estocástica. Se ϵ_t e u_t possuírem distribuição normal, então o filtro de Kalman pode ser utilizado para calcular a verossimilhança do DSGE:

$$p(Y_T|\theta) = \int_{\mathbb{R}^k} \dots \int_{\mathbb{R}^k} \prod_{t=1}^{T} p\left(y_t|\alpha_t; \theta\right) p\left(\alpha_t|\alpha_{t-1}; \theta\right) \, d\alpha_1 \dots d\alpha_T$$

As variáveis com ”chapéu”estão denotando desvios do estado estacionário de variáveis de estado. O u_t irá denotar um erro de medida, mas ele pode ou não ser utilizado. A ideia é que a relação das variáveis com outras variáveis da economia é

determinística, porém o u_t está modelando as diferenças entre o que sai de um modelo e o que é observado.

Caso os choques sejam gaussianos, então como estamos fazendo uma aproximação linear, pode-se resolver a verossimilhança via filtro de Kalman. A solução do modelo DSGE dá uma estrutura em modelo de estado de espaços; a tecnologia e o capital dão os estados da economia e as outras variáveis, como consumo, emprego, investimento, etc, podem ser escritas como função dos estados da economia, onde tudo isso se liga via matriz Z. Temos então uma equação de estados em (310) e uma equação de medida em (311). No capítulo anterior, tínhamos definido a equação de transição dos estados como umpasseio aleatório e portanto ao invés de T (em 7), tínhamos a matriz identidade. Agora temos uma matriz qualquer.

A verossimilhança será um subproduto do problema de filtragem, que é a constante de integração da densidade a posteriori. Se temos a verossimilhança, podemos maximizá-la e fazer os cálculos seguindo a abordagem clássica.

8.2.4 Estimação bayesiana de DSGE linearizado

Suponha que queiramos estimar $\theta = \{\beta, \sigma, \lambda, \alpha, \delta, \rho, \sigma_\epsilon^2\}$ do modelo RBC usando séries do produto e emprego. Por que usar a série do produto e do emprego? No problema de maximização, se temos o consumo e o investimento, então conseguimos calcular o produto. Não há preocupações com os valores nominais pois estamos falando da economia real.

Como obter a priori? Precisamos de uma distribuição a priori para θ e podemos usar informações não contidas na amostra de produto e emprego para isso. Por exemplo, a média de longo prazo dos juros reais, da relação capital/trabalho e investimento como participação do PIB são informativos em relação a β, α e δ. Estudos microeconométricos podem nos informar a respeito da oferta de trabalho e, consequentemente, de λ. Deve-se escolher distribuições a priori que satisfaçam as restrições $0 < \alpha, \beta, \delta, \rho < 1$, $\sigma, \sigma_\epsilon^2 > 0$, entre outras restrições teóricas para os parâmetros. Uma das possibilidades é considerar o seguinte:

- β é o parâmetro de como o consumidor representativo desconta o futuro, logo a taxa de juros da economia ajuda para determinar uma priori para ele;
- A relação capital trabalho traz informações a resperito do α, coeficiente da Cobb-Douglas. $(1-\alpha)$ dá a participação do trabalho no produto. Essa informação vem das contas nacionais, de forma que novamente não precisa usar informações amostrais;
- O investimento menos o capital fornecem a informação a respeito da depreciação;
- O parâmetro λ dá a sensibilidade da função de utilidade às horas de trabalho. Com informações de estudos microeconômicos da sensibilidade do trabalho às variações de salário ajudam a determinar o valor do parâmetro λ (informações a respeito da oferta de trabalho).

Observe que na abordagem clássica todas essas informações a respeito da economia seriam ignoradas, pois só a informação amostral seria utilizada para fazer as estimativas.

Aqui vamos ter um problema no momento de ”virar”a verossimilhança: por causa da não linearidade do problema ao enxergarmos a verossimilhança como função para os parâmetros, não irá aparecer nenhuma densidade conhecida, de forma que não será possível determinar uma priori que será conjugada natural a ela. Isto não ocorria no modelo de regressão que estudamos nos capítulos anteriores.

Também devemos cuidar com o suporte dos parâmetros, pois muitos deles estão entre zero e um ou ainda existem parâmetros estritamente positivos (no caso das variâncias) e portanto as densidades a priori devem incorporar esses conhecimentos teóricos e outros conhecimentos que não estão contidos na amostra.

Por causa da relação não linear entre θ, T, R, Z, a posteriori conjunta não terá nenhuma densidade conhecida (por causa

do problema da verossimilhança falada anteriormente). Por isso, métodos numéricos são utilizados. O primeiro método utilizado foi amostragem por importância, mas daí entramos no problema de conseguir achar o amostrador por importância ótimo. Então o algoritmo de Metropolis-Hastings passeio aleatório para os problemas de estimação em DSGE. Apesar de sua não eficiência (pois fica percorrendo todo o suporte), ele acaba sendo mais indicado (a não ser que se conheça o amostrador por importância que resolva tudo). Existe ainda um trabalho do Chibb que faz um algoritmo específico para cada problema. A vantagem dele é que obriga a pessoa a pensar no modelo, a desvantagem é essa mesma (é um algoritmo de *Metropolis within Gibbs*). Usualmente, a distribuição candidata é $\mathcal{N}(\widehat{\theta}, c^2\widehat{\Sigma})$, com $\widehat{\theta}$ obtido através da maximização de $\log p(Y_T|\theta) + \log p(\theta)$ e Σ é o inverso do negativo do Hessiano calculado em $\widehat{\theta}$ e c é calibrado de forma a obter-se uma taxa de aceitação próxima de 50%.

A soma do log da verossimilhança com o log da priori é justamente o log da posteriori. A lógica é a mesma da maximização de verossimilhança: estima-se o máximo e encontra-se qual é a curvatura nesse ponto. O valor de c é utilizado para ter uma taxa de aceitação próxima de 50%. A inicialização do algoritmo pode ser feita em qualquer ponto (pois ele converge de qualquer jeito), porém começar do máximo da posteriori é uma forma de fazer o método ficar mais rápido.

Um problema comum em modelos DSGE é encontrar distribuições que sejam multimodais na verossimilhança, de forma que o algoritmo de maximização pode se perder nos máximos locais. Uma forma de evitar isso é iniciar o algoritmo de maximização em pontos distintos do suporte.

Algoritmo 4: Algoritmo RWMH para estimação de DSGE.

início

1. Use um otimizador numérico para maximizar $\log p(Y_T|\theta) + \log p(\theta)$, e calcular $\widehat{\theta}$ e $\widehat{\Sigma}$.
2. Amostre $\theta^{(0)} \sim \mathcal{N}(\widehat{\theta}, c^2\widehat{\Sigma})$.
3. Para $s = 1, \ldots, S$, amostre $\vartheta \sim \mathcal{N}(\theta^{(s-1)}, c^2\widehat{\Sigma})$ e defina $\theta^{(s)} = \vartheta$ com probabilidade $\min\{1, r(\theta^{(s-1)}, \vartheta|Y_T)\}$, ou $\theta^{(s)} = \theta^{(s-1)}$ caso contrário. Sendo que:
   $r(\theta^{(s-1)}, \vartheta|Y_T) = \frac{p(Y_T|\vartheta)p(\vartheta)}{p(Y_T|\theta^{(s-1)})p(\theta^{(s-1)})}$.
4. Defina o *burn in B* e use a amostra $\{\theta^s\}_{s=B}^{S}$ para realizar inferência a a respeito de $p(\theta|Y_T)$.

fim

Eliminamos as primeiras B amostras do burn in e então fazemos inferências, com as observações que ficaram, a respeito da posteriori de θ, sendo possível inclusive calcular as funções de impulso resposta do DSGE através da posteriori. Pode-se calcular os momentos, etc.

Observe que no passo (3) do Algoritmo (5) a equação de qualquer forma vai precisar de um método de filtragem, a dúvida é se vai ser um filtro de Kalman ou um filtro de Partículas. A cada realização das S amostras vamos precisar fazer o filtro de Kalman duas vezes, tornando o processo bem lento computacionalmente. A estimação de modelos DSGE se apoia fortemente no uso da função de verossimilhança, a diferença é que na abordagem bayesiana ao invés de resolver um problema de maximização vamos querer resolver um problema de integração, onde para isso utilizamos métodos de Monte Carlo (em particular, RWMH).

Normalmente utilizamos a média e a moda da priori como chute inicial no passo 1, porém o melhor é pegar vários pontos (por exemplo pontos distribuídos nos quartis da priori). De qualquer maneira, a dica para o chute inicial é sempre olhar para a priori e idealmente em vários pontos. Tem algumas análises bayesianas que param apenas no passo de determinar a máxima posteriori, onde não precisa determinar toda a posteriori, apenas fica no passo 1.

Resumindo o que vumos até o momento: vimos de onde que surgem os modelos DSGE em geral e falamos do modelo de ciclo real de modelos, considerado o primeiro modelo DSGE. Ele é um problema de otimização intertemporal condicionada e seu diferencial é incorporar conhecimento teórico econômico em um modelo econométrico.

Para estimação deste modelo, iremos utilizar um metropolis hastings passeio aleatório pois a verossimilhança não tem a forma de uma distribuição conhecida, impossibilitando encontrar uma conjugada natural ou até mesmo uma condicional conhecida, pois os parâmetros θ tem uma relação altamente não linear entre si. Sendo assim, a forma que conhecemos de amostrar de algo que não conhecemos é o RWMH. Observe que poderíamos usar amostragem por importância, porém como não se conhece nada da distribuição, fica difícil encontrar um bom amostrador. A vantagem do RWMH será justamente poder usar uma candidata para amostrar em todo o suporte, ficando em determinados locais que apresentem maior massa de probabilidade a posteriori.

8.3 DSGE e má especificação

Desde a década de 90, bancos centrais vem utilizando os modelos DSGE pois eles dão base teórica para entender ciclos econômicos e desenvolver política econômica. Os modelos puramente estatísticos como VAR não permitem modelagem de políticas econômicas, por exemplo, incorporar contrafactuais, utilizando as funções de reação. Os DSGE formam uma plataforma que permite fazer experimentos macroeconômicos, o que na prática não pode ser feito.

O problema é que estes modelos são uma simplificação da realidade, no sentido de que o modelo DSGE não é o modelo gerador dos dados macroeconômicos. Uma vez que toda a ideia de estimação por máxima verossimilhança envolve assumir que o modelo é gerador dos dados, acaba sendo uma limitação usar um modelo DSGE para a estimação dos parâmetros, porém ele é útil para fazer a parte de calibragem do modelo. Justamente por "combinar" a abordagem de calibração e estimação, a estimação Bayesiana de DSGEs ganhou tantos adeptos. O trabalho de Smets and Wouters (2003) chamou a atenção de diversos bancos centrais e mostrou que pode ser possível obter o melhor dos dois mundos: um modelo teoricamente fundamentado e ajustado à evidência empírica.

8.4 DSGE e previsão

Como o DSGE não previa muito bem (em relação aos VARs), Smets and Wouters (2003) tentaram colocar um monte de choques aleatórios (nas preferências, no markup dos preços, etc, etc). Só que isso acabou descaracterizando o modelo em relação à ideia original de Lucas, em ter um modelo microfundamentado (pois não havia nada que justificasse esses choques todos).

A ideia de Del Negro and Schorfheide (2004) foi a de, ao invés de usar uma priori de Minnesota que até tem alguma coisa de ideia econômica (por exemplo, fazer os passeios aleatórios), usar o DSGE como priori para o VAR. DSGEs possuem forte base teórica e têm obtido bom desempenho empírico, sendo candidatos óbvios para a priori.

8.4.1 DSGE como priori para o VAR

Podemos usar dados simulados do DSGE ao invés de utilizar as séries de dados originais. Obviamente a verossimilhança se mantém e agora teremos uma "verossimilhança"calculada com base nos dados simulados do DSGE, que pode ser interpretada agora como uma priori conjugada natural.

Como vimos, um modelo VAR(p) pode ser escrito como:

$$Y = XA + E, \quad \mathcal{MN}(0, I_T, \Sigma) \tag{312}$$

sendo que X contém termos determinísticos e as p defasagens de Y, e função de verossimilhança deste VAR é dada por:

$$p(Y|A,\Sigma) \propto |\Sigma|^{\frac{-T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\left(Y'Y - A'X'Y - Y'XA + A'X'XA\right)\right]\right\} \tag{313}$$

Já vimos que (313) é proporcional a uma densidade Normal-Wishart. Como o número de parâmetros θ do DSGE é usualmente muito menor do que os do VAR, o modelos DSGE pode ser usado para impor restrições em A e Σ.

Se simularmos $T^* = \lambda T$ observações (Y_{T^*}, X_{T^*}) do DSGE, então teremos uma verossimilhança para a amostra simulada:

$$\begin{aligned} &p(Y(\theta)_{T^*}|A,\Sigma) \\ &\propto |\Sigma|^{\frac{-\lambda T}{2}} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}\left(Y'_{T^*}Y_{T^*} - A'X'_{T^*}Y_{T^*} - Y'_{T^*}X_{T^*}A + A'X'_{T^*}X_{T^*}A\right)\right]\right\} \end{aligned} \tag{314}$$

Note que a verossimilhança (314) é idêntica à (313) e pode ser vista como uma a priori conjugada natural. (314) ainda depende dos parâmetros do modelos DSGE, portanto, pode ser interpretada como uma a priori hierárquica. Mas se usássemos o mesmo DSGE, com o mesmo θ, para gerar duas amostras diferentes (Y_{T^*}, X_{T^*}) e $(\widetilde{Y}_{T^*}, \widetilde{X}_{T^*})$, teríamos duas a prioris diferentes apesar de usarmos a mesma informação.

Sendo assim, existe uma crítica ao DSGE como priori pois a cada nova retirada de uma amostra, teremos valores diferentes coletados. Para contornar isso, ao invés de utilizar as quantidades amostrais, iremos utilizar na priori os momentos populacionais, que saem da densidade a posteriori cuspida pelo DSGE. Sendo assim, não é necessário simular nada.

Dessa forma, para eliminar a variação estocástica dada pela amostra finita (Y_{T^*}, X_{T^*}), substitui-se os momentos amostrais $Y'_{T^*}Y_{T^*}$, $X'_{T^*}Y_{T^*}$ e $X'_{T^*}X_{T^*}$ pelos seus respectivos momentos populacionais dados pelo DSGE.

Denotando $E_\theta[i_t j'_t] = \Gamma^*_{ij}(\theta)$, para $i, j \in \{y, x\}$, podemos rescrever (314) como:

$$p(A,\Sigma|\theta) \propto |\Sigma|^{\frac{-T}{2}} \exp\left\{-\frac{1}{2} tr\left[\lambda T \Sigma^{-1}\left(\Gamma^*_{yy} - A'\Gamma^*_{xy} - \Gamma^*_{yx}A + A'\Gamma^*_{xx}A\right)\right]\right\} \tag{315}$$

Se $\lambda T > M(p+1)$ e $(\Gamma^*_{xx})^{-1}$ existir, então (315) é uma a priori Normal-Wishart para um dado valor dos hiperparâmetros θ. A a priori para o modelo VAR possui portanto a seguinte estrutura hierárquica:

$$p(A,\Sigma) = p(A,\Sigma|\theta)p(\theta).$$

8.4.2 A posteriori para $A, \Sigma|\theta$

Como a a priori (315) é conjugada natural, $p(A,\Sigma|Y,\theta)$ também será Normal-Wishart, mais especificamente:

$$\begin{aligned} \Sigma^{-1}|Y,\theta &\sim \mathcal{W}([(\lambda+1)T\widetilde{\Sigma}(\theta)]^{-1}, (1+\lambda)T - M(p+1)), \\ A|Y,\Sigma,\theta &\sim \mathcal{N}(\widetilde{A}(\theta), \Sigma \otimes (\lambda T \Gamma^*_{xx}(\theta) + X'X)^{-1}), \end{aligned}$$

sendo que $\widetilde{A}(\theta)$ e $\widetilde{\Sigma}(\theta)$ são estimativas de MQO baseadas nas amostras simulada e verdadeira, com os momentos populacionais da amostra simulada. λ agora atua como o n atuava antes em relação ao peso dado na posteriori: se $\lambda \to 0$, não damos

confiança para o DSGE e quando $\lambda \to \infty$, o peso é todo para o DSGE. Por exemplo, se formos calcular $\hat{A} = (X'X)^{-1}X'Y$, estamos usando os momentos amostrais $(X'X)^{-1}$ e $X'Y$. Podemos fazer uma troca usando $(\Gamma^*_{XX})^{-1}(\theta)$ para representar $(X'X)^{-1}$ e $\Gamma^*_{XY}(\theta)$ para substituir $X'Y$.

O λ dá o peso dado para o modelo, no sentido que em trabalhos antigos como o do Goldberg, esse valor era utilizado para saber qual o tamanho da amostra simulada como fração de T. Só que, no momento que trocarmos o momento amostral pelo momento populacional, nós perdemos o impacto do número de observações (que estava presente em X original): mesmo que $(X'X)$ seja de tamanho $k \times k$, este produto considerou a quantidade de T observações. Isso some quando usamos o momento populacional e por isso vamos usar o λ para poder fazer uma ponderação e corrigir o viés que aparece ao usarmos o momento populacional. O lambda pode ser calculado através da verossimilhança marginal (que é usada para outras coisas, como por exemplo comparação de modelos). A verossimilhança marginal pode ser usada para selecionar λ e também para comparar a performance de diferentes modelos.

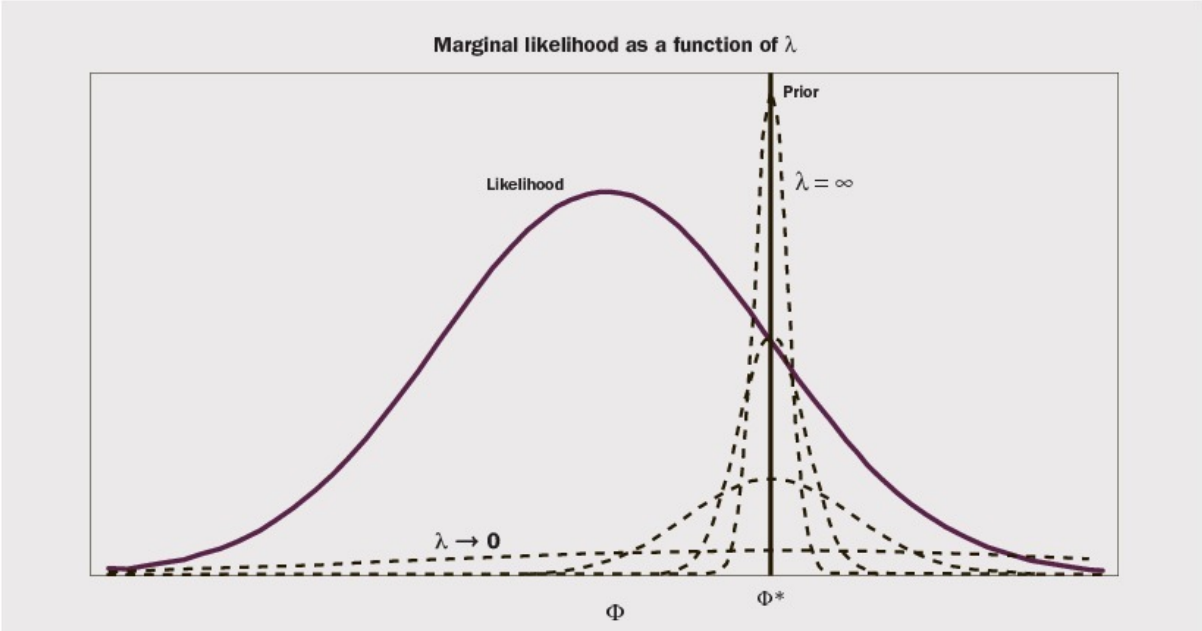


Figura 33: Verossimilhança e a Priori em um modelo DSGE.

8.5 Aprendendo a respeito de θ

Se temos informações a priori a respeito dos hiperparâmetros θ, toda a inferência a respeito do VAR pode ser feita analiticamente. Porém, normalmente estamos interessados em estimar θ e Del Negro & Schorfheide (2004) desenvolvem um algoritmo MCMC que permite aprender a respeito de θ, A e Σ simultaneamente. Como θ são hiperparâmetros, este procedimento pode ser visto como um método empírico de Bayes. Assim como na estimação de DSGEs, a posteriori para θ não possui fórmula analítica e será necessário utilizar o algoritmo Metropolis-Hastings dado em (5).

8.5.1 O algoritmo de Del Negro e Schorfheide (2004)

O algoritmo de Del Negro and Schorfheide (2004) parte de um espaço paramétrico finito $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ e realiza a estimação de θ, A e Σ q vezes para cada valor λ_i. Note que a verossimilhança marginal:

$$p_\lambda(Y) = \int p_\lambda(Y|\theta)p(\theta)d\theta$$

pode ser usada para determinar a densidade a posteriori de λ. O valor normalizado de $p_{\lambda_i}(Y)$ pode ser visto como a probabilidade a posteriori de λ_i em Λ para uma a priori com pesos iguais para todo $\lambda_i \in \Lambda$. Dado λ_i, utiliza-se o RWMH em (5) para gerar $\theta_s \sim p_{\lambda_i}(\theta|Y)$ e, dado esse θ_s, calcula-se $p(A, \Sigma|Y, \theta_s)$ e então amostrar A_s e Σ_s.

Algoritmo 5: Algoritmo de Del Negro e Schorfheide para estimação de DSGE.

início

para $\lambda \in \{1, \ldots, \Lambda\}$ **gere**

1. Para $s = 1, \ldots, S$, use o Algoritmo RWMH em (**??**) para amostrar $\theta_s \sim p_{\lambda_i}(\theta|Y)$;
2. Resolva o modelo DSGE para θ_s e calcule os momentos populacionais $\Gamma^*_{yy}(\theta_s)$, $\Gamma^*_{xy}(\theta_s)$, $\Gamma^*_{xx}(\theta_s)$;

fim

3. Calcule $p_{\lambda_i}(Y)$ usando o estimador de Geweke (1999) e encontre $\widehat{\lambda}$ com maior probabilidade a posteriori.
4. Selecione $\{\theta_s\}_{s=1}^S$ associado a $\widehat{\lambda}$ e use os resultados da a priori conjugada natural para amostrar $\{A_s\}_{s=1}^S$ e $\{\Sigma_s\}_{s=1}^S$ de $p(A, \Sigma|Y, \theta_s)$.
5. Use $\{A_s\}_{s=1}^S$ e $\{\Sigma_s\}_{s=1}^S$ para fazer inferência a posteriori a respeito do DSGE-VAR.

fim

A abordagem DSGE-VAR foi desenvolvida com o intuito de usar informações teóricas para melhorar a previsão de modelos VAR, pois estes são altamente parametrizados e geram estimativas imprecisas. Entretanto, Del Negro and Schorfheide (2004), Smets and Wouters (2007) também utilizam essa metodologia para comparar modelos DSGE distintos. Note que $p_\lambda(Y)$ calculada para diferentes modelos permite identificar qual deles melhor se ajusta aos dados. Del Negro and Schorfheide (2009) usam a metodologia para analisar a sensibilidade de políticas econômicas à má especificação de modelos.

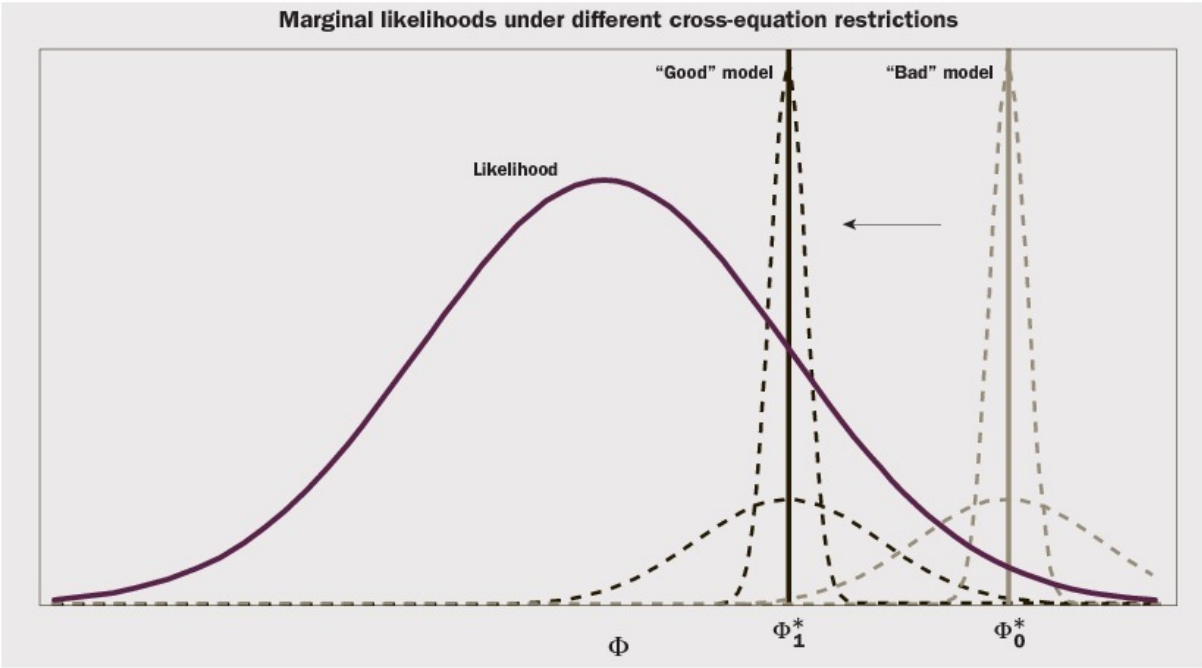


Figura 34: DSGE-VAR e comparação de modelos.

9 Parte 9 - Modelos VAR com parâmetros variando no tempo

Um modelo VAR pode ser expresso como

$$y_t = A \cdot y_{t-1} + u_t \quad u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), t = 1, 2, \ldots, T$$

Note que a suposição de que A é estático é uma condição para identificação do modelo, pois quando fazemos uma estimação por MQO, só podemos fazer isso pois assumimos que existe uma relação estável entre as variáveis.

No entanto, Lucas (1976) criticou os modelos econoétricos da época dizendo que não é razoável que os parâmetros (que representam relações econômicas) sejam estáveis, pois os agentes ao sofrerem choques vão mudar suas respostas ótimas. Por exemplo, nos EUA em 1970-1980 houve um período de inflação e desemprego altos. Cogley e Sargent, em 2001, estudaram esse período especificamente para entender o que aconteceu na gestão do Paul Volker quando o FED reagiu fortemente à inflação na época.

O modelo de Cogley e Sargent (2001) pode ser escrito como

$$y_t = A_t \cdot y_{t-1} + u_t \qquad u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma^1) \tag{316}$$

$$A_t = A_{t-1} + \eta_t \qquad \eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega^{-1}) \tag{317}$$

Observe que a equação (317) significa que esperamos que, em média, os coeficientes da equação de medida sejam iguais aos do período anterior mais um ruído. O sistema (316)-(317) pode ser visto como um modelo na forma de espaço de estados, onde os estados latentes (não observáveis) são dados pelas matrizes A_t, $t = 1, \ldots, T$.

Os parâmetros do sistema (316)-(317) são Σ, Ω e $A_1, A_2, \ldots, A_T$, isto é, temos uma matriz de coeficientes a cada instante de tempo. A proposta de Cogley e Sargent é estimar o modelo utilizando um amostrador de Gibbs.

1. Supondo que Σ e Ω sejam conhecidos, o sistema é linear e gaussiano e sabemos como usar o filtro e o suavizador de Kalman com o Algoritmo de Carter and Kohn (1994) para amostrar valores de $A_1, A_2, \ldots, A_T$.
2. Supondo que $\underline{A}_T = \{A_t\}_{t=1}^T$, $y_t - A_t \cdot y_{t-1} = u_t$ é conhecido (pois os y são as nossa observáveis). Isto é, u_t é observável e vem de uma distribuição normal. Podemos usar a priori conjugada natural $p(\Sigma) \sim \mathcal{W}(S_0, \nu_0)$ para obter as amostras de $p(\Sigma|\underline{A}_T, y)$ amostrando da densidade Wishart correspondente dada por $\mathcal{W}((S_0^{-1} + \underline{u}_T'\underline{u}_T)^{-1}, T + \nu_0)$ (a matriz de escala é o inverso do inverso da matriz de escala da priori mais a soma de quadrado dos resíduos e os graus de liberdade são o número de observações mais os graus de liberdade da priori).
3. Quando $\underline{A}_T$ são conhecidos, $A_t - A_{t-1} = \eta_t$ é conhecido e $p(\Omega|y, \underline{A}_T)$ quando baseada em uma priori $\mathcal{W}(Q_0, \nu_{q0})$ será dada por $\mathcal{W}((Q_0^{-1} + \underline{\eta}'\underline{\eta})^{-1}, T + \nu_{q0})$.

Após o trabalho de Cogley and Sargent (2001), houve uma crítica de Sims dizendo que, baseado em trabalhos dele e do Bernanke, há evidências de que a volatilidade está variando e o modelo de C&S estaria mal especificado. Após essa discussão, Cogley e Sargent em 2005 propõe um modelo que permite tanto o parâmetro variando no tempo quanto a presença de volatilidade estocástica. Neste segundo artigo eles concluem que existe tanto mudança em A_t quanto em Σ_t.

O modelo (316)-(317) na presença de volatilidade estocástica passa a ser dado por uma equação de medida e uma equação de transição dos estados similar ao que tínhamos antes mas agora é necessário uma outra transição dos estados para a matriz Σ:

$$y_t = A_t \cdot y_{t-1} + u_t \qquad u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_t^1) \tag{318}$$

$$A_t = A_{t-1} + \eta_t \qquad \eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega^{-1}) \tag{319}$$

$$\Sigma_t^{-1} = \lambda^{-1}\Sigma_{t-1}^{-1/2}\Theta_t\Sigma_{t-1}^{-1/2'} \qquad \Theta \overset{\text{i.i.d.}}{\sim} \mathcal{B}_M(n/2, 1/2) \tag{320}$$

Essa proposta foi feita por Uhlig (1997) e sua ideia é que a esperança da precisão em t dada a precisão em $t-1$ seria dada por:

$$\mathbb{E}[\Sigma_t^{-1}|\Sigma_{t-1}^{-1}] = \Sigma_{t-1}^{-1} \qquad \text{se}\lambda = \frac{n}{n+1}$$

Ou seja, a nossa esperança para a volatilidade é da mesma forma que a esperança para a equação de transição dos estados A_t, na ideia de um passeio aleatório. O choque Beta é motivado pela ideia que a Beta multivariada tem esperança dada por $\mathbb{E}[\Theta_t] = \frac{n}{n+1}\mathbb{I}$. Obs: note que é diferente de um Garch, pois lá temos muito mais valores para estimar e, ao mesmo tempo, a sua esperança condicional será conhecida se tivermos as variáveis passadas. O modelo de volatilidade estocástica permite que no período t tenhamos um choque que nos afaste da média incondicional do modelo. Este modelo então tem menos parâmetros e é mais flexível que um Garch. Note, no entanto, que o modelo (316)-(317) em relação a (318)-(320) tem mais parâmetros. Vamos adaptar então o nosso Gibbs anterior:

1. Supondo que $\underline{\Sigma}_T \equiv \{\Sigma_t\}_{t=1}^T$ e Ω sejam conhecidos, o sistema é linear e gaussiano e sabemos como usar o filtro e o suavizador de Kalman. A partir disso, $(A_1, A_2, \ldots, A_T|y, \underline{\Sigma}_T, \Omega)$ podem ser amostrados usando o algoritmo de Carter and Kohn (1994).
2. Supondo que $\underline{A}_T = \{A_t\}_{t=1}^T$, $y_t - A_t \cdot y_{t-1} = u_t$ é conhecido (pois os y são as nossa observáveis). Isto é, u_t é observável e sua distribuição é Normal porém com heterocedasticidade. Windle and Carvalho (2014) mostram como "adaptar"o algoritmo de forward filtering e backward sampling para a equação (320).
3. Quando $\underline{A}_T$ são conhecidos, $A_t - A_{t-1} = \eta_t$ é conhecido e $p(\Omega|y, \underline{A}_T)$, quando baseada em uma priori $\mathcal{W}(Q_0, \nu_{q0})$ será dada por $\mathcal{W}((Q_0^{-1} + \underline{\eta}'\underline{\eta})^{-1}, T + \nu_{q0})$.

Referências

Luc Bauwens, Michel Lubrano, and Richard Jean-François. *Bayesian inference in dynamic econometric models*. Univ. Press, 2003. 30, 32, 50, 60, 87, 113

Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to probability*. Athena Scientific, 2008. 15

Olivier Jean Blanchard and Charles M Kahn. The solution of linear difference models under rational expectations. *Econometrica: Journal of the Econometric Society*, pages 1305–1311, 1980. 139

João F Caldeira, Guilherme V Moura, and André AP Santos. Previsões macroeconômicas baseadas em modelos tvp-var: evidências para o brasil. *Revista Brasileira de Economia*, 69(4):407–428, 2015. 38

Chris K Carter and Robert Kohn. On gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994. 126, 135, 147, 148

George Casella and Roger L. Berger. *Statistical inference*. Wadsworth & Brooks, Cole, 1990. 36

Siddhartha Chib. Introduction to simulation and MCMC methods. In John Geweke, Gary Koop, and Herman Van Dijk, editors, *The Oxford Handbook of Bayesian Econometrics*, chapter 5, pages 183–217. Oxford University Press, Oxford, 2013. 96

Timothy Cogley and Thomas J Sargent. Evolving post-world war ii us inflation dynamics. *NBER macroeconomics annual*, 16:331–373, 2001. 38, 147

David N DeJong and Chetan Dave. *Structural macroeconometrics*. Princeton University Press, 2011. 139

Piet DeJong and Neil Shephard. The simulation smoother for time series models. *Biometrika*, 82(2):339–350, 1995. 126, 136

Marco Del Negro and Frank Schorfheide. Priors from general equilibrium models for vars. *International Economic Review*, 45(2):643–673, 2004. 142, 145

Marco Del Negro and Frank Schorfheide. Monetary policy analysis with potentially misspecified models. *American Economic Review*, 99(4):1415–50, 2009. 145

Marco Del Negro and Frank Schorfheide. Bayesian macroeconometrics. In John Geweke, Gary Koop, and Herman Van Dijk, editors, *The Oxford Handbook of Bayesian Econometrics*, chapter 7, pages 293–389. Oxford University Press, Oxford, 2013. 112

Thomas Doan, Robert Litterman, and Christopher Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100, 1984. 38, 119

James Durbin and Siem Jan Koopman. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616, 2002. 126, 135

Sylvia Frühwirth-Schnatter. Applied state space modelling of non-gaussian time series using integration-based kalman filtering. *Statistics and Computing*, 4(4):259–269, 1994. 126

Andrew Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 131–144. Chapmann & Hall CRC Press, Florida, 1996. 100, 101

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992. 100

J Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, Dawid A. P., and Smith A. F. M., editors, *Bayesian 4 - Proceedings of the Fourth Valencia International Meeting: Dedicated to the Memory of Morris H. DeGroot*, pages 169–193. Oxford University Press, Oxford, 1992. 99

John Geweke. Bayesian econometrics and forecasting. *Journal of Econometrics*, 100(1), 2001. 38

John Geweke, Gary Koop, and Herman Van Dijk, editors. *The Oxford handbook of Bayesian econometrics*. Oxford University Press, 2011. 38

Edward Greenberg. *Introduction to Bayesian econometrics*. Cambridge University Press, 2008. 5, 30, 94, 109

William H Greene. *Econometric analysis*. Pearson Education India, 2003. 39

Lars Peter Hansen and Thomas J Sargent. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control*, 2:7–46, 1980. 137

Dagoberto Adriano Rizzotto Justo, Esequia Sauter, Fabio Souto de Azevedo, Leonardo Fernandes Guidi, and Pedro Henrique de Almeida Konzen, editors. *Cálculo Numérico - um livro colaborativo*. REAMAT - Recursos Educacionais Abertos de Matemática, UFRGS, 2018. URL https://www.ufrgs.br/reamat/CalculoNumerico/livro-py/livro-py.pdf. 72

K Rao Kadiyala and Sune Karlsson. Numerical methods for estimation and inference in bayesian VAR-models. *Journal of Applied Econometrics*, pages 99–132, 1997. 117

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1): 35–45, 1960. 131

Gregor Kastner, Sylvia Frühwirth-Schnatter, and Hedibert Freitas Lopes. Efficient bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, (just-accepted), 2017. 38

Paul Klein. Using the generalized schur form to solve a multivariate linear rational expectations model. *Journal of Economic Dynamics and Control*, 24(10):1405–1423, 2000. 139

Gary Koop. *Bayesian econometrics*. Wiley, 2003. 5, 39, 40, 45, 50, 58, 67, 74, 100, 102, 106, 113, 117, 128, 153, 154, 155, 156

Finn E Kydland and Edward C Prescott. Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, pages 1345–1370, 1982. 137, 138

Robert B Litterman. Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986. 119

Robert E. Lucas. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1: 19–46, 1976. doi: 10.1016/s0167-2231(76)80003-6. 137

Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines and emerged triumphant from two centuries of Controversy*. Yale University Press, 2011. 38

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. 89, 95

Mario J Miranda and Paul L Fackler. *Applied computational economics and finance*. MIT press, 2004. 139

Ron Mittelhammer. *Mathematical statistics for economics and business*. Springer, 2013. 19, 23, 24

Guilherme Valle Moura. *Efficient importance sampling in applied econometrics*. PhD thesis, 2010. 71, 72, 73, 74

Charles R Nelson and Charles R Plosser. Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics*, 10(2):139–162, 1982. 118

Carlos Daniel Mimoso. Paulino, Turkman M. Antonia Amaral., and Bento Murteira. *Estatística Bayesiana*. Fundação Calouste Gulbenkian, 2003. 5

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. *The matrix cookbook*, volume 7. 2008. 39

Alexander Philipov and Mark E Glickman. Multivariate stochastic volatility via wishart processes. *Journal of Business & Economic Statistics*, 24(3):313–328, 2006. 38

Constantine Pozrikidis. *An introduction to grids, graphs, and networks*. Oxford University Press, 2014. 62

Christian P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007. 73, 87

Christian P Robert and George Casella. *Introducing monte carlo methods with r*, volume 18. Springer, 2010a. 87, 93

Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer, 2010b. 80, 85, 95, 110

Sheldon M. Ross. *A first course in probability*. Pearson Prentice Hall, 2010. 73

Reuven Y. Rubinstein. *Simulation and the Monte Carlo method*. 1981. 78

Shayle R Searle. *Matrix algebra useful for statistics*. John Wiley & Sons, 1982. 49, 63, 64

Christopher Sims. Second order accurate solution of discrete time dynamic equilibrium models. Technical report, Working Paper, Princeton University, 2000. 139

Christopher A Sims. Money, income, and causality. *The American economic review*, pages 540–552, 1972. 137

Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48, 1980. 112

Frank Smets and Raf Wouters. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European economic association*, 1(5):1123–1175, 2003. 142

Frank Smets and Rafael Wouters. Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, 97(3):586–606, 2007. 145

Rafael Stern and Rafael Izbicki. *Introducao àa Teoria das Probabilidades e Processos Aleatorios*. UFSCAR, 2016. 153

Harald Uhlig. Bayesian vector autoregressions with stochastic volatility. *Econometrica: Journal of the Econometric Society*, pages 59–73, 1997. 148

Harald Uhlig et al. A toolkit for analysing nonlinear dynamic stochastic models easily. *QM&RBC Codes*, 1998. 139

Jesse Windle and Carlos M Carvalho. A tractable state-space model for symmetric positive-definite matrices. *Bayesian Analysis*, 9(4):759–792, 2014. 148

Arnold Zellner. *An introduction to Bayesian inference in econometrics*. 1 edition, 1971. 38

Arnold Zellner. Applications of bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:23–34, 1983. 38

Arnold Zellner. Bayesian econometrics. *Econometrica: Journal of the Econometric Society*, pages 253–269, 1985. 38

10 Anexo 1 - Principais distribuições de probabilidade

Este conteúdo foi adaptado de Stern and Izbicki (2016) e Koop (2003).

10.1 Distribuições discretas

1. **Variável Aleatória Binomial** - $X \sim \text{Binomial}(n, p)$.

   Considere um experimento cujo desfecho é binário (sim x não, cara x coroa, alto x baixo, etc) onde há apenas uma probabilidade de sucesso envolvida, denotada por p (então o evento complementar sempre tem probabilidade $1 - p$). Este experimento é chamado *ensaio de Bernoulli*.

   Para uma variável aleatória X ter distribuição binomial, ela deve contar o número de sucessos em n experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. Por exemplo: total de faces caras observadas em 10 lançamentos de uma moeda.

   Caracterização:

   - **Notação:** Y segue uma distribuição Binomial de parâmetros n e p: $Y \sim Binomial(n, p)$
   - **Parâmetros:** $n \in \mathbb{N}$ é o número de ensaios de Bernoulli (também chamado de número de tentativas ou número de lançamentos) e $p \in [0, 1]$ denota a probabilidade de sucesso (que deve ser constante a cada repetição).
   - **f.m.p:** $\binom{n}{y} p^y (1-p)^{n-y}$, $y \in \{1, 2, \ldots, n\}$
   - **Média:** $\mathbb{E}(Y) = np$
   - **Variância:** $Var(Y) = np(1-p)$
   - **Mediana:** $Med(Y) = np$
   - **Moda:** $Mod(Y) = \lfloor (n+1)p \rfloor$
   - **f.g.m.:** $[pe^t + (1-p)]^n$
   - **Verossimilhança:**
   - **Obs:** Quando $n = 1$, temos uma distribuição $Bernoulli(p)$ e quando $n \to \infty$ e $p = \frac{\lambda}{n}$, temos uma $Poisson(\lambda)$.

2. **Variável Aleatória Geométrica** - $X \sim \text{Geom}(p)$.

   Considere uma série de experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. X é o número de experimentos até o primeiro "sucesso". Por exemplo, número de lançamentos de uma moeda até observar a primeira cara.

   Caracterização:

   - **Notação:** Y segue uma distribuição Geométrica de parâmetro p: $Y \sim Geometrica(p)$
   - **Parâmetros:** $p \in [0, 1]$ denota a probabilidade de sucesso (que deve ser constante a cada repetição).
   - **f.m.p:** $p(1-p)^{y-1}$, $y \in \{1, 2, \ldots, n\}$
   - **Média:** $\mathbb{E}(Y) = \frac{1}{n}$
   - **Variância:** $Var(Y) = \frac{1-p}{p^2}$
   - **Mediana:** $Med(Y) = \lceil \frac{-1}{log_2(1-p)} \rceil$

- **Moda:** $Mod(Y) = 1$
- **f.g.m.:** $\frac{pe^t}{1-(1-p)e^t}$

3. **Variável Aleatória Binomial Negativa** - $X \sim$ Binomial Negativa(r, p).

   Considere uma série de experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. X é o número de experimentos até obtermos r "sucessos". Por exemplo, número de lançamentos de uma moeda até observar 5 caras. Por definição, Binomial Negativa$(1, p)$ é o mesmo que $Geometrica(p)$.

   - **Notação:** Y segue uma distribuição Binomial Negativa de parâmetros r e p: $Y \sim BinomialNegativa(r, p)$
   - **Parâmetros:** $r \in \{1, 2, \ldots, n\}$ é o número de sucessos e $p \in [0, 1]$ denota a probabilidade de sucesso (que deve ser constante a cada repetição).
   - **f.m.p:** $\binom{y-1}{r-1}p^r(1-p)^{y-r}$, $y \in \{r, r+1, r+2, \ldots\}$
   - **Média:** $\mathbb{E}(Y) = r\frac{1}{p}$
   - **Variância:** $Var(Y) = r\frac{1-p}{p^2}$

4. **Variável Aleatória Poisson** - $X \sim$ Poisson(λ).

   A família de distribuições de Poisson frequentemente dá um bom modelo para o número de eventos (em particular, eventos raros) que ocorrem num período de tempo fixo (ou outra unidade fixa). Por exemplo, o número de clientes chegando em uma hora, número de reivindicações de seguro em um mês, número de terremotos num ano, número de erros de digitação numa página. λ é o número médio de eventos. A distribuição de Poisson pode ser vista como o limite de uma distribuição Binomial quando $n \to \infty$ e $p = \frac{\lambda}{n}$. Para a derivação completa, veja: https://goo.gl/zMAzu2.

   - **Notação:** Y segue uma distribuição de Poisson de parâmetro λ: $Y \sim Poisson(\lambda)$
   - **Parâmetros:** $\lambda \in \mathbb{N}$ o número médio de ocorrências em um determinado intervalo de tempo.
   - **f.m.p:** $\frac{\lambda^y e^{-\lambda}}{y!}$, $y \in \{1, 2, \ldots, n\}$
   - **Média:** $\mathbb{E}(Y) = \lambda$
   - **Variância:** $Var(Y) = \lambda$
   - **Mediana:** $Med(Y) \approx \lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \rfloor$
   - **Moda:** $Mod(Y) = \lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
   - **f.g.m.:** $e^{\lambda e^t - \lambda}$

5. **Distribuição Multinomial**[66] - $X \sim M(T, p)$. Um vetor aleatório $Y \in \mathbb{R}^n$, $Y = (Y_1, Y_2, \ldots, Y_n)'$ segue uma distribuição multinomial com parâmetros T e p se a sua função densidade de probabilidade é dada por:

$$f_M(y|T, p) = \begin{cases} \frac{T!}{y_1!\ldots y_N!}p_1^{y_1}\cdots p_n^{y_n}, \text{ se } y_i = 0, 1, \ldots, T \text{ e } \sum_{i=1}^{N} y_i = T \\ 0, \text{ caso contrário.} \end{cases}$$

   Onde $p = (p_1, \ldots, p_n)'$, $0 \leq p_i \leq 1$ para $i = 1, \ldots, N$, $\sum_{i=1}^{N} p_i = 1$ e T é um inteiro positivo.

   Se $Y \sim M(T, p)$, então $\mathbb{E}(Y_i) = Tp_i$ e $Var(Y_i) = Tp_i(1 - p_i)$ para $i = 1, \ldots, N$.

[66] Tradução de Koop (2003)

A distribuição multinomial é uma generalização da Binomial, onde um experimento com N possíveis resultados é repetido T vezes. O vetor aleatório Y conta o número de vezes que cada desfecho ocorre.

10.2 Variáveis aleatórias contínuas

Consulte o final da lista para a expressão das funções densidade.

1. **A Variável Aleatória Uniforme** - $X \sim \mathrm{U}(a, b)$.

   Todos os subconjuntos de (a, b) com o mesmo comprimento são equiprováveis. A distribuição é normalmente usada quando todos os pontos em (a, b) são "igualmente prováveis".

2. **A Variável Aleatória Exponencial** - $X \sim \mathrm{Exp}(\lambda)$.

   Este distribuição é usada, frequentemente, para modelar o tempo até um certo evento ocorrer. É a única distribuição contínua com propriedade de perda de memória.

3. **A Variável Aleatória Gama** - $X \sim \mathrm{Gama}(k, \lambda)$.

   A Gama(1, λ) é uma Exponencial(λ) e, assim, a distribuição Gama é a generalização da Exponencial. Se k é um número natural, $X = \sum_{i=1}^{k} Y_i$, onde Y_i são variáveis aleatórias independentes e $Y_i \sim$ Exponencial(λ).

   O livro do Koop tem uma parametrização diferente da Gama usual (que é a que está na tabela, porém com $\alpha = k$ e $\beta = 1/\lambda$). Para chegar na Gama do Casella e Berger , use na densidade do Koop os valores $\mu = \frac{\alpha}{\beta}$ e $\nu = 2\alpha$.

   Usando a densidade do Koop, temos que a qui-quadrado é um caso particular da Gama quando $\nu = \mu$ e denotamos por $Y \sim \chi^2(\nu)$. A distribuição exponencial é uma gama com $\nu = 2$. A inversa gama tem a propriedade de que se Y segue essa distribuição, então $\frac{1}{Y}$ tem distribuição gama.

4. **A Variável Aleatória Beta** - $X \sim \mathrm{Beta}(\alpha, \beta)$.

   Como a distribuição Beta assume valores em $(0, 1)$, ela é frequentemente usada para modelar frequências, probabilidades e razões.

5. **A Variável Aleatória Normal (Gaussiana)** - $X \sim \mathrm{N}(\mu, \sigma^2)$.

   Usando o Teorema do Limite Central, a distribuição Normal é frequentemente usada para aproximar a distribuição da média de uma sequência de variáveis aleatórias independentes e identicamente distribuídas.

6. **A distribuição normal multivariada**[67] - $Y \sim N(\mu, \Sigma)$.

   Um vetor aleatório $Y \in \mathbb{R}^k$, $Y = (Y_1, \ldots, Y_k)'$ tem uma distribuição normal multivariada com média μ ($\mu \in \mathbb{R}^k$) e matriz de covariância Σ ($\Sigma \in M_{k\times k}$ e Σ é positiva definita), se a sua função densidade de probabilidade é dada por:

$$f_N(y|\mu, \Sigma) = \frac{1}{2\pi^{\frac{k}{2}}}|\Sigma|^{-\frac{1}{2}} exp\left[\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right] \tag{321}$$

   Obs: Quando $k = 1$, $\mu = 0$ e $\Sigma = 1$, temos a normal padrão.

   Marginais e distribuições condicionais:

   Suponha que o vetor $Y \sim (\mu, \Sigma)$ pode ser particionado como:

[67] Traduzido de Koop (2003)

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix}$$

Onde $Y_{(i)}$ é um vetor de comprimento k_i para $i = 1, 2$ com $k_1 + k_2 = k$ e particionamos μ e Σ da seguinte maneira:

$$\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$$

e

$$\Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma'_{(12)} & \Sigma_{(22)} \end{pmatrix}$$

Então temos os seguintes resultados:

- A distribuição marginal de $Y_{(i)}$ é $N(\mu_{(i)}, \Sigma_{(ii)})$ para $i = 1, 2$;
- a distribuição condicional de $Y_{(1)}$ dado $Y_{(2)} = y_{(2)}$ é $N(\mu_{(1|2)}, \Sigma_{(1|2)})$ onde:

$$\mu_{(1|2)} = \mu_{(1)} + \Sigma_{(12)}\Sigma^{-1}_{(22)}(y_{(2)} - \mu_{(2)})$$
$$e$$
$$\Sigma_{(1|2)} = \Sigma_{(11)} - \Sigma(12)\Sigma^{-1}_{(22)}\Sigma'_{12}$$

Teorema B.10: Seja $Y \sim N(\mu, \Sigma)$ um vetor aleatório de k entradas e A uma matriz $m \times k$ não aleatória com $posto(A) = m$, então $AY \sim N(A\mu, A\Sigma A')$.

Teorema B.11: Suponha que o vetor Y de dimensão k seja tal que $Y \sim N(\mu, \Sigma)$. Então, a variável aleatória $Q = (Y - \mu)'\Sigma^{-1}(Y - \mu)$ tem distribuição qui-quadrado com k graus de liberdade, isto é, $Q \sim \chi^2(k)$ ou ainda $Q \sim G(k, k)$.

7. **A distribuição t multivariada**[68] - $Y \sim t(\mu, \Sigma, \nu)$.

   Um vetor $Y = (Y_1, \ldots, Y_n)' \in \mathbb{R}^k$ segue uma distribuição t multivariada com parâmetros $\mu \in \mathbb{R}^k$, $\Sigma \in M_{k\times k}$ positiva definida e ν, escalar positivo que representa os graus de liberdade, se a sua função densidade de probabilidade é dada por:

$$f_t(y|\mu, \Sigma, \nu) = \frac{1}{c_t}|\Sigma|^{-\frac{1}{2}}[\nu + (y - \mu)'\Sigma - 1(y - \mu)]^{-\frac{\nu+k}{2}} \tag{322}$$

   onde

$$c_t = \frac{\pi^{\frac{k}{2}}\Gamma(\frac{\nu}{2})}{\nu^{\frac{\nu}{2}}\Gamma\left(\frac{\nu+k}{2}\right)}$$

   Teorema B.12: Média e variância da t:

   Se $Y \sim t(\mu, \Sigma, \nu)$, então $\mathbb{E}(Y) = \mu$ se $\nu > 1$ e $Var(Y) = \frac{\nu}{\nu-2}\Sigma$ se $\nu > 2$. A Cauchy por exemplo, não tem média nem os

[68] Traduzido de Koop (2003)

outros momentos bem definidos, mesmo que sua f.d.p. tenha uma forma funcional e os quantis existam.

Variável Aleatória (Y)	fdp: $f_Y(y)$	$\mathbb{E}[Y]$	$Var[Y]$
Uniforme(a, b)	$\frac{1}{b-a}$ $y \in (a, b)$, $-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial(λ)	$\frac{1}{\lambda} e^{-\frac{y}{\lambda}}$ $y \in \mathbb{R}^+$	λ	λ^2
Gama(k, λ)	$\frac{1}{\Gamma(k)\lambda^k} y^{k-1} e^{-\frac{y}{\lambda}}$ $y \in \mathbb{R}^+$	$k\lambda$	$k\lambda^2$
Gama(μ, v) (Koop)	$\left(\frac{2\mu}{v}\right)^{v/2} \Gamma(\frac{v}{2}) y^{\frac{v-2}{2}} e^{-\frac{yv}{2\mu}}$ $y \in \mathbb{R}^+$	μ	$\frac{2\mu^2}{v}$
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $y \in (0, 1)$	$\frac{\alpha}{\alpha+\beta}$	$r \cdot \frac{1-p}{p^2}$
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$ $y \in \mathbb{R}$	μ	σ^2

Tabela 1: Revisão das Distribuições Contínuas

11 Anexo 2 - Propriedades de esperança, variância e covariância de v.a.'s

Definições básicas

Definição 11.0.1. 1 Seja X uma variável aleatória discreta com função massa de probabilidade denotada por p_X e que assume valores $x \in \chi$. O *valor esperado* ou *esperança matemática* ou *média* de X é definida por:

$$\mathbb{E}[X] = \sum_{x \in \chi} x_i p_X(x_i) \tag{323}$$

Se X é uma variável aleatória contínua com função densidade de probabilidade denotada por f_X, sua esperança será dada por:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{324}$$

Definição 11.0.2. 2 O *k-ésimo momento* da variável aleatória X é dado pela esperança de X elevada à potência k, isto é, $\mathbb{E}[X^k]$ (desde que essa quantidade esteja bem definida), para $k \in \{1, 2, \cdots\}$. Se a esperança de X for um número finito μ, isto é, se $\mathbb{E}[X] = \mu < \infty$, então definimos $\mathbb{E}[(X - \mu)^k]$ como o *k-ésimo momento central de X*, desde que essa quantidade esteja bem definida.

Definição 11.0.3. 3 Seja X uma variável aleatória com média finita denotada por μ. Sua variância é dada pelo momento central de ordem 2 de X:

$$Var[X] = \mathbb{E}[(X - \mu)^2] \tag{325}$$

Definição 11.0.4. 4 Sejam X e Y duas variáveis aleatórias definidas no mesmo espaço de probabilidade. A *covariância* entre elas será dada por:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{326}$$

Propriedades

As seguintes propriedades podem ser demonstradas a partir das definições básicas e a maioria pode ser adaptada para o caso de variáveis aleatórias contínuas. Neste caso, ao invés de somatórios, teremos as integrais correspondentes.

Propriedades da Esperança

1	Esperança da soma	$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$	É a soma das esperanças
2	Esperança do produto	$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ se X e Y indep.	É o produto das esperanças, desde que X e Y sejam independentes
3	Esperança de um escalar	$\mathbb{E}[\alpha] = \alpha,\ \alpha \in \mathbb{R}$	É o próprio escalar
4	Esperança de X vezes um escalar	$\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X],$ $\alpha \in \mathbb{R}$	É o escalar vezes a esperança de X
5	Esperança de um escalar mais X	$\mathbb{E}[\beta + X] = \beta + \mathbb{E}[X],$ $\beta \in \mathbb{R}$	É escalar mais a esperança de X
6	Esperança de uma função de X Lei do Estatístico Inconsciente	$\mathbb{E}[g(X)] = \sum_{x\in\chi} g(x_i)p_X(x_i)$ se X é discreta	$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx$ se X é contínua
7	Lei das Expectativas Iteradas	$\mathbb{E}[\mathbb{E}[X\|Y]] = \mathbb{E}[X]$	O valor esperado da esperança de X dado Y é a esperança de X.
8	Forma alternativa da esperança	$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i)$	Se X assume valores positivos.
9	Lei da esperança total	$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X\|A_i] \cdot \mathbb{P}(A_i)$	Se $A_1, \ldots, A_n$ uma partição de Ω e X uma v.a. discreta.

Tabela 2: Propriedades do valor esperado

Propriedades da Variância e Covariância

1	Forma alternativa da variância 1	$Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$	Variância de X é a esperança de X ao quadrado menos a média de X ao quadrado
2	Forma alternativa da variância 2	$Var[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$	
3	Forma alternativa da variância 3	$Var[X] = \mathbb{E}[(X-d)^2] - (\mathbb{E}[X]-d)^2,\ d \in \mathbb{R}$	se $Var[X] < \infty$
4	Variância da soma	$Var[X+Y] = Var[X] + Var[Y]$ Se X e Y indep.	É a soma das variâncias, se X e Y independentes
5.1	Variância da soma (caso geral)	$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$	É a soma das variâncias mais duas vezes a covariância
5.2	Variância da diferença (caso geral)	$Var[X-Y] = Var[X] + Var[Y] - 2Cov[X,Y]$	É a soma das variâncias menos duas vezes a covariância
6	Variância de um escalar	$Var[\alpha] = 0,\ \alpha \in \mathbb{R}$	É zero pois um número é sempre ele mesmo (não há variação)
7	Variância de X vezes um escalar	$Var[\beta X] = \beta^2 Var[X],\ \beta \in \mathbb{R}$	É o escalar ao quadrado vezes a variância de X
8	Variância de um escalar mais um escalar vezes X	$Var[\alpha + \beta X] = \beta^2 Var[X],\ \alpha, \beta \in \mathbb{R}$	
9	Forma alternativa da covariância	$Cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$	A covariância de X e Y é a esperança do produto menos o produto das esperanças.
10	Covariância de X com X	$Cov[X,X] = Var[X]$	Logo, se X é constante, $Cov[X,X] = 0$
11	Simetria da covariância	$Cov[X,Y] = Cov[Y,X]$	
12	Covariância de um escalar vezes X	$Cov[\alpha X, Y] = \alpha Cov[X,Y]$ $Cov[\alpha X, \beta Y] = \alpha\beta Cov[X,Y]$ $\alpha, \beta \in \mathbb{R}$	
13	Linearidade na 1^a entrada	$Cov[\alpha X + \beta Y, Z] = \alpha Cov[X,Z] + \beta Cov[Y,Z]$	
13.1	Linearidade na 1^a entrada e multiplicação por escalar	$Cov[\alpha X + \beta Y, \gamma Z + \delta W] = \alpha\gamma Cov[X,Z] + \alpha\delta Cov[X,W] + +\beta\gamma Cov[Y,Z] + \beta\delta Cov[Y,W]$	
14	Covariância de sequências de variáveis aleatórias	$Cov[\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_i Y_i] = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j Cov[X_i, Y_j]$	Suponha $\{X_1, \cdots, X_n\}$ e $\{Y_1, \cdots, Y_m\}$ duas sequências de variáveis aleatórias e $\{\alpha_1, \cdots, \alpha_n\}$ $\{\beta_1, \cdots, \beta_n\}$ sequências de constantes.

Tabela 3: Propriedades da variância e Covariância