UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Roberto Sousa Rocco

Markov Chain Monte Carlo e algumas aplicações

São Carlos
2019

Roberto Sousa Rocco

Markov Chain Monte Carlo e algumas aplicações

Monografia apresentada ao Curso de Engenharia Elétrica com Ênfase em Sistemas de Energia e Automação, da Escola de Engenharia de São Carlos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Engenheiro Eletricista.

Orientador: Prof. Dr. Carlos Dias Maciel

USP – São Carlos
2019

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Rocco, Roberto Sousa
R671m Markov Chain Monte Carlos e algumas aplicações / Roberto Sousa Rocco; orientador Carlos Dias Maciel. São Carlos, 2019.

Monografia (Graduação em Engenharia Elétrica com ênfase em Sistemas de Energia e Automação) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2019.

1. Markov Chain Monte Carlo. 2. MCMC. 3. Metropolis-Hastings. 4. Filtro de Kalman. 5. Gibs. 6. Inferência Bayesiana. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE APROVAÇÃO

Nome: Roberto Sousa Rocco

Título: "Markov Chain Monte Carlo e algumas aplicações"

Trabalho de Conclusão de Curso defendido e aprovado em 26 /06 /2019 ,

com NOTA 7,0 (sete, zero), pela Comissão Julgadora:

Prof. Associado Carlos Dias Maciel - Orientador - SEL/EESC/USP

Prof. Dr. Danilo Hernane Spatti - SSC/ICMC/USP

Mestre Talysson Manoel de Oliveira Santos - Doutorando - SEL/EESC/USP

Coordenador da CoC-Engenharia Elétrica - EESC/USP:
Prof. Associado Rogério Andrade Flauzino

Dedicado à minha família.

Agradecimentos

A Deus pelo dom da minha vida e por sempre ter enviado anjos para me ajudarem em minha jornada.

Aos meus pais, pelo amor incondicional, dedicação constante e por serem meus maiores exemplos de vida.

A Universidade de São Paulo , por garantir o direito a um ensino superior de qualidade a todos.

A Escola de Engenharia de São Carlos, na figura de seus professores e funcionários, pela empenho em transformar a vida de muitos jovens por meio do conhecimento.

Ao meu orientador, não só pelas conversas e ensinamentos passados, mas também pela paciência e compreensão.

Aos meus irmãos, por demonstrarem em todos os momentos que mais importante que “ser família” é “se fazer família” nas situações mais complexas.

E a todos os meus amigos e colegas, que participaram dessa jornada, com os quais pude compartilhar diversos momentos e muitos aprendizados.

"Não sabendo que era impossível, ele foi lá e fez."

JEAN COCTEAU

Resumo

ROCCO, R. S. ***Markov Chain Monte Carlo* e algumas aplicações**. 2019. 51 p. Monografia (Trabalho de Conclusão de Curso) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos–SP, 2019.

Os métodos de reamostragem de *Markov Chain Monte Carlo* (MCMC) foram essenciais para a evolução da inferência bayesiana, uma vez que a inferência dos parâmetros poderá se apresentar *a posteriori* de diversos modos e muitas vezes analiticamente não-trivial. Assim, a modelagem MCMC tem um papel fundamental na aplicação dos métodos bayesianos em diversas áreas como a engenharia, ciências médicas, finanças, inteligência artificial, dentre outras. Este trabalho de conclusão de curso tem por objetivo discutir o pensamento Bayesiano, incorporando os métodos MCMC, além de exemplificar a sua utilização via modelagem de regressão utilizando o filtro de Kalman. Foram analisados e comparados os seus desempenhos por meio de estudos com dados reais, a fim de observar a adequabilidade dos paradigmas sob análise, utilizando para tal, os ferramentais de ciências de dados providos pelos ambientes de programação R e Python. Os achados reverberam o potencial dos métodos discutidos para a resolução de problemáticas contemporâneas.

Palavras chave: Inferência Bayesiana; Cadeias de Markov; Monte Carlo; Métodos de reamostragem; Modelo dinâmico.

Abstract

ROCCO, R. S. **Markov Chain Monte Carlo and some applications**. 2019. 51 p. Monografia (Trabalho de Conclusão de Curso) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos–SP, 2019.

Markov Chain Monte Carlo (MCMC) resampling methods were essential in the evolution of Bayesian inference, since the *a posteriori*inference of the parameters may be often analytically non-trivial. MCMC has a fundamental role in the application of Bayesian methods in several areas such as engineering, medical sciences, finance, artificial intelligence, etc. This work aims to discuss Bayesian paradigma, incorporating the MCMC methods, exemplifying it through regression model (whereas Kalman filter). Their performance was analyzed in studies with real data, in order to observe the suitability of the paradigms under analysis using the software such as R and Python. The findings reverberate the potential of the methods discussed to solve contemporary problems.

Keywords: Bayesian Inference; Markov chains; Monte Carlo; Resampling methods; Dynamic model.

Lista de Figuras

Lista de Abreviaturas e Siglas

ARIMA	Autoregressivos integrados e de médias móveis
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
CO_2	Dióxido de Carbono
FFBS	Forward Filtering Backward-Sampling
MCMC	Markov Chain Monte Carlo
MLE	Estimador de Máxima Verossimilhança

Sumário

Capítulo 1

Introdução

Atualmente, na era dos dispositivos inteligentes que a todo instante adquirem dados por meio de seus diversos sensores, com isso, gerando grandes bases de dados, tem-se a disposição um elemento com potencial inerente: a informação disponível, que traz consigo uma larga e complexa estrutura dinâmica (CORCHADO *et al.*, 2014). O levantamento de registros de dados nas últimas décadas, despertou várias inquietações acompanhadas de tentativas de pesquisas e desenvolvimentos no campo da mineração de dados. Visto isso, é desejável utilizar a informação obtida em uma amostra (parte de um todo ou conjunto) com o intuito de generalizar e modelar os fenômenos observados ou tomar decisões. Esse procedimento é chamado inferência estatística para a tomada de decisão.

No contexto da inferência estatística, o paradigma bayesiano associa exclusivamente, a partir da aplicação lógica indutiva, o calculo de probabilidades a essas inferências (LINDLEY *et al.*, 1956). O'Hagan e Forster (1994) complementam essa assertiva, defendendo que qualquer questão envolvendo probabilidades, tem uma e apenas uma resposta, embora possa haver múltiplas maneiras de derivar essa resposta. Assim, a atribuição de distribuições de probabilidades a quantidades incertas, sejam observáveis (dados) ou não (modelos estatísticos), possibilita incorporar os conceitos de subjetividade como representação dos graus de credibilidade, que é condicionada à informação disponível.

Diante da complexidade que um modelo estatístico pode apresentar, muitas vezes relacionado à alta-dimensionalidade dos dados sob modelagem, uma aproximação numérica é necessária a fim de se obter essa distribuição de probabilidades relacionada à tomada de decisão do incerto. O método de Monte Carlo, criado pelo matemático russo Andrei Markov, em 1906, combinado com o conceitual das cadeias de Markov, que foi apresentado por Lindley, em 1965, originaram algoritmos considerados bastante eficientes, como o amostrador de Gibbs (ELLIOTT *et al.*, 1984) e o Metropolis-Hasting (HASTINGS, 1970), para a obtenção dessas aproximações numéricas e que, hoje, graças ao poder computacional

dos dispositivos, são utilizados corriqueiramente.

Os modelos de espaço-estado contemplam uma classe de modelos estatísticos ampla e capaz de incorporar a dinâmica de eventos, mesmo que esses eventos provenham de dados sintéticos, ou seja, que não foram observados diretamente. Um caso particular bastante utilizado atualmente em aplicações nas áreas de engenharia, ciências médicas, finanças, inteligência artificial, dentre outras, é o filtro de Kalman (KALMAN, 1960).

O filtro de Kalman se mostra como uma alternativa competitiva quanto a modelagem de um problema, dado a sua flexibilidade e rapidez na convergência dos dados. Kim e Kim (2009), Xie, Sudhakar e Zhuang (1995) apresentaram o filtro de Kalman (e algumas de suas variações) como um método de rastreamento de elementos visuais em imagens sequenciadas (neste caso, uma bola) e movimentos oculares utilizando câmeras de vídeo. Os resultados dos trabalhos demonstram que o filtro de Kalman tem a vantagem de ser não intrusivo, barato, e com alto grau de automatização.

Desse modo, este trabalho de conclusão de curso em Engenharia Elétrica com Ênfase em Sistemas de Energia e Automação, tem como objetivo central discutir o pensamento Bayesiano, incorporando os métodos de reamostragem, conhecidos como *Markov Chain Monte Carlo* (MCMC), além de exemplificar via modelagem com o filtro de Kalman, que é um caso particular de modelos dinâmicos. Foram analisados e discutidos os desempenhos de cada uma das técnicas, em estudos de caso aplicados sobre dados sintéticos e reais, a fim de observar a adequabilidade dos paradigmas em análise.

Para garantir uma melhor fluidez do texto, este trabalho foi estruturado em sete capítulos, como segue:

- Inicialmente, no Capítulo 2, será discutido um breve histórico sobre o surgimento do pensamento Bayesiano, seguido de uma sumarização do paradigma Bayesiano, bem como do método de reamostragem MCMC;
- Os Capítulos 3 e 4 contém os fundamentos teóricos Bayesianos e de MCMC;
- Posteriormente, o Capítulo 5 apresenta em detalhes o filtro de Kalman;
- O Capítulo 6 contém uma aplicação sobre conjuntos de dados amostrados, ou seja, de mundo real;
- Por fim, o Capítulo 7 apresenta as discussões dos resultados obtidos, algumas limitações e possíveis sugestões para serem consideradas em trabalhos futuros.

Capítulo 2

Uma breve história

> "O teorema de Bayes é, para muitos, um dos poucos resultados da matemática que se propõe caracterizar a aprendizagem com a experiência, isto é, a modificação da atitude inicial em relação aos "antecedentes", depois de ter a informação adicional de que certo acontecimento(s) se realizam (depois de conhecer os dados do experimento ou observação)." (PAULINO, TURKMAN & MURTEIRA, 2003 - pg. 10)

O termo "bayesiano" refere-se aos conceitos matemáticos e estatísticos desenvolvidos pelo inglês Thomas Bayes (1702–1761), que provou um caso particular daquele que hoje é chamado de Teorema de Bayes, que, no entanto, foi fundamentado por Richard Price (1723–1791) (BAYES, 1991). Foi Pierre-Simon Laplace (1749–1827) quem introduziu uma versão geral do teorema de Bayes, contemplando os rigores matemáticos do paradigma em problemas de mecânica celeste, estatística médica, confiabilidade e jurisprudência. A inferência Bayesiana inicial era chamada de "probabilidade inversa" (visto que infere a partir de observações sobre parâmetros, ou de efeitos para causas).

A medida em que a revolução computacional inundava o mundo moderno com dados dos mais variados tipos, adquiridos por meio das mais diversas fontes sensoras, o teorema de Bayes enfrentava uma de suas maiores crises dentro de seus quase de 250 anos de existência, por volta dos anos de 1950–1980. Estimar o relacionamento entre as múltiplas variáveis de um conjunto de dados, e também determinar o efeito que uma perturbação em uma variável é refletida nas outras e no todo. Assim, a dimensionalidade dos dados atormentava tanto os estatísticos frequentistas quanto os bayesianos.

Na década de 1980, houve um crescimento expressivo tanto na pesquisa quanto em aplicações dos métodos Bayesianos, principalmente aqueles atribuídos à descoberta da cadeia Markov e métodos de Monte Carlo (*Markov Chain Monte Carlo*, MCMC). Esse crescimento foi impulsionado pelo avanço da tecnologia e a consequente remoção (ou ao menos redução) das diversas barreiras computacionais existentes, o que viabilizou o já

seminal interesse em aplicações não-padrão, ou complicadas. Os métodos Bayesianos são amplamente aceitos e usados na atualidade, ganhando espaço em análises de dados, especialmente na era da Inteligência Computacional.

Entre os grandes pesquisadores e alguns dos seus respectivos trabalhos, que contribuíram para o desenvolvimento de aplicações dos métodos bayesianos, ao longo do século XX, e que não poderiam deixar de serem citados neste texto, estão Ramsey (1926), Jeffreys (1946), Finetti (1937), Good (1950), Savage *et al.* (1962), Lindley *et al.* (1956).

Deve ser destacado que as informações históricas apresentadas aqui foram extraídas do livro *The theory that would not die*, de autoria de McGrayne (2011). As seções apresentadas a seguir foram inspiradas pelo livro *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*, elaborado por Gamerman e Lopes (2006).

2.1 Uma visão geral bayesiana

Este mundo apresenta incertezas, seja da própria natureza do fenômeno observado ou vinculada à sua métrica. As ferramentas utilizadas para medições de incertezas são derivadas da teoria das probabilidades. Desse modo, pode-se dizer que um modelo estatístico é aquele que quantifica matematicamente as relações entre as variáveis (aleatórias e/ou determinísticas), por meio da especificação de distribuições de probabilidade.

A inferência estatística pode ser definida como a ciência dedicada a tirar conclusões sobre os dados por meio de medidas quantitativas. A partir da inferência estatística é possível atingir a generalização dos resultados obtidos por meio de uma amostragem, isto é, uma parte de um conjunto total chamado de população.

É comum encontrar incertezas associadas às medições, sejam elas ocasionadas por algum grau de imprecisão inerente aos dispositivos de captura dos dados, ou mesmo devido ao processo sob o qual essas quantificações se tornam disponíveis, que pode não ser totalmente controlado ou compreendido.

Considere um exemplo oriundo de um estudo que relaciona a eficiência da comunicação entre um transmissor e um receptor de um determinado sinal, ou seja, deseja-se obter uma estimativa da informação. Suponha ainda que semanalmente deseja-se descrever a distribuição energética a fim de analisar e descrever as operações sobre estes sinais em uma dada rede.

Espera-se que o transmissor de canal (x) tenha o sinal recebido, isto é, a comunicação seja bem efetuada traduzindo a informação emitida. Assim, de um modo simplificado, é possível assumir uma relação determinística (ou relação linear) vinculando esse efeito à

probabilidade do funcionamento do receptor (y).

$$y = \alpha + \beta x$$

Logo, essa probabilidade associa a estimação desse volume de informação recebida, bem como a sua eficiência. Seja $y \in [0, 1]$, faz-se necessário garantir que a resposta do modelo estará também inserida dentro do domínio $[0, 1]$. Uma transformação comum que pode ser aplicada é a *logit*:

$$logit(y) = log\left(\frac{y}{1-y}\right) = \alpha + \beta x$$

O parâmetro β associará a eficiência da transmissão quanto relativamente ao aumento à propagação da informação. Assim, quanto maior for o valor desse parâmetro β, melhor será a eficácia do transmissor. A coleta das possíveis relações, juntamente com as especificações de probabilidade para as porcentagens dessas informações transmitidas, define um modelo estatístico.

Inferências poderão ser conduzidas sobre os dados a partir da definição do modelo estatístico. A inferência é desenhada ao se construir uma distribuição de probabilidade conjunta abrangendo todas as quantidades não observadas, com base em tudo aquilo que é conhecido sobre as mesmas.

Dentro do paradigma bayesiano são associadas incertezas à todas as quantidades inicialmente desconhecidas, bem como há a inclusão de informação *a priori* no processo de estimação. A estatística bayesiana combina informações vividas sobre os fenômenos em estudo, além de também se basear em valores de quantidades observadas, quando estas estão disponíveis. Neste caso, a distribuição de probabilidade relacionada às incertezas condicionadas às informações conhecidas é chamada de distribuição posterior, pois é obtida após os dados terem sido observados. As quantidades desconhecidas, originadas de observações futuras, são chamadas de "previsão" e sua distribuição marginal é chamada de distribuição preditiva.

A experiência com o sistema de transmissão de sinais pode fornecer algumas informações básicas. A fonte de informação é dada pelo resultado da informação recebida, uma vez que elas são realizadas e seus porcentuais se tornam conhecidos. Dessa forma, a inferência bayesiana fornece as ferramentas para combinar esses conhecimentos de modo a obter a distribuição posterior de α e β, baseado no conhecimento prévio e nas observações coletadas. Com isso, é possível prever os resultados futuros das transmissões e assim gerar uma previsão de eficiência/confiabilidade do sistema em eventos futuros.

Obter a distribuição posterior é um passo importante, mas não a etapa final do processo. Deve-se conseguir extrair as informações significativas dessa distribuição e traduzi-las em termos de seu impacto dentro do escopo do estudo. A principal preocupação envolvida nessa etapa se baseia na avaliação de resumos pontuais como média, mediana ou moda, ou também resumos de intervalares dados por intervalos de probabilidade. A extração ou sumarização podem ser realizadas analiticamente, o que significa que poderá ser feita uma avaliação exata ou por método de reamostragem da situação.

Voltando ao estudo sobre transmissão de sinal, o principal interesse nesse cenário está relacionado com a avaliação do valor de β. Se sua distribuição está concentrada com grande probabilidade em torno de valores positivos, o estudo confirma que a informação recebida aumenta a medida que o sistema é mais eficiente. Porém, a quantificação também é uma informação importante: quanto maiores forem os valores de β, melhor poderá ser a captação da informação, no que se refere ao aumento da "tradução" da mesma.

Na maioria dos casos, entretanto, a complexidade da modelagem acaba por impedir que essa simples operação possa ocorrer. Essa complexidade geralmente é causada pela combinação das fontes de informação disponíveis ligadas a uma determinada quantidade. Em outros casos, pode ser causada pelo grande volume de quantidades necessárias para uma descrição adequada dos fenômenos estudados. Em alguns casos, pode ainda ser causada por uma combinação de muitas quantidades com muitas fontes de informação.

No exemplo anterior, uma descrição mais adequada do processo será fornecida por um modelo que permita que as ligações entre transmissor de canal e funcionamento do receptor mudem com o decorrer do tempo. Diferentes variações, sendo elas observadas diretamente ou não corroboram com uma modelagem dinâmica. Uma representação possível é permitir que as quantidades α e β variem com o passar do tempo. A relação entre a probabilidade de reconhecimento ϕ_t e a despesa x_t, no período t, torna-se:

$$logit(y_t) = \alpha_t + \beta_t x_t$$

onde as quantidades desconhecidas α_t e β_t agora podem mudar com decorrer dos intervalos de tempos em que os eventos estão discretizados. Isso leva automaticamente a um aumento expressivo no número de quantidades desconhecidas. Além disso, deve-se esperar algum grau de similaridade entre os próximos períodos. Um modo conveniente para especificar as similaridades é:

$$\alpha_t = \alpha_{t-1} + \omega_{1t}$$
$$\beta_t = \beta_{t-1} + \omega_{2t}$$

Observe que a quantidade apresentada no período t, tanto para α e β, poderá ser explicada pela observação mais próxima $(t-1)$ adicionado a um ruído (aleatório). Note também que o número de quantidades desconhecidas (complexidade do modelo) aumentou de 2 para $2n$, onde n é o número de períodos de tempo considerados no estudo. A distribuição das quantidades desconhecidas tornou-se, consequentemente, mais difícil para se lidar.

Buscar aproximações que possam fornecer, ao menos, uma "boa" aproximação da resposta exata, se faz necessário face ao aumento na complexidade do modelo. Algumas pesquisas apresentadas na literatura trabalharam no sentido de indicar possíveis modos de se obter boas aproximações, com maior intensidade de pesquisas ocorrendo a partir da década de 1980. Esse fenômeno pode ser explicado graças ao aumento no poder de computação, o que permite soluções mais sofisticadas e baseadas em cálculos iterativos. Essas soluções podem ser amplamente divididas em aproximações analíticas (determinísticas), enquanto outras são baseadas em aproximações numéricas (estocásticas).

Uma perspectiva completamente diferente para extrair informações relevantes contidas em uma determinada distribuição é fornecida pela simulação estocástica. Neste caso, a abordagem é usar valores simulados da distribuição de interesse. Uma coleção desses valores forma uma amostra que define uma distribuição discreta concentrada nos valores da amostra. A distribuição desses valores é uma aproximação da distribuição referência usada para a simulação. Então, todos os cálculos relevantes com a distribuição referência podem ser feitos de modo aproximado com a distribuição da amostra.

Em particular, a amostra pode ser agrupada dentro de intervalos e o histograma de frequências relativas exibido graficamente. Se uma grande quantidade desses valores for simulado, o histograma resultante será uma aproximação muito semelhante à densidade da distribuição de interesse, que é o foco deste trabalho.

As técnicas de simulação estocásticas, ou de Monte Carlo, têm algumas características atraentes que podem explicar seu recente sucesso dentro do contexto da Inferência Estatística. Primeiro, elas têm forte apoio em resultados de probabilidade, como a lei dos grandes números, por exemplo. Asseguram que a aproximação se torna cada vez melhor à medida que o número de valores simulados aumenta. Esse número é controlado pelo pesquisador e apenas será considerado o tempo e o custo que podem impedir uma

aproximação praticamente livre de erros. Além disso, em qualquer estágio do processo de simulação, os erros de aproximação podem ser medidos probabilisticamente usando o teorema do limite central.

Logo, a motivação deste trabalho está inserida no contexto da descrição de técnicas dedicadas a realizar inferência bayesiana com base em simulação estocástica. Usando essa gama de técnicas, é possível criar esquemas de simulação para modelar valores a partir da distribuição de α e β (na configuração de modelo estático). Entretanto, estes parâmetros não fornecem soluções adequadas para o caso mais elaborado que inclui a variação temporal, α_t e β_t. Essas técnicas tendem a ser ineficientes à medida que a dimensionalidade das quantidades desconhecidas aumenta, obrigando a aplicação de técnicas mais sofisticadas de simulação.

2.2 Uma visão geral sobre MCMC

Atualmente, existem muitos problemas de interesse que se enquadram na categoria dos modelos de grandes dimensões. Configurações dinâmicas (contendo parâmetros variando no tempo) são apenas um tipo de exemplo neste contexto. Outros exemplos também surgem no que se refere aos modelos de efeitos hierárquicos ou aleatórios para dados espaciais. O primeiro grupo trata a variação aproximada adicional não-estruturada, enquanto o segundo grupo trabalha com variações relacionadas à estruturas vizinhas. Modelos que consideram os erros de medição e uma mistura ou combinação de diferentes modelos também podem ser configurações para modelos de grandes dimensões.

Markov Chain Monte Carlo (MCMC) define uma área da estatística que fornece uma resposta para o difícil problema da simulação, a partir da distribuição altamente dimensional das quantidades desconhecidas que aparecem em modelos complexos. É possível dizer, em termos muito amplos, que as cadeias de Markov são processos que descrevem trajetórias em que as quantidades sucessivas são descritas probabilisticamente, de acordo com o valor de seus resultados predecessores e imediatos. Em muitos casos, esses processos tendem a um equilíbrio e as quantidades limitantes seguem uma distribuição invariante.

As técnicas baseadas em MCMC permitem a simulação a partir de uma distribuição, incorporando essa simulação como uma distribuição limitante de uma cadeia de Markov até que esta cadeia se aproxime de um ponto de equilíbrio. Antes de compreender a problemática da simulação por meio das cadeias de Markov, ou MCMC, é importante que a fundamentação e as propriedades das cadeias de Markov sejam bem compreendidas.

A introdução das cadeias de Markov nos esquemas de simulação é de grande

interesse, pois permite o manuseio de distribuições complicadas, como as que surgem nos modelos de grandes dimensões, mencionados logo acima. Há também a questão do trabalho extra envolvido na simulação de um único valor pelo método MCMC: é necessária a simulação de uma sequência completa de valores de uma cadeia até que esta atinja seu equilíbrio, e somente o valor de equilíbrio pode ser tomado como um valor simulado da distribuição limitante. Os resultados também existem e são análogos aos da lei dos grandes números e do teorema do limite central para as cadeias de Markov. Eles garantem que a maioria dos valores simulados de uma cadeia possa ser usada para gerar informações sobre a distribuição de interesse.

Há ainda a questão de como construir uma cadeia de Markov cuja distribuição limite seja exatamente a distribuição de interesse, ou seja, a distribuição de todas as quantidades desconhecidas do modelo. Note que isso não é apenas possível, mas também existem grandes classes de ferramentas e técnicas que fornecem essas respostas.

Uma dessas técnicas, conhecida como "amostragem de Gibbs", baseia-se em uma cadeia de Markov cuja dependência do elemento predecessor é governada pelas distribuições condicionais que surgem do modelo. Entretanto, muitos modelos têm uma distribuição conjunta consideravelmente difícil de ser obtida, embora por construção, (algumas) das suas distribuições condicionais sejam relativamente simples de serem calculadas. Assim, a estratégia da amostragem de Gibbs explora esse ponto, sendo capaz de fornecer soluções simples para muitos problemas.

Existem diversas maneiras de se utilizar o MCMC em situações arbitrárias. A principal restrição que deve ser levada em consideração é quanto a computação eficiente. Entende-se por eficiência computacional como a medida de "facilidade" com que uma amostra simulada é obtida. Esta medida leva em consideração muitos aspectos importantes, como a escolha das distribuições condicionais a serem aplicadas no modelo, a necessidade de transformações das quantidades e o custo computacional, que está diretamente atrelado ao tempo necessário para que uma simulação atinja a estabilidade das soluções obtidas.

Outra estratégia interessante é dada pelos algoritmos Metropolis-Hastings. Estes algoritmos são baseados em uma cadeia de Markov cuja dependência do elemento antecessor é dividida em duas partes: uma proposta e uma aceitação da proposta. As propostas sugerem em um próximo passo arbitrário da trajetória da cadeia de Markov, enquanto a aceitação garante que a direção das limitação apropriada seja mantida, rejeitando movimentos indesejados que podem ser tomados durante o processo. Essas duas partes que constituem o algoritmo fornecem uma solução quando até mesmo as distribuições condicionais de interesse são complicadas, embora seu uso não seja restrito somente a esses casos. Os algoritmos de Metropolis-Hastings podem ser apresentados em variadas formas.

Algumas dessas formas podem ser categorizadas como generalizações da amostragem de Gibbs.

Deve ser destacado também, que a inferência bayesiana não necessariamente é completada após resumir informações sobre quantidades desconhecidas de um dado modelo. Podem haver outras operações relevantes para se realizar dentro de uma modelagem abrangente e compreensiva, como a própria avaliação do modelo em uso e comparações entre modelos, algo que envolveria a aplicação de mais de um modelo. É de particular interesse neste cenário, a consideração conjunta de um número (grande) de possíveis modelos. A inferência bayesiana (e MCMC) podem ser arranjados para lidar com essas questões, ainda que modelos alternativos também possam ser utilizados como dispositivos auxiliares no projeto de um método MCMC para um modelo específico, de modo garantir a eficiência do processo por meio da redundância.

Capítulo 3

Fundamentos Bayesianos

> "The fundamental problem towards which the study of statistics is adressed is that of inference. Some data are observed and we wish to make statements, inferences, about one or more unknown features of the physical system which gave rise to these data." (O'HAGAN, 1994)

Neste capítulo serão apresentados os principais elementos no que se refere ao processo de tomada de decisão por meio do paradigma Bayesiano. Dessa maneira, é possível trazer ao processo de modelagem a incorporação de conhecimentos prévios adicionados aos já obtidos e provenientes dos dados coletados, o que auxilia na tomada de uma decisão. Isso é combinando à quantificação da incerteza que está presente na informação disponível que será adicionada à conhecimentos já existentes (*a priori*).

3.1 O dilema da estatística

A análise estatística pode ser dividida em dois grandes grupos de processos: estatística descritiva das observações e o procedimento inferencial. A análise descritiva visa sumarizar os resultados por meio de tabelas e gráficos. Enquanto que a inferencial visa descrever os dados por meio de um modelo estatístico que estima e testa hipóteses sobre os resultados aferidos pelo modelo adotado. Assim, o objetivo principal da estatística, muitas vezes, está em fazer inferências ou predições de novas observações, bem como na descrição do fenômeno observado. Dessa forma, esse procedimento de estimação estatística poderá ser subdividido em dois principais paradigmas: Frequentista (também conhecidos como clássico) e Bayesiano (subjetivo).

A inferência clássica obtém por meio da indução, as informações referentes ao conhecimento obtido. A inferência Bayesiana obtém as informações por meio da dedução do processo de conhecimento. O primeiro paradigma decorre apenas do cálculo de

probabilidades, enquanto o outro paradigma transforma o modelo em um problema inverso de probabilidade, utilizando o teorema de Bayes.

As conclusões ou informações retiradas dos dados inferidos por meio da modelagem (por vezes traduzida em parâmetros), enquadram-se, em geral, na lógica indutiva e a justificativa via processo indutivo é um dos problemas mais controversos da filosofia. Enquanto do ponto de vista Bayesiano, diferentes graus de incerteza são introduzidos no processo de modelagem e representados por meio de modelos probabilísticos para os parâmetros.

A inferência Bayesiana fundamenta-se por duas motivações: define-se como *inferência*, o ato de analisar em termos probabilísticos um conjunto de fenômenos, visando, com isso, desenvolver um conhecimento científico. Enquanto a *decisão* pode ser definida como a ação prática realizada por meio do processamento da informação provida a partir do conhecimento.

3.2 Inferência Bayesiana

Antes de aprofundar na discussão teórica específica sobre a temática da inferência bayesiana, os conceitos de probabilidade e coerência devem ser definidos. Compreende-se por probabilidade, um valor aferido que descreve um grau de confiança de uma dada proposição que não se sabe, *a priori*, se é verdadeira ou falsa (FINETTI, 1937). A coerência é a representação da incerteza de algo ser razoável, satisfazendo os axiomas da probabilidade (RAMSEY, 1926).

A seguir, serão elucidados os principais resultados acerca da teoria de probabilidades, utilizando a roupagem da coerência, no contexto das apostas finitas (para maiores detalhes consultar Esteves e Stern (2017)). Considere um espaço de probabilidades $(\Omega, \mathscr{F}, P)$ em que Ω define um espaço fundamental não-vazio, A é um conjunto de elementos tal que $A \subseteq \Omega$, $\mathscr{F}$ é uma família dos acontecidos dotados de probabilidade (álgebra ou *sigma*-álgebra) e P é uma medida de probabilidade.

Adicionando formas de incertezas no contexto de novos aprendizados, o sistema de apostas poderá incorporar a extensão do conceito de probabilidade condicional. Os axiomas de probabilidade são: i) Não-negatividade; ii) Aditividade; iii) Normalização; e iv) Multiplicação. Ou seja:

- Para todo A, $0 \leq P(A) \leq 1$.
- Se $P(\Omega) = 1$.

- Se existem A e B disjuntos, $P(A \cup B) \neq P(A) + P(B)$.
- Se existem A e B não disjuntos, $P(A \cap B) = P(B)P(A|B)$.

De acordo com os resultados apresentados e utilizando a aditividade e a probabilidade condicional, será obtido:

$$P(B) = \sum_i P(A_i \cap B) = \sum_i P(B|A_i)P(A_i) \tag{3.1}$$

Note que:

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B) \tag{3.2}$$

Derivando assim o teorema de Bayes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)} \tag{3.3}$$

Uma interpretação que decorre deste teorema é que dadas as situações antecedentes (*a priori*, A_i), adicionado a informação do acontecimento (conjunto de dados, B), o investigador irá obter as suas probabilidades revisadas (a posteriori, $A_i|B$). Isto é, na próxima iteração a probabilidade revisada anteriormente (a posteriori) passará a ser a nova probabilidade *a priori*.

3.2.1 Atualização de incerteza

O grau de incerteza será refletido por meio de uma probabilidade, associando esta a diferentes proposições. Note que essa tarefa não é trivial, visto que a obtenção dos dados pode considerar várias hipóteses. Seja uma probabilidade da hipótese condicionada os dados observados,

$$P(\text{Hipótese} \mid \text{Dados})$$

Utilizando o teorema de Bayes poderemos desenvolver essa probabilidade em termos do aprendizado sobre as hipóteses consideradas utilizando os dados obtidos. Ou seja,

$$P(\text{Hipótese} \mid \text{Dados}) = \frac{P(\text{Hipótese})P(\text{Dados} \mid \text{Hipótese})}{P(\text{Dados})}$$

uma vez que

$$P(\text{Dados}) = \int_{\text{Hipótese}} P(\text{Hipótese})P(\text{Dados} \mid \text{Hipótese})$$

Definido por Esteves e Stern (2017), "o modelo estatístico oferece uma tradução padronizada de problemas envolvendo incertezas para um espaço das probabilidades" que é composto por dois elementos aleatórios; dados observados (X) e uma quantidade desconhecida de interesse (θ), comumente chamado de "parâmetro do modelo". Utilizando o teorema de Bayes,

$$P(\theta \mid x) = \frac{P(\theta)P(x \mid \theta)}{P(x)} = \frac{P(\theta)P(x \mid \theta)}{\int_\Theta P(\theta)P(x \mid \theta)d\theta}$$

Note que X está associado a um vetor aleatório e x a um vetor observado, em relação aos dados coletados (já observados). Compreende-se como Θ o espaço de todas as quantidades desconhecidas, também chamado de espaço paramétrico. $P(\theta|x)$ é definido como a probabilidade a posteriori. $P(x)$ é conhecida como constante de normalização, não depende de θ e confere à probabilidade a posteriori o valor inteiro 1. $P(\theta)$ é conhecida como a probabilidade *a priori* e $P(x|\theta)$ é a verossimilhança.

A explicação está fundamentada sob a existência de uma função que descreve a relação entre os dados e essa quantidade desconhecida, chamada de função de probabilidade (ou densidade) conjunta ($P(x, \theta)$).

O que difere o paradigma Bayesiano do paradigma clássico, é a consideração dessa quantidade desconhecida de interesse (θ) como sendo uma variável aleatória, possibilitando, com isso, atribuir uma medida (distribuição) de probabilidade sobre a mesma. Assim, para determinar a distribuição de $\theta|X$, uma vez que os dados foram observados e $X = x$, é necessário observar os fatores que dependeram apenas de θ (ou seja, os demais serão considerados constantes), dizendo "proporcional a":

$$P(\theta|x) \propto P(\theta)P(x|\theta)$$

A hipótese de independência condicional é comumente encontrada em problemas estatísticos (EHLERS, 2017). Entende-se que duas variáveis sejam independentes condicionalmente quando, dado uma terceira variável, a ocorrência ou não-ocorrência das duas primeiras variáveis sejam eventos independentes se condicionadas à distribuição de probabilidade daquela terceira variável. Suponha que foi observado uma amostra $X = x_1, x_2, ..., x_n$, independente dado θ:

$$
\begin{aligned}
P(\theta|x_1) &\propto P(\theta)P(x_1|\theta) \\
P(\theta|x_1, x_2) &\propto P(\theta)P(x_1, x_2|\theta) = P(\theta)P(x_1|\theta)P(x_2|\theta) \\
&\vdots \\
P(\theta|x_1, x_2, ..., x_n) &\propto P(\theta)\prod_i P(x_i|\theta)
\end{aligned}
$$

Assim, pode-se dizer que a ordenação em que as observações são processadas pelo teorema de Bayes não é considerada relevante. Complementar a isso, fazer inferências sobre θ será relevante apenas ao que foi observado. Essa postulação é conhecida como princípio da verossimilhança (DEGROOT, 1986).

3.3 Decisão Bayesiana

A decisão Bayesiana está vinculada à distribuição *a posteriori* do parâmetro θ, visto que a mesma contém toda a informação probabilística a respeito deste parâmetro. É natural verificar a função de densidade *a posteriori* como sendo a melhor descrição do processo de inferência (por exemplo, de modo gráfico).

A compactação da informação apresentada pela *a posteriori* é uma estratégia comumente utilizada. A estimativa pontual, sobre o(s) parâmetro(s) poderá(ão) ser sumarizada(s) por um único número. Estimativas intervalares e testes de hipótese também podem ser frequentemente aplicados na tomada de decisão Bayesiana. Essa é a escolha quanto ao parâmetro θ, elucidado sob a ótica da teoria da decisão.

A teoria da decisão tem por objetivo fazer uso ativo do conjunto de ações possíveis, auxiliando na escolha da melhor ação a ser tomada. Três tipos de espaços são utilizados na resolução dessa problemática: i) Espaço paramétrico (Θ); ii) Espaço experimental (Ω); e iii) Espaço das possíveis ações (A).

Uma regra no contexto da tomada de decisão (tratada também como "utilidade") representa o quanto é desejável obter a possibilidade θ, uma vez feita uma escolha a. Ou seja, a utilidade será um par composto por uma alternativa ($a \in A$) e uma possibilidade ($\theta \in \Theta$). Assim, para cada decisão a e para cada possível valor do parâmetro θ (dado que o mesmo é desconhecido), é possível associar uma perda $L(a, \theta)$ assumindo valores positivos. Logo, o risco de uma regra de decisão, denotado por $R(a)$, será a perda esperada *a posteriori*, ou seja, $R(a) = E_{\theta|x}(L(a, \theta))$. A regra de decisão mínima é denominada regra de Bayes ou risco de Bayes.

3.3.1 Estimadores de Bayes

Dentro do arcabouço das teorias da estimativa e da decisão, pode-se dizer que um estimador de Bayes, ou uma ação de Bayes, é um estimador ou regra de decisão que minimiza o valor esperado posterior de uma função de perda (ou seja, a perda esperada posterior). Equivalentemente, esse estimador maximiza a expectativa posterior de uma função de utilidade. Uma maneira alternativa de formular um estimador dentro da estatística bayesiana é a estimativa máxima *a posteriori*.

Um problema de inferência será iniciado, conforme apresentado na seção 2.1, a partir dos valores observados na amostragem, com o objetivo de estimar o valor do parâmetro θ. Considere uma amostra aleatória $X_1, X_2, ..., X_n$, tomada de uma distribuição com função de densidade de probabilidade $P(x|\theta)$, sendo θ desconhecido. Pode-se dizer que $\theta \in \Theta$. Então, os valores obtidos por um estimador também serão $\delta(X) \in \Theta$.

Cada valor de θ calculado por meio da estimativa a poderá ter associado a ele uma perda $L(a, \theta)$, no qual uma pequena distância representará um bom estimador. Com isso, tem-se que a alta probabilidade de que um bom estimador apresente um erro $\delta(X) - \theta$ será próxima de zero. Assim, a regra de Bayes consiste em escolher a estimativa que minimiza a perda esperada, que é dada por,

$$E[L(a, \theta)|x] = \int L(a, \theta)P(\theta|x)d\theta$$

As funções de perda simétricas são frequentemente adotadas na literatura, sendo as principais: a função de perda quadrática ($L(a, \theta) = (a - \theta)^2$) e absoluta ($L(a, \theta) = |a - \theta|$). Pelo fato da resposta para um problema de estimação ser um número único, a média *a posteriori* é o estimador ótimo para a perda quadrática e a mediana *a posteriori* é o estimador ótimo para a perda obtida pelo desvio absoluto. Note que aqui não está sendo indicado o quanto se tem de certeza de que o parâmetro está próximo desse número, mas sim há a sumarização da informação contida na distribuição em um único número.

O uso da distribuição *a posteriori* em uma estimativa pontual poderá ser aplicado sob o conceito de intervalo de credibilidade (também conhecido por intervalo de confiança Bayesiano), que é baseado na distribuição. A principal restrição da estimação pontual é que ao se estimar um parâmetro por meio de um valor numérico único, então toda a informação presente na distribuição *a posteriori* será resumida por meio deste número. Assim, é importante associar alguma informação sobre o quão precisa será a especificação deste número. Para os estimadores que serão vistos aqui, a literatura aponta como as medidas de incerteza mais usuais: a variância (ou coeficiente de variação para a média a

posteriori); a medida para a moda *a posteriori* da informação observada de Fisher; e a distância para a mediana *a posteriori* entre quartis.

Com isso, tem-se C como sendo um intervalo de credibilidade de $100(1-\alpha)\%$, ou nível de credibilidade (ou confiança) $1-\alpha$, para $\theta \geq P(\theta \in C) \geq 1-\alpha$ (EHLERS, 2017).

3.3.2 Valor da informação

Existem três tipos de informação: i) a informação conhecida inicialmente (*a priori*); ii) a informação dada por novas observações; e iii) a informação que se obtém de acordo com as consequências das ações. A quantificação probabilística de crenças *a priori* visa incorporar dentro da análise, a informação apriorística que se tem sobre os dados, por exemplo, o especialista no domínio de aplicação. Além do mais, essas distribuições *a priori* também poderão ser conjugadas (adoção de uma forma funcional) ou não-informativas (conhecimento *a priori* "vago" ou "pouco significativo" quando comparado à informação amostral). Para maiores detalhes consulte Paulino, Turkman e Murteira (2003).

Dada a importância de aplicações em modelos normalizados, em problemas de uma ou várias amostras ou mesmo de regressão, serão comparadas as implicâncias de *a priori* conjugadas e não-informativas. Como exemplo, considere uma amostral aleatória $X_1, X_2, ..., X_n$ proveniente da distribuição normal com média desconhecida (μ) e variância conhecida (σ^2), ou seja, $X|\mu \sim N(\mu, \sigma^2)$. Seja ainda $\tau^2 = \sigma^{-2}$:

$$
\begin{aligned}
P(\mu|x) &\propto P(\mu)P(x|\mu) \\
&= exp\left(-\frac{\tau_0^2(\mu-\mu_0)^2}{2}\right) exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\
&= exp\left(-\frac{\tau_0^2\mu^2 - 2\tau_0^2\mu\mu_0 + \tau_0^2\mu_0^2}{2}\right) exp\left(-\frac{\tau^2x^2 - 2\tau^2x\mu + \tau^2\mu^2}{2}\right) \\
&\propto exp\left(-\frac{\tau_0^2(\mu^2 - 2\mu\mu_0) + \tau^2(\mu^2 - 2x\mu)}{2}\right) \\
&= exp\left(-\frac{\mu^2(\tau_0^2+\tau^2) - 2\mu(\tau_0^2\mu_0 + \tau^2x)}{2}\right)
\end{aligned}
$$

com $a = \sqrt{\tau^2 + \tau_0^2}$ e $b = \frac{\tau_0^2\mu_0 + \tau^2x}{\sqrt{\tau^2+\tau_0^2}}$, então

$$
\begin{aligned}
P(\mu|x) &\propto exp\left(-\frac{a^2\mu^2 - 2ab\mu}{2}\right) \\
&\propto exp\left(-\frac{a^2\mu^2 - 2ab\mu + b^2}{2}\right) \\
&= exp\left(-\frac{(a\mu - b)^2}{2}\right) = exp\left(-\frac{a^2(\mu - \frac{b}{a})^2}{2}\right)
\end{aligned}
$$

com isso, a variância *a posteriori* será a combinação entre a variância *a priori* e a variância dos dados, assim como a média *a posteriori* será uma média ponderada:

$$
\mu|X \sim N\left(\frac{\tau_0^2\mu_0 + \tau^2 x}{\tau_0^2 + \tau^2}, \tau_0^2 + \tau^2\right)
$$

Esse modelo também é chamado de normal-normal, ou seja, quando *a priori*os dados forem provenientes de distribuições normais, implica que *a posteriori* o modelo terá a forma analítica fechada como a de uma distribuição normal.

Suponha dois cenários. Seja o parâmetro real $p = 0.15$ e considere *a posteriori* de uma função *a priori* não-informativa, em que basta imputar uma variância tendendo ao infinito ($\sigma_0^2 \to \infty$, na prática um número bem grande). O outro cenário sendo uma *a priori* com $Beta(5, 5)$ especificada. A Figura 3.1 compara graficamente as curvas *a priori*, em cada cenário, e como *a posteriori* está sendo influenciada em cada um dos casos.

Uma observação importante é que *a posteriori* estará sempre entre a curva *a priori* e a curva da verossimilhança, ou seja, sendo o resultado da combinação delas. Para o caso da *priori* não informativa é possível notar que *a posteriori* é guiada quase que exclusivamente pelos dados (verossimilhança). Já no caso da *priori* informativa, a sua influência será condicionada a força da verossimilhança.

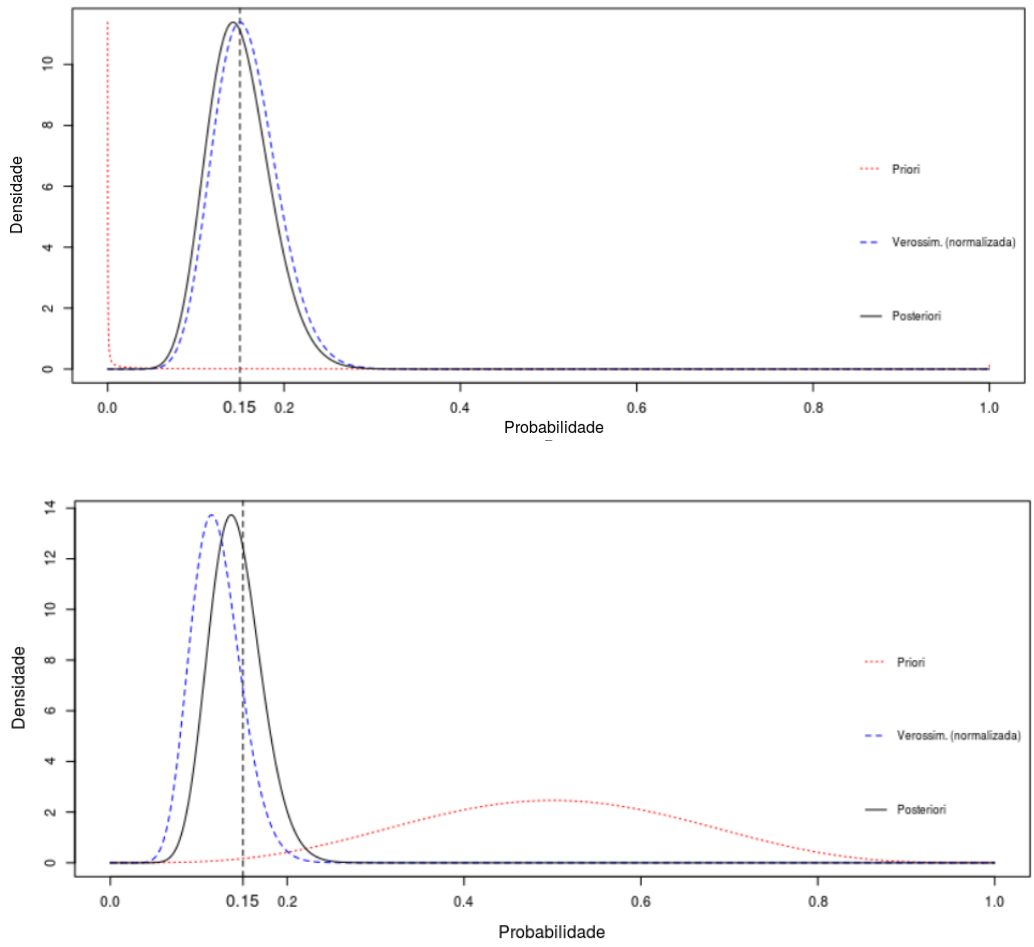

Figura 3.1 – Densidade com *a priori* não-informativa (topo) *vs.* informativa (inferior).

Ou seja, a força que guiará *a posteriori* proveniente da *priori* está condicionada ao tamanho da amostral (tamanho da verossimilhança), quanto maior o tamanho amostral menor influência irá vir da *priori*.

Até o momento, foi citado um modelo em que era possível calcular analiticamente a distribuição *a posteriori* para o parâmetro do modelo estatístico. Para isso, utiliza-se a expressão obtida a partir do Teorema de Bayes. Geralmente não é possível calcular diretamente ou aproximar a expressão $P(\theta|x)$ partindo de métodos determinístico de integração (especialmente se o parâmetro tiver alta dimensionalidade). Assim, para realizar uma análise Bayesiana apropriadamente, será necessário desenvolver outros modos de avaliação *a posteriori*, que serão discutidos no próximo capítulo.

Capítulo 4

Métodos de Amostragem - MCMC

> "Statistics needs a foundation by which I mean a framework of analysis within which any statistical investigation can theoretically be planned, performed, and meaningfully evaluated. The word any and theoretically are key, in that the framework should apply to any situation but may only theoretically be implementable. Practical difficulties or time limitations may prevent complete (or even partial) utilisation of such framework, but the direction in which truth could be found would at least to be known." (BERGER, 1984)

Neste capítulo serão discutidos os métodos de simulação de Monte Carlo, bem como os algoritmos MCMC definidos como rotinas de simulação de Monte Carlo via cadeias de Markov. Esses algoritmos têm sido intensamente utilizados em Estatística Bayesiana, área geralmente dependente de processamento computacional intenso. Também serão discutidos quais, entre os algoritmos apresentados, são os mais precisos, eficientes e viáveis sob o ponto de vista computacional.

O principal foco deste estudo é a média *a posteriori*, em que os resultados calculados de modo analítico são então comparados com aproximações e simulações obtidas para vários parâmetros e tamanhos amostrais. De modo geral, a simulação de Monte Carlo e os algoritmos baseados em MCMC são rápidos e eficientes, porém estes últimos possuem construção mais elaborada e requerem algumas particularidades. No decorrer do texto serão apresentadas algumas ressalvas importantes sobre os métodos que foram levantadas.

4.1 Computação Bayesiana

Em várias situações, que podem permear praticamente todas as áreas do conhecimento, ter definida ou mesmo supostamente conhecida a distribuição de probabilidades ou o delineamento probabilístico de um conjunto de dados, é de grande importância. Com base em tal conhecimento, calculam-se probabilidades de ocorrência de eventos como, por

exemplo, o bloqueio de um servidor em uma rede baseada em filas (Barbosa e Cruz, 2008), estimativas de confiabilidade, estatísticas de interesse e muitos outros assuntos relativos aos dados ou instrumentos modelados.

Algumas situações exigem a atualização dos modelos definidos *a priori*. Esse novo ajuste é feito por meio da necessidade de se promover uma modelagem mais adequada, ou mais precisa dos dados. Ou mesmo, devido às modificações que ocorreram durante o processo e que se fazem significativas em todos os cálculos durante a modelagem. Nessas situações é preciso calcular uma integral que pode ser intratável analiticamente, ou também é necessário estimar uma quantidade amostral de uma nova distribuição que ainda não é previamente conhecida, ou mesmo estimar parâmetros desconhecidos. Para facilitar essas tarefas e torná-las menos árduas, podem ser aplicados métodos numéricos específicos que são apropriados para cada situação em que se está trabalhando.

Assim, os métodos de simulação de Monte Carlo, bem como os algoritmos de simulação de Monte Carlo via cadeias de Markov, vêm demonstrando suas capacidades de aplicação em Estatística Bayesiana, área que quase sempre necessita de cálculos intensos durante seus processos de análise. Serão discutidos quais desses métodos são mais coerentes, precisos e viáveis no cálculo da média *a posteriori*. Serão utilizados nos algoritmos dois tamanhos de amostra, além de diferentes valores para a probabilidade de sucesso θ. Os resultados obtidos pelos métodos de simulação serão comparados com os valores analíticos exatos, que são calculados a partir da distribuição *a posteriori*.

4.1.1 Métodos de Simulação Monte Carlo

Para ilustrar os métodos de simulação de Monte Carlo (Gamerman, 1997), seja:

$$h(\theta) = \frac{f(\theta)}{\int f(\theta) d\theta}$$

uma função a partir da qual se quer obter amostras de $h(\theta)$ sem resolver a integração. Suponha que uma amostra possa ser facilmente gerada a partir de uma função $g(\theta)$, chamada de função de referência, de quem é desejado obter uma amostra de $h(\theta)$, sendo que a função $h(\theta)$ deve ser positiva e padronizável. Dado o cenário da situação anterior, é possível gerar uma amostra $h(\theta)$, sabendo apenas a forma funcional de $f(\theta)$ e tendo uma amostra de $g(\theta)$. As técnicas para tal geração amostral que podem ser aplicadas são os métodos da rejeição e SIR (*Sampling Importance Resampling* ou *Bootstrap* Bayesiano).

Métodos da rejeição

Suponha que exista m, tal que:

$$\frac{f(\theta)}{g(\theta)} \leq m, \forall \theta$$

Agora, o objetivo é encontrar $g(\theta)$, da qual é possível gerar amostras e, com $g(\theta)m$, seja $f(\theta)$. Para isso, considere o algoritmo a seguir:

1. Gere $\theta_i \sim g(\theta)$ para $i = 1, 2, ..., T$;
2. Gere $u \sim uniforme\ (0, 1)$;
3. Se $u \leq \frac{f(\theta)}{mg(\theta)}$, se aceita θ_i, onde m é tal que $\frac{f(\theta)}{g(\theta)} \leq m$;
4. Volte ao passo 1.

O resultado obtido será de uma amostra de $h(\theta), \theta_1, ..., \theta_k$, $k \leq T$. A distribuição $g(\theta)$ será obtida e conterá caudas pesadas, além da amostra gerada que eventualmente será grande o suficiente para varrer todo o espaço paramétrico definido.

Método de SIR (*Sampling Importance Resampling*)

Se $m(12)$ não estiver disponível no procedimento anterior, uma possibilidade é usar o método de SIR, descrito no algoritmo a seguir:

1. Escolha uma função de referência;
2. Gere uma amostra $\theta_1, \theta_2, ..., \theta_n$ de $g(\theta)$;
3. Para cada $i = 1, 2, ..., n$ calcule:
   $w_i = \frac{f(\theta)}{g(\theta)}\ e\ q_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$
4. Selecione uma amostra $\theta_1^*, \theta_2^*, ..., \theta_n^*$ de $g(\theta)$ da amostra original $\theta_1, \theta_2, ..., \theta_n$ de $g(\theta)$, assumindo $p(\theta = \theta_i) = q_i$;
5. Gere $u \sim uniforme\ (0, 1)$ e observe se:
   - $u \in (0, q_1)$, escolhe θ_1.
   - $u \in (q_1, q_1 + q_2)$, escolhe θ_2.
   - $u \in (q_1 + q_2, q_1 + q_2 + q_3)$, escolhe θ_3.

As mesmas limitações que foram observadas no método da Rejeição também são válidas para o método de SIR.

4.1.2 Métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC)

A idéia central do método é a construção de uma Cadeia de Markov, da qual seja fácil gerar uma amostra e que haja uma distribuição de equilíbrio, $h(\theta)$, dada pela distribuição de interesse. Para tal, as seguintes condições devem ser observadas:

1. Seja $\theta_1, \theta_2, ..., \theta_p \sim p(\theta)$ com $p(\theta_1, \theta_2, ..., \theta_p)$ definido em $\Phi \subset \Re^p$;
2. Devemos supor uma Cadeia de Markov homogênea, irredutível e aperiódica, com espaço de estado Phi, e cuja distribuição de equilíbrio, $p(\theta)$, possa ser construída. Ou seja, deve ser possível construir uma Cadeia de Markov com probabilidade de transição invariante no tempo, onde cada estado possa ser visitado de qualquer outro, com um número finito de interações, e não pode haver estado absorvente. A distribuição estacionária deve ser $p(\theta)$;
3. As amostras das probabilidades de transição devem ser geradas facilmente.

A seguir, é apresentado um dos métodos mais comumente utilizados.

Algoritmo Metropolis-Hastings

Para descrever o algoritmo, suponha que a distribuição alvo é a distribuição *a posteriori* $(\theta|x)$ com $\theta = (\theta_1, \theta_2, ..., \theta_S)$. Além disso, considere que as condicionais completas *a posteriori* com $(\theta|\theta_{-1}, x)$, $i = 1, 2, ..., n$ estejam disponíveis, mas que, no entanto, não se saiba gerar amostras diretas de cada uma. Suponha também que as amostras de um valor novo de θ_i, são geradas a partir da distribuição proposta condicionalmente ao valor atual de θ_i, $q(\theta_i^{(p)}|\theta_i^{(a)})$, em que $\theta_i^{(p)}$ é o valor proposto e $\theta_i^{(a)}$ é o valor atual, para $i = 1, ..., n$. Para facilitar o entendimento, um esquema de amostragem é apresentado a seguir:

1. Inicialize $\theta^{(O)} = (\theta_1^{(O)}, ..., \theta_S^{(O)} ek = 1$
2. Obtenha um novo valor para $\theta^{(k)}$ a partir de $\theta^{(k-1)}$ por meio de gerações sucessivas de valores. Para $i = 1$ até S, faça:
   - Gere uma proposta para $\theta^{(k)}$ de $\theta^{(p)} \sim q(\theta_i \mid \theta_i^{(k-1)})$
   - Não rejeite a proposta de probabilidade de aceitação dada por: $\alpha = min1, \frac{p(\theta_i^{(p)})|\theta_i^{(a)})q(\theta_i^{(k-1)})|\theta_i^{(p)})}{p(\theta_i^{(k-1)})|\theta_i^{(a)})q(\theta_i^{(p)})|\theta_i^{(k-1)})}$ onde, $\theta_{-i}^{(a)} = \left(\theta_1^{(k)}, ..., \theta_{i-1}^{(k)}, \theta_{i+1}^{(k)}, ..., \theta_S^{(k)}\right)$
3. Faça $K = k + 1$, volte para o passo 2 e repita até alcançar uma convergência

O Algoritmo de Metropolis-Hastings é considerado bastante geralista e pode, ao menos em tese, ser desenvolvido com qualquer distribuição condicional *a posteriori* completa e para qualquer proposta. Porém, sob um aspecto prático, a escolha adequada da proposta é crucial para uma boa evolução do algoritmo, ou seja, para atingir sua convergência *a posteriori* na direção da distribuição. Algumas propostas mais comuns são apresentadas a seguir.

- Cadeias Simétricas

Esta situação ocorre quando a distribuição proposta é simétrica em torno da iteração anterior, isto é:

$$p(\theta_i^{(p)}) \mid \theta_i^{(k-1)}) = p(\theta_i^{(k-1)}) \mid \theta_i^{(p)})\alpha = 1, \frac{p(\theta_i^{(p)}) \mid \theta_i^{(a)}), x}{p(\theta_i^{(k-1)}) \mid \theta_i^{(a)})}$$

- Cadeias Independentes

As cadeias são independentes se a proposta não depender do passo anterior, ou seja:

$$q(\theta_i^{(p)}) \mid \theta_i^{(k-1)}) = q(\theta_i^{(p)})$$

e a probabilidade de aceitação seja dada por:

$$\alpha = [1, \frac{p(\theta_i^{(p)}) \mid \theta_i^{(a)})q(\theta_i^{(k-1)})}{p(\theta_i^{(k-1)}) \mid \theta_i^{(a)})q(\theta_i^{(p)})}]$$

Um cenário particular em que as cadeias são independentes é dado quando a distribuição proposta é a distribuição *a priori* para θ_i. Neste cenário, a probabilidade de aceitação é somente dada pela função de verossimilhança, ou seja:

$$\alpha = [1, \frac{x \mid p(\theta_i^{(p)}) \mid \theta_i^{(a)})}{p(x \mid \theta_i^{(k-1)}), \theta_i^{(a)})}]$$

Outro contexto particular de cadeias independentes se dá quando a distribuição proposta é a própria distribuição condicional completa *a posteriori*, isto é, $q(\theta_i^{(p)}) = q(\theta_i^{(p)} \mid \theta_i^{(a)}, x)$. Fazendo isso, a probabilidade de aceitação será igual a 1. Essa característica de geração da condicional completa e aceitação sempre por um algoritmo iterativo é a definição do *Gibbs Sampler*, que nada mais é do que uma particularização do algoritmo de Metropolis-Hastings.

Verificação de Convergência

Os métodos de MCMC são ótimas ferramentas para resolução de muitos problemas práticos na análise Bayesiana, bem como se verifica em muitas outras áreas do conhecimento. Porém, deve ser destacado que algumas questões relacionadas à convergência destes métodos ainda são terreno fértil para pesquisas.

Uma pergunta que pode surgir naturalmente é "quantas iterações o processo de simulação deve ter para garantir que a cadeia convergiu para um estado de equilíbrio?". A resposta para este questionamento talvez não possa ser dada facilmente, uma vez que a distribuição estacionária será, na prática, desconhecida. Entretanto, sempre será possível avaliar a convergência das cadeias ao se detectar problemas fora do período de preparação.

Uma análise de convergência dentro do contexto de métodos de simulação poderá ser feita preliminarmente, simplesmente ao se analisar os gráficos ou medidas descritivas dos valores simulados da quantidade de interesse θ. Dentre as representações gráficas mais frequentes, estão o gráfico da estimativa da distribuição *a posteriori* de θ e um gráfico de θ ao longo das iterações, além do histograma ou uma densidade de *Kernel*.

As estatísticas usualmente empregadas neste contexto são a média, o desvio padrão e os quantis (2,5%; 50%; 97,5%). Na detecção do período de *burn-in*, usam-se gráficos como a média ergódica (Rodrigues et al., 2009) e funções de auto-correlação. No gráfico da média ergódica, quando não houver variabilidade significativa, então houve convergência.

Capítulo 5

Filtro de Kalman

"All models are wrong, but some are useful" (George E. P. Box)

Em meados da década de 1960, o pesquisador Rudolph Kalman publicou um artigo descrevendo um procedimento dedicado à resolução de problemas lineares por meio da observação e estimação de seus estados (KALMAN, 1960). Inicialmente, apenas os problemas definidos em tempo discreto foram cobertos pelo pesquisador. Então, ainda na década de 1960 e agora em conjunto com Richard Bucy, foi desenvolvida uma variante do método definido anteriormente, agora para operar com informações em tempo contínuo.

O filtro de Kalman é um estimador instantâneo para estados determinados que é perturbado por ruídos, dentro de um sistema linear dinâmico. É considerado uma abordagem estatisticamente eficiente com relação a qualquer função de estimativa de erros quadrática (GREWAL; ANDREWS, 2015). Suas equações formam um processo recursivo que é capaz de reduzir a soma dos quadrados das diferenças entre os valores medidos e os valores estimados. Em geral, o filtro de Kalman é considerado como um propagador de distribuições probabilísticas, um vez que fornece uma caracterização completa do estado atual de um sistema, o que inclui suas referências passadas, sem a necessidade efetiva dos valores aferidos anteriormente.

O filtro de Kalman é utilizado nas mais diversas aplicações, desde a estimação de trajetórias de corpos celestes, de cursos de rios, em GPSs (*Global Positioning System*), na predição de commodities, em sinais de sensores e em grande parte dos dispositivos de telecomunicações. O filtro de Kalman ainda pode ser aplicado em modelagens de equações de estados, uma forma corriqueira para a descrição de aspectos físicos. Face ao significativo aumento do poder computacional nos últimos anos, o filtro de Kalman tem sido bastante aplicado na teoria de controle, incluindo aplicações relacionadas à

problemáticas não-lineares.

Destaca-se que os modelos de espaço de estados podem ser utilizados para modelar séries temporais univariadas ou multivariadas, na presença de não-estacionariedade, mudanças estruturais e padrões irregulares, com a finalidade de desenvolver possíveis aplicações dos modelos de espaço de estados na análise de séries temporais.

5.1 Breve Introdução ao filtro de Kalman

Esse filtro tem o objetivo de estimar o estado de um sistema linear que é corrompido por incertezas, que podem incluir ruídos de diversos tipos e também imprecisões na aquisição das amostras. Uma das principais vantagens descritas pelo filtro de Kalman é a sua capacidade de estimar com precisão os estados de um sistema que apresenta ruído Gaussiano branco (ABREU, 2008). Para tal, o filtro faz uso ativo de um método recursivo que é capaz de reduzir a soma dos quadrados das diferenças entre os valores reais e os valores que são estimados, considerando um processo de duas fases: a predição e a atualização.

A predição, que também é conhecida como a fase da estimativa *a priori*, estima o estado atual utilizando apenas os dados estimados até o passo anterior, o que não inclui os dados que foram observados no tempo atual. A seguir, o estado atual do sistema é incrementado com uma atualização que irá corrigir a estimativa *a priori* utilizando a observação do tempo atual. Esse processo resulta em uma estimativa combinada que é conhecida como estimativa *a posteriori* (AIUBE, 2005).

Dada a estrutura de um modelo de regressão (ou sistema linear), visto como equações de estado, isto é, segundo (KALMAN, 1960):

$$y_k = B_k x_k + z_k,$$
$$x_k = A_k x_{k-1} + w_k$$

sendo que o tempo atual é representado por k, a observação do k-ésimo tempo y_k, o ruído da medida z_k, o estado do k-ésimo tempo x_k, o ruído do processo w_k e as entradas de controle dos sistemas que serão A e B.

É importante destacar que em diversas aplicações que envolvem sistemas lineares dinâmicos, incluindo a aplicação desejada, são considerados ruídos Gaussianos brancos e não-correlacionados os ruídos w_k e z_k, que provém de possíveis erros de modelagem e

também erros dos sensores de medição que adquirem os dados, respectivamente. Assim, tais ruídos assumem média zero e covariância R_k e Q_k. Ou seja,

$$z_k \sim N(0, R_k)$$
$$w_k \sim N(0, Q_k)$$

Na fase da predição, são estimados os valores de interesse do sinal ($\hat{x}_{k/k-1}$) e também a matriz de covariância do erro ($P_{k/k-1}$), no tempo k, que são baseados nos valores do passo anterior, ou seja, no tempo $k-1$. A estimativa inicial do sinal de interesse, que também é chamada de *a priori*, baseia-se na estimativa no passo anterior do próprio sinal ($\hat{x}_{k-1}$), que é ponderada por uma matriz de transição de estados de controle do sistema sob aplicação, A_k. Ou seja, $\hat{x}_{k/k-1} = A_k\hat{x}_{k-1}$.

A matriz de covariância do erro *a priori* ($P_{k/k-1}$) se baseia no valor de si mesma estimado no passo anterior (P_{k-1}), na matriz A_k e também no ruído de processamento inerente ao sistema Q_k. Assim, $P_{k/k-1} = A_k P_{k/k-1} A_k^T + Q_k$. Então, são processados os cálculos de correção e também são estimados os valores do sinal de interesse ($\hat{x}_k$), bem como da matriz de covariância do erro (P_k), agora com a atualização destes erros a partir de um fator conhecido como "ganho de Kalman".

O ganho de Kalman (K_k) é um fator que está baseado no valor estimado da matriz de covariância do erro ($P_{k/k-1}$), sendo que este valor foi encontrado na fase da predição, numa matriz B_k resultante da equação de estados do sistema em processamento e também no ruído encontrado nas aferições do sistema (R_k), que é oriundo da aquisição do sinal medido.

$$K_k = P_{k/k-1}B_k'(B_k P_{k/k-1} B_k' + R_k)^{-1}$$

Pode-se dizer que a finalidade desse parâmetro é minorar o valor estimado na matriz de covariância do erro (P_k), com a estimativa o valor do sinal de interesse. Após essa etapa, uma nova estimativa do sinal de interesse é calculada ($\hat{x}_k$). Essa estimativa é conhecida como *a posteriori* e vem para atualizar a estimativa que foi encontrada no passo anterior do processo ($\hat{x}_{k/k-1}$), tendo por base o ganho de Kalman (K_k), a matriz B_k e a medição calculada no passo atual do processamento (y_k), relativa ao tempo k.

$$\hat{x}_k = \hat{x}_{k/k-1} + K_k(y_k - B_k\hat{x}_{k/k-1})$$

Finalmente, é calculada uma estimativa *a posteriori* nova que está relacionada com a matriz de covariância do erro (P_k). Deve ser levado em conta nessa fase do processo, a estimativa obtida na predição ($P_{k/k-1}$), a matriz B_k e também o ganho de Kalman, que tem por objetivo a minimização do erro da seguinte maneira:

$$P_k = (I - K_k B_k) P_{k/k-1}$$

A partir das equações descritas acima, pode-se dizer que quando ocorre um elevado nível de ruído das medições, o estimador irá conferir pouco crédito à medição atual no próximo passo da estimativa, uma vez que a medição que foi obtida no passo atual pode não ser confiável, o que acaba por tornar a convergência do sistema um tanto mais lenta. Por outro lado, quando há ruído considerado de pequena monta nas medidas, será dado muito crédito à medição atual na estimativa seguinte, dado que esta medição possui uma relevância considerada mais expressiva pois o ruído é pequeno, fato este que torna o sistema de convergência mais rápida.

O ruído de processamento é inserido no sistema propositalmente devido aos possíveis erros no modelo. Quando este ruído de processamento é considerado elevado, então será dado maior nível de crédito à medição atual na estimativa seguinte. Essa ação se deve ao fato de que os distúrbios no modelo utilizado são relevantes, o que torna a convergência geral do sistema um tanto mais rápida de ser obtida. Por outro lado, quando esse ruído é de menor monta, então pouco crédito será conferido à medição atual que ocorrerá na próxima estimativa, algo que confere maior relevância ao modelo, tornando a convergência do sistema mais lenta.

Como todas as distribuições relevantes são gaussianas, elas são completamente determinadas por suas médias e variações (CAMPAGNOLI; PETRIS; PETRONE, 2009). Campagnoli, Petris e Petrone (2009) afirmaram que a solução do problema de filtragem para modelos lineares dinâmicos (DLMs) pode ser dada pelo célebre filtro de Kalman.

O filtro de Kalman via máxima verossimilhança não explica as incertezas sobre a precisão dos parâmetros do modelo. Isso pode ser um problema, já que a precisão em modelos de espaço de estados são sempre observados, adicionado à restrições numéricas do MLE. Assim, a abordagem bayesiana oferece uma maneira natural de lidar com a incerteza dos parâmetros em um modelo de espaço de estados.

Na seção a seguir, será apresentada uma simulação do espaço de estados via filtro de Kalman, combinado com os conceitos de inferência bayesiana utilizando métodos de reamostragem de MCMC, já discutidos nas seções 3 e 4.

5.2 Espaço de estados via Metropolis-Hastings

As exemplificações que serão apresentada aqui foram estimadas utilizando a linguagem de programação *Python*, em que seus códigos-fonte poderão ser encontrados no anexo 7.2. Utilizou-se o pacote *tsa.statespace*, bem como o PyMC para estimar os parâmetros do modelo de espaço de estados com base no algoritmo Metropolis-Hastings. O modelo selecionado será o *Integrated Moving Average* (IMA), que considera a diferença estocástica de ordem 1 e uma média móvel de ordem 1, ou seja, um IMA(1,1).

Suponha uma série temporal $Y_T \equiv \{y_t\}_{t=0}^{T}$, modelada como processo de nível local:

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma\ \varepsilon^2) \tag{5.1}$$

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{5.2}$$

$$\tag{5.3}$$

Neste modelo, existem dois parâmetros desconhecidos, que serão coletados em um vetor ψ, de modo que: $\psi = (\sigma_\varepsilon^2, \sigma_\eta^2)$. Definem-se seus valores verdadeiros como segue (denotado com o subscrito 0):

$$\psi_0 = (\sigma_{\varepsilon,0}^2, \sigma_{\eta,0}^2) = (3, 10)$$

Finalmente, também deve-se especificar o $\mu_0 \sim N(m_0, P_0)$ para inicializar o filtro de Kalman. Seja uma série de tamanho $T = 1000$, conforme a figura 5.1

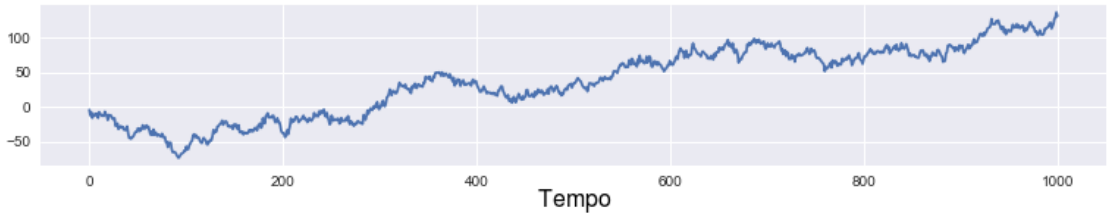

Figura 5.1 – Série simulada IMA(1,1) contendo 1000 observações

Acontece que será conveniente escrever o modelo em termos da precisão de ε, definido como $H^{-1} \equiv \sigma_\varepsilon^2$, e a razão das variâncias: $Q \equiv \sigma_\eta^2/\sigma_\varepsilon^2$ para que $QH^{-1} = \sigma_\eta^2$.

Então, os termos de erro podem ser escritos:

$$\varepsilon_t \sim N(0, H^{-1})$$
$$\eta_t \sim N(0, QH^{-1})$$

E os valores verdadeiros são:

$H_0^{-1} = 1/3$ e $Q = 10/3$

Para adotar uma abordagem Bayesiana a esse problema, assume-se que ψ é uma variável aleatória com objetivo de aprender sobre os valores de ψ baseado nos dados Y_T; na verdade, objetiva-se uma densidade $p(\psi|Y_T)$. Para fazer isso, usa-se a regra de Bayes para escrever:

$$p(\psi|Y_T) = \frac{p(Y_T|\psi)p(\psi)}{p(Y_T)}$$
$$\underbrace{p(\psi|Y_T)}_{\text{posterior}} \propto \underbrace{p(Y_T|\psi)}_{\text{probabilidade}} \underbrace{p(\psi)}_{\text{prior}}$$

O objeto de interesse é *a posteriori* e, para alcançá-lo, é necessário especificar uma densidade prévia para os parâmetros desconhecidos, bem como a função de verossimilhança do modelo.

Priori - Serão utilizados os seguintes *priors*:

Precisão - Como a precisão deve ser positiva, mas não tem limite superior teórico, usa-se um Gamma como *a priori*. Ou seja, $H \sim \text{Gamma}(\alpha_h, \beta_h)$, para ser específico, a densidade é escrita:

$p(H \mid \alpha_h, \beta_h) = \frac{\beta_h^{\alpha_h}}{\Gamma(\alpha_h)} H^{\alpha_h - 1} e^{-\beta_h H}$ e considerando os hiperparâmetros como $\alpha_h = 2, \beta_h = 2$. Nesse caso, têm-se $E(H) = \alpha_h/\beta_h = 1$ e também $E(H^{-1}) = E(\sigma_\varepsilon^2) = 1$.

Razão de Variações - De modo semelhante, a proporção de variâncias deve ser positiva, mas não há um limite superior teórico, então, novamente, usa-se um Gamma (independente) *a priori*:

$Q \sim \text{Gamma}(\alpha_q, \beta_q)$ com mesmos hiperparâmetros, então $\alpha_q = 2, \beta_q = 2$. Como $E(q) = 1$, o anterior é de variações iguais, então temos $E(\sigma_\eta^2) = E(QH^{-1}) = E(Q)E(H^{-1}) = 1$.

Estado inicial *a priori*- Como mencionado acima, o filtro de Kalman deve ser

inicializado com $\mu_0 \sim N(m_0, P_0)$. Será utilizado o seguinte aproximadamente difuso antes:

$$\mu_0 \sim N(0, 10^6)$$

Probabilidade - Para determinar os parâmetros, a probabilidade deste modelo pode ser calculada por meio da decomposição do erro de previsão usando uma aplicação das iterações do filtro de Kalman.

Simulação Posterior: Metropolis-Hastings - Uma opção para descrever o posterior é por meio dos métodos de simulação *a posteriori* do MCMC. O algoritmo de Metropolis-Hastings é simples e requer apenas a capacidade de avaliar as densidades prévias e sua probabilidade. Os antecedentes têm densidades conhecidas e a função de verossimilhança pode ser calculada utilizando os modelos de espaço de estados do pacote *Statsmodels tsa.statespace*. O pacote PyMC foi utilizado para simplificar a especificação de *priors* e *sampling*, no caso Metropolis-Hastings.

O pacote do espaço de estados facilita a especificação e avaliação de modelos de espaço de estados. Abaixo, é destacada uma nova classe LocalLevel. Entre outras coisas, ela herda do MLEModel, que é um método loglike utilizado para avaliar a probabilidade em vários parâmetros.

Estimação Bayesiana: Metropolis-Hastings Agora que os antecedentes e a probabilidade foram especificados, aplica-se um amostrador *PyMC Model* e MCMC. A Figura 5.2 sumariza o resultado proveniente da amostragem a partir da *a posteriori*. Observe que as médias das densidades posteriores estão próximas dos valores reais dos parâmetros.

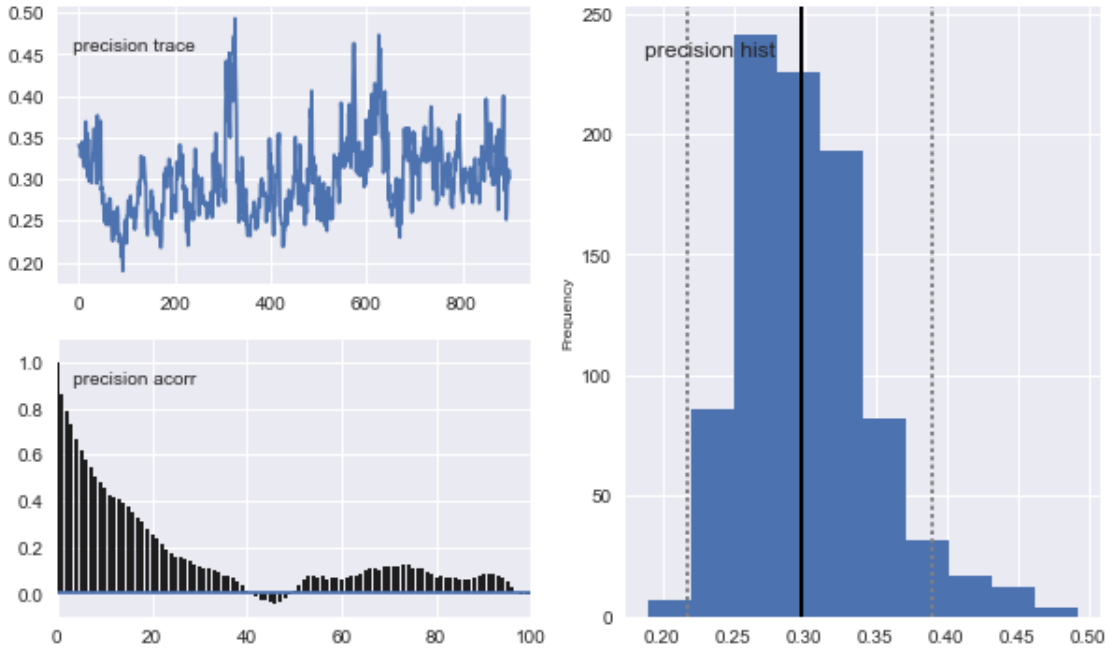

Figura 5.2 – Desempenho das cadeias de Markov e posteriori estimada

Capítulo 6

Estudo de aplicação em dados reais

> "Nature has established patterns originating in the return of events, but only for the most part. New illnesses flood the human race, so that no matter how many experiments you have done on corpses, you have not thereby immposd a limit on the nature of events so that in the future they could not vary." (Gottfried Leibniz)

Neste capítulo será descrita uma problemática considerando dados reais. O primeiro conjunto de dados registra a concentração mensal de CO_2 atmosférico, que foi amostrado desde 1960, enquanto o segundo conjunto de dados se refere à observações sobre a vazão anual do rio Nilo, em uma cidade no Egito, entre os anos de 1871 e 1970. O objetivo dessas análises são os de realizar uma descrição dos fenômenos observados e, com isso, realizar previsões (inferir comportamento futuro) para os mesmos. Fez-se uso do filtro de Kalman, utilizando o paradigma Bayesiano e algoritmos MCMC, como modelagem estatística. Também serão apresentados os resultados sobre o desempenho dessas cadeias, mais precisamente, informações sobre o custo computacional gerado por essa modelagem. Para análise foram desenvolvidos *scripts* utilizado a linguagem de programação Python.

6.1 Análise Empírica

6.1.1 Concentração de CO_2 atmosférico

Desde o final da década de 1950, o observatório de Mauna Loa, no estado norte americano do Hawaii, tem medido regularmente os níveis de CO_2 atmosférico. No final dos anos 1950, o pesqquisador Charles Keeling desenvolveu uma maneira bastante precisa de medir a concentração de CO_2 na atmosfera. Desde então, as medições de CO_2 foram registradas quase que continuamente no observatório de Mauna Loa. Maiores informações a

respeito do processo e dos dados registrados estão disponíveis em , acessado em 10/05/2019.

Assim, este trabalho considerou as médias mensais desde 1960 até 2017, como período de ajuste de modelo e os meses de Janeiro à Agosto de 2018 utilizados como teste de adesão na previsão fornecida pelo modelo.

Inicialmente, não se sabia ao certo sobre o impacto que a queima de combustíveis fósseis teria sobre o clima até o final dos anos 50. Os primeiros dois anos de coleta de dados mostraram que os níveis de CO_2 subiam e caíam, normalmente após as estações do verão e inverno (hemisfério norte). Este fato acompanhou o crescimento e a decadência da vegetação no hemisfério norte, conforme ilustra a figura 6.1. Além do mais, o mês de Maio apresenta quase sempre uma maior concentração de CO_2, em cada ano.

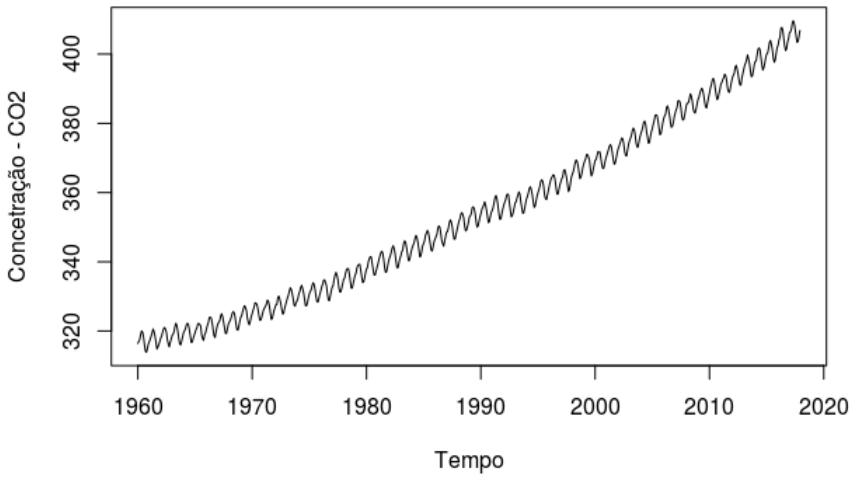

Figura 6.1 – Série histórica dos níveis de CO_2

É possível observar também, com o passar dos anos, que a constante tendência ascendente tornou-se cada vez mais orientada. Contando com mais de 70 anos de dados atmosféricos coletados, a curva *Keeling* pode ser um importante indicador climático.

Com evidências apontando para a presença de uma tendência de sazonalidade nas observações, e com fins descritivos da série temporal, adotou-se uma decomposição aditiva para a série do CO_2. Os resultados são apresentados na figura 6.2, em que o primeiro gráfico apresenta os dados coletados, o segundo está relacionado ao componente da tendência determinística, o terceiro ao componente da sazonalidade determinística e o quarto com os componentes não explicados (erro).

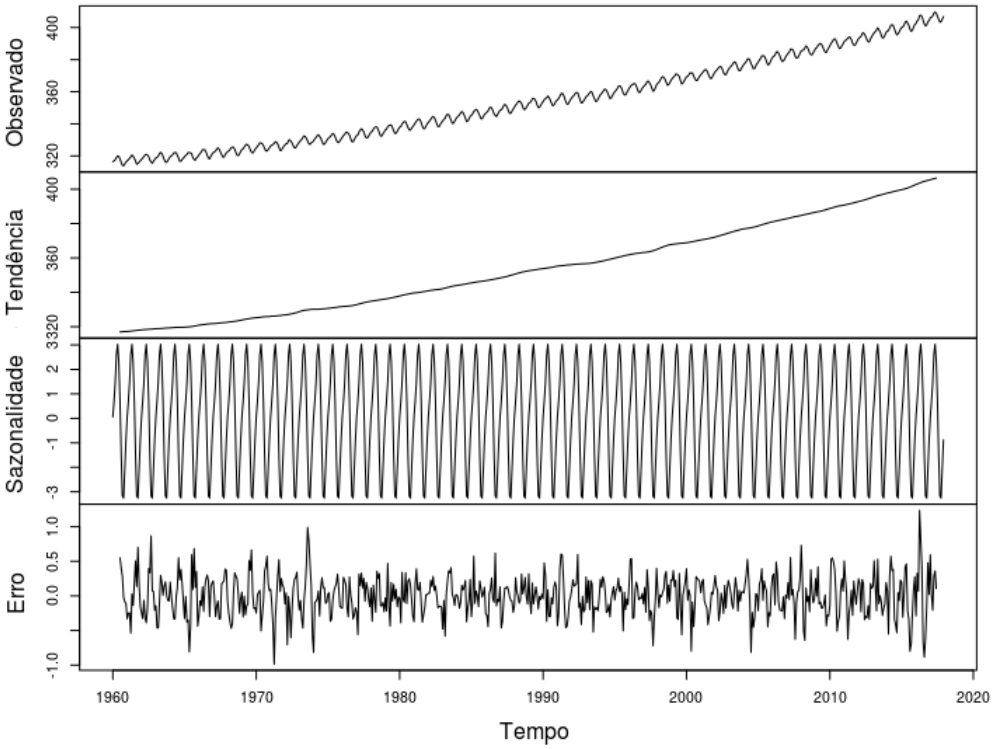

Figura 6.2 – Segregação da série temporal dos níveis de CO_2

Assim, para fins de previsão e considerando o período de 2000 à 2017, foram adotados os modelos clássicos para tratamento de séries temporais, sendo eles a metodologia Box & Jenkins (NAYLOR; SEAKS; WICHERN, 1972), ARIMA (autoregressivos integrados e de médias móveis) e o filtro de Kalman (modelo dinâmico), conforme mostram as figuras **??**.

Note que a série mensalizada das concentrações de CO_2 apresenta um comportamento relativamente suave (com ciclos e tendências aparentemente bem definidas). Desse modo, os dados são ditos "bem comportados". O modelo ARIMA se aproxima do padrão descrito pelos dados, entretanto algumas variações não são contempladas por esse modelo (observe as regiões próximas de alguns picos, na figura **??**). Enquanto o filtro de Kalman ficou bem ajustado na série, com a modelagem praticamente se sobrepondo na série histórica. A série observada foi modificada para o padrão em pontos, para melhor observação.

Um vez selecionado o modelo de estados, foram feitas previsões para o ano de 2018, conforme apresentado na figura 6.5.

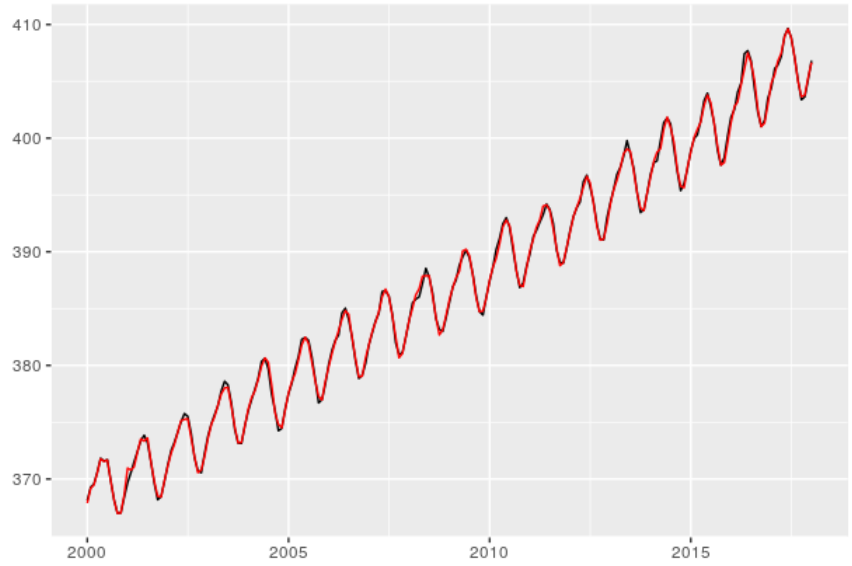

Figura 6.3 – Modelagem ARIMA

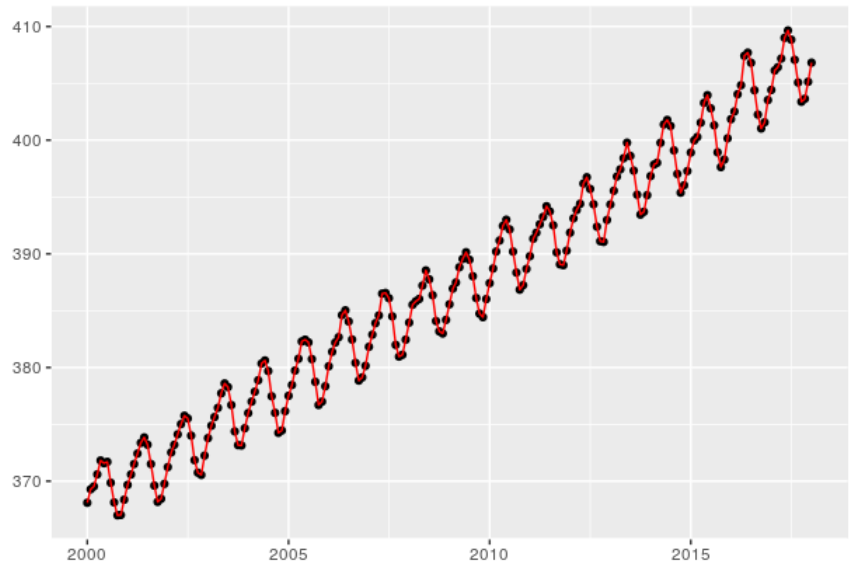

Figura 6.4 – Modelagem filtro de Kalman (direita)

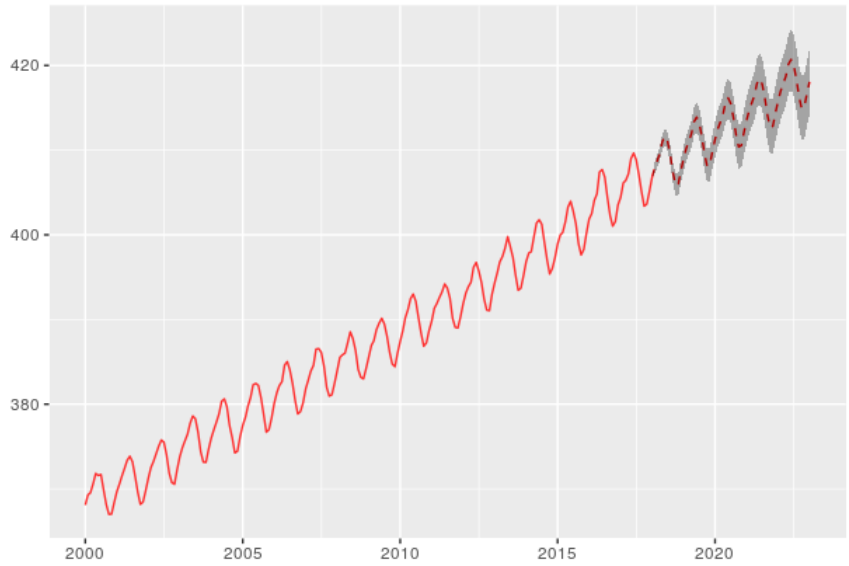

Figura 6.5 – Previsão para o ano de 2018

A seguir, outra série temporal irá comparar os modelos ARIMA e Kalman, agora considerando uma série contendo maiores irregularidade ao longo do tempo.

6.1.2 Vazão do rio Nilo

Os dados a serem explorados neste cenário são de medições do volume anual de descarga do rio Nilo, em Aswan, durante os anos de 1871 a 1970, apresentado em Cobb (1978). As medições são significativas em termos meteorológicos, apresentando evidências de uma possível mudança abrupta no regime de chuvas próximo à virada do século passado.

As informações registradas revelam que houve uma redução acentuada no volume anual após o ano de 1898. À primeira vista, é possível associar essa queda à presença de uma barragem que iniciou seu funcionamento em 1902, ou mesmo à correções aplicadas aos dados para remover esse efeito da barragem. Entretanto, evidências independentes apontam que a mudança é um problema real, também podendo ser encontrada nos registros pluviométricos de diversas estações meteorológicas localizadas em locais próximos aos trópicos, conforme mostra figura 6.6.

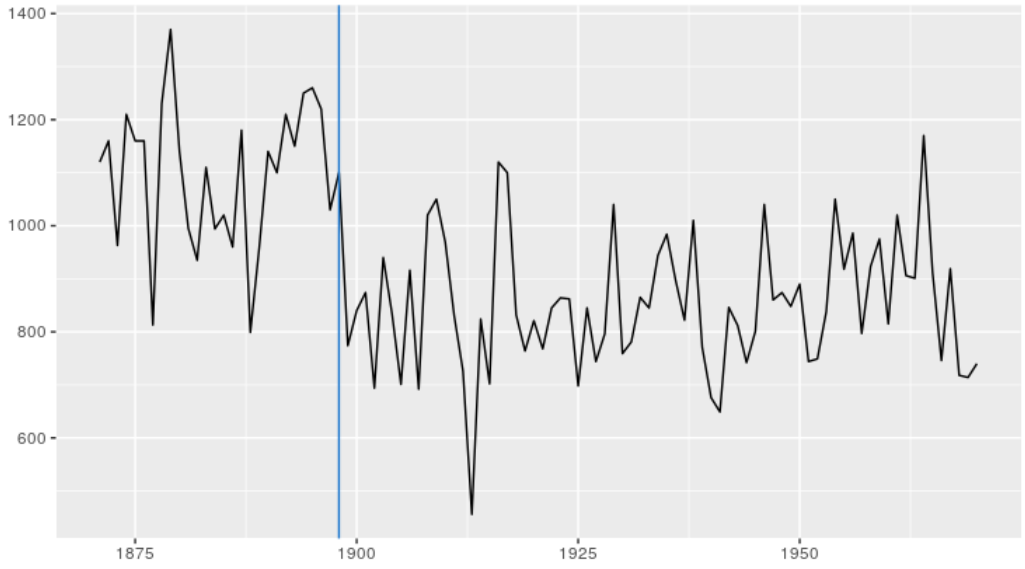

Figura 6.6 – Vazão do rio Nilo antes e depois da construção de uma barragem

Vale destacar que o exemplo trazido aqui foi inspirado pelo trabalho de Petris e Petrone (2011). Usualmente, as aplicações de suavização Bayesiana em modelos de espaço de estados utilizam o pacote dlm (conforme apresentou a subseção anterior).

A amostragem completa de Gibbs é fornecida para o caso básico de um modelo de espaço de estados univariados (de ordem 1), com variância de observação desconhecida e matriz de covariância de evolução diagonal desconhecida (dV e dW), porém com prévios de inversão de gamma independentes (função dlmGibbsDIG). As ferramentas para a análise da saída do MCMC também são fornecidas. Assim, uma amostragem de Gibbs da articulação posterior para as matrizes de covariância é implementada pela função *dlmGibbsDIG*.

Um amostrador de Gibbs pode ser obtido por amostragem iterativa a partir da distribuição condicional completa dos estados dos parâmetros, juntamente com os dados, e a condição completa dos parâmetros, o que deve levar em conta os estados e os dados. Utilizando o algoritmo *Forward Filtering Backward-Sampling* (FFBS) será possível amostrar *a posteriori*.

Foi adotado um limitante de 100.000 (cem mil) iterações para o algoritmo MCMC. Além disso, o método armazena uma amostra a cada 10 iterações (thin = 9), com a finalidade de reduzir a autocorrelação. O pacote dlm inclui algumas ferramentas para facilitar o diagnóstico básico de convergência da saída do MCMC. A Figura 6.7 mostra as médias das amostras em execução, bem como as funções de autocorrelação empíricas para as amostras de MCMC das variâncias. Após o descarte dos primeiros 1000 sorteios com o *burn-in*.

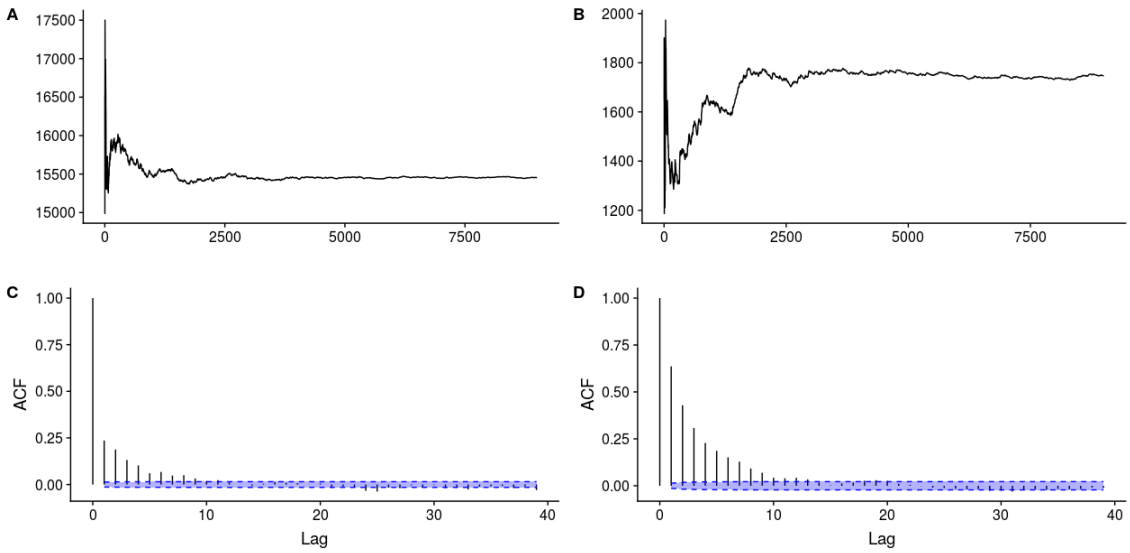

Figura 6.7 – Diagnóstico de convergência da saída do MCMC

Os gráficos de diagnóstico do MCMC auxiliam na análise visual do desempenho das médias amostrais e funções de autocorrelação das amostras (cadeia) MCMC. A Figura 6.8 ilustra as estimativas MCMC das densidades posteriores da variância de observação e da variância de evolução, respectivamente. Esses resultados sugerem, a partir de sua densidade posterior conjunta, que existe uma evidente e alta correlação, que reflete uma mistura bastante lenta do amostrador de Gibbs.

As estimativas Bayesianas das variâncias desconhecidas (dV e dW), com relação à perda quadrática, são dadas por suas expectativas posteriores, juntamente com a estimativa MCMC e com os erros padrão de Monte Carlo. A Figura 6.8 ilustra as densidades *a posteriori* para os parâmetros desconhecidos (variâncias).

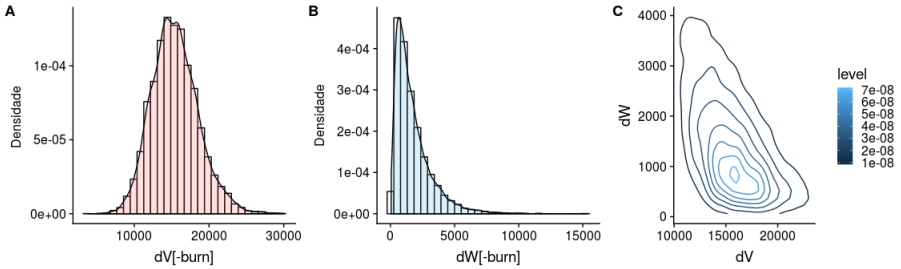

Figura 6.8 – Distribuições estimada *a posteriori* via MCMC

Uma vez que foram obtidas as densidades *a posteriori* dos parâmetros desconhecidos, ajustou-se o filtro de Kalman com essas estimativas. Este é o modelo dinâmico mais simples de ordem 1 para fins ilustrativos, conforme apresenta a figura 6.9, em que a linha contínua representa a média do processo estimado dado o modelo ajustado, e as linhas tracejadas representam o intervalo de credibilidade de 95% associado ao parâmetro da

média do processo.

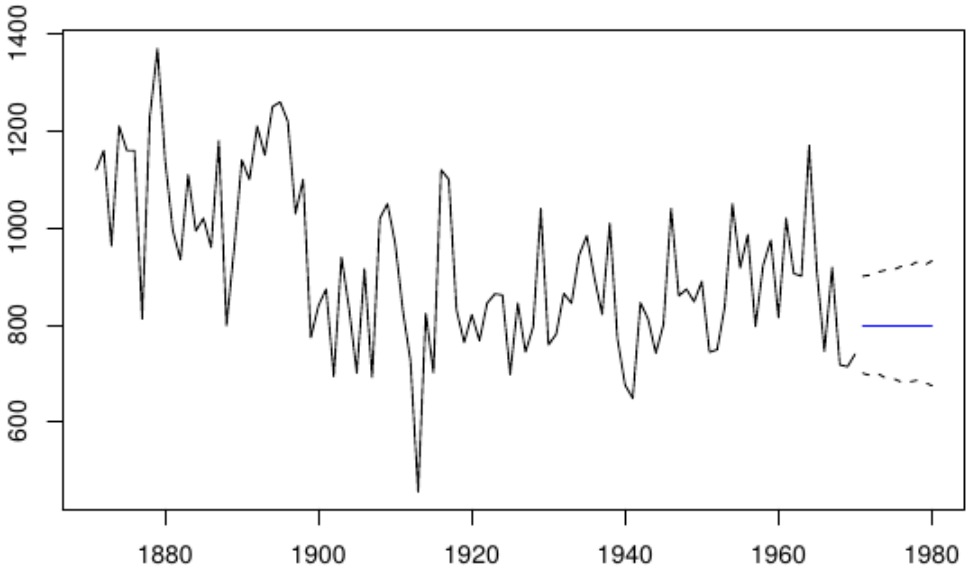

Figura 6.9 – Previsão da vazão do rio Nilo

Vale destacar que para a filtragem e previsão Bayesiana on-line, no entanto, o MCMC não é eficiente. Isso se deve ao fato de que q amostragem de Gibbs tem que ser executada novamente conforme novos dados se tornam disponíveis. Soluções fechadas são possíveis desde que tenham algumas premissas restritivas e pré-conjugadas, geralmente em conjunto com técnicas de fatores de desconto (WEST; HARRISON, 2006). Dessa forma, após ilustrar uma aplicação rudimentar do MCMC (via Gibbs) para a problemática de modelos dinâmicos, utilizou-se a análise conjugada Bayesiana com fatores de desconto, que mostrou-se computacionalmente mais eficiente.

Considere uma série temporal, como aquela saída de um sistema dinâmico que foi perturbado por ruídos aleatórios. Esses ruídos permitem uma interpretação da natureza dessa série temporal, por meio da combinação de vários componentes, como os de tendência, os sazonais ou mesmo os regressivos.

Ao mesmo tempo, esses componentes têm uma estrutura probabilística poderosa, oferecendo uma estrutura flexível para uma ampla gama de aplicações. Com isso, as inferências podem ser implementadas por algoritmos recursivos. O problema de estimativa e previsão é resolvido por meio do cálculo recursivo da distribuição condicional das quantidades de interesse, dadas as informações disponíveis (estrutura Bayesiana).

Dessa forma, um filtro de Kalman mais sofisticado foi adotado, obtendo as respostas a seguir. As estimativas de máxima verossimilhança para parâmetros desconhecidos de um modelo de espaço de estados arbitrário, dada a função de atualização e considerando

o método BFGS com valores estimados de variância inicial igual a 10,26. A Figura 6.10 apresenta o desempenho do filtro de Kalman ajustado.

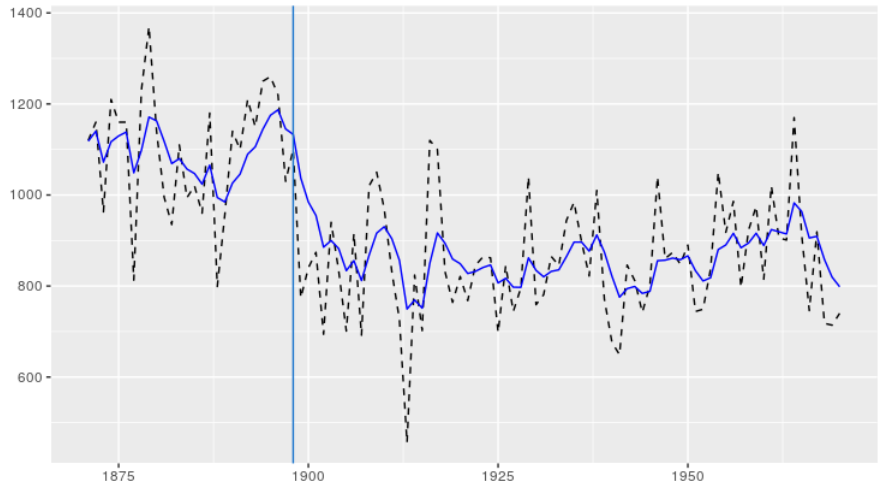

Figura 6.10 – Série filtrada (via Kalman) da vazão do rio Nilo

Também é possível adicionar a essa análise o alisamento da série, comparando a aderência do filtro de Kalman ao se considerar o processo bayesiano recursivo. A Figura 6.11 apresenta a estimação do estado da vazão do rio Nilo via filtragem e alisamento. A Filtragem estima o estado atual dos dados, levando em conta as observações até presente ponto. A suavização estima o estado a qualquer momento, condicionada em todos os dados.

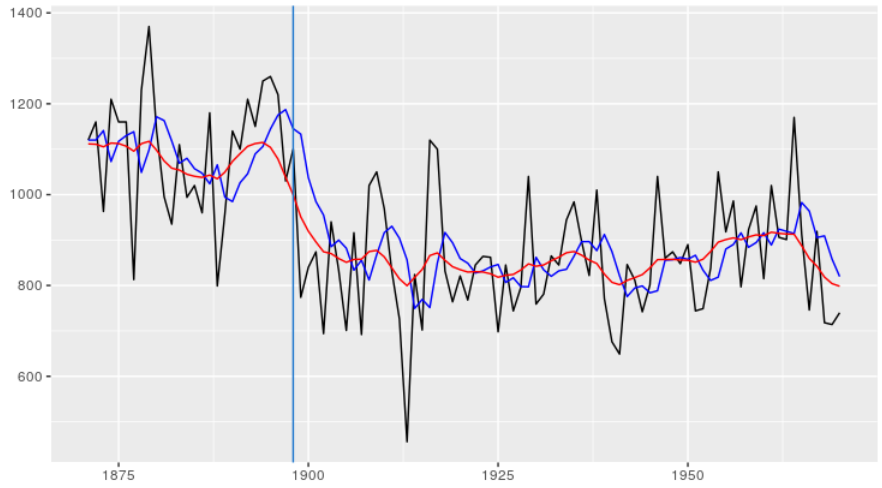

Figura 6.11 – Série filtrada e alisada (via Kalman) da vazão do rio Nilo

De acordo com resultados obtidos, é possível notar que as abordagens consideradas aqui são alternativas competitivas em modelar dados complexos, como os de séries temporais (sinais), até aquelas que apresentam não-linearidades.

Capítulo 7

Conclusão

7.1 Discussões

A literatura a respeito da inferência bayesiana teve seu início bem antes do métodos frequentistas, embora sua expressiva notoriedade tenha ocorrido apenas a partir dos anos 1950. A literatura apresenta uma vasta quantidade de trabalhos utilizando métodos bayesianos, a exemplo o *survey* de Fragoso, Bertoli e Louzada (2018), que combina o método de MCMC.

A utilização do filtro de Kalman, a obtenção de parâmetros iniciais a partir uma rápida convergência originada pela filtragem do sinal de entrada, contemplando uma classe de modelos lineares e não-lineares. Adicionados a isso, a alteração dos parâmetros de ruído das medidas e também de processo garante o controle do período transitório dos sinais de saída.

Algumas vantagens do método proposto aqui envolvem o baixo esforço computacional exigido pelo filtro de Kalman, além da possibilidade de sincronismo em sinais que contenham variação de frequência. Também podem ser citadas outras características interessantes, que também estão presentes em outros métodos, como a filtragem de harmônicos e a facilidade de implementação dos algoritmos.

7.2 Trabalhos Futuros

Mesmo com os consideráveis esforços empreendidos, seria inviável proceder aqui uma análise sobre todas as técnicas relevantes dentro do universo da estatística bayesiana. Ainda que delimitando o escopo apenas no que toca ao MCMC, pode-se realizar mais revisões da bibliografia disponível, no quesito das recentes inovações sobre a temática. Como exemplo,

pode-se citar os métodos de reamostragem utilizando técnicas Hamiltonianas (Monte Carlo Hamiltoniano).

Além disso, uma área de pesquisas que também vem ganhando notoriedade, é a das Redes Neurais Artificiais (RNA). RNAs são sistemas conexionistas que realizam uma determinada tarefa, aprendendo sobre exemplos sem que haja conhecimento prévio sobre a tarefa. Geralmente, as RNAs usam estimativas pontuais, como pesos, e apresentam desempenho satisfatório quando há grandes conjuntos de dados. No entanto, essas estimativas pontuais não expressam as incertezas em regiões com poucos dados, podendo levar a decisões excessivamente confiantes. Por esse motivo, *Bayesian Deep Neuro Networks* pode ser um conceito alternativo que também incorpora uma medida para incertezas e regularização, trazendo interpretabilidade ao modelo (SHRIDHAR; LAUMANN; LIWICKI, 2019).

Referências Bibliográficas

BAYES, T. An essay towards solving a problem in the doctrine of chances. 1763. *MD computing: computers in medical practice*, v. 8, n. 3, p. 157, 1991.

CAMPAGNOLI, P.; PETRIS, G.; PETRONE, S. *Dynamic Linear Models with R*. [S.l.]: Springer, 2009.

COBB, G. W. The problem of the nile: Conditional solution to a changepoint problem. *Biometrika*, Oxford University Press, v. 65, n. 2, p. 243–251, 1978.

CORCHADO, E.; WOŹNIAK, M.; ABRAHAM, A.; CARVALHO, A. C. D.; SNÁŠEL, V. Recent trends in intelligent data analysis. *Neurocomputing*, n. 126, p. 1–2, 2014.

DEGROOT, M. H. *Probability and Statistics Addison*. [S.l.]: Wesley Publishing Company, Reading, Massachusetts, 1986.

EHLERS, R. S. *Introdução a Inferência Bayesiana*. São Carlos, SP - Brasil, 2017.

ELLIOTT, H.; DERIN, H.; CRISTI, R.; GEMAN, D. Application of the gibbs distribution to image segmentation. In: IEEE. *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.], 1984. v. 9, p. 678–681.

ESTEVES, L. G.; STERN, R. B. *Inferência Bayesiana*. São Carlos, SP - Brasil, 2017.

FINETTI, B. D. La prévision: ses lois logiques, ses sources subjectives. In: *Annales de l'institut Henri Poincaré*. [S.l.: s.n.], 1937. v. 7, n. 1, p. 1–68.

FRAGOSO, T. M.; BERTOLI, W.; LOUZADA, F. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, Wiley Online Library, v. 86, n. 1, p. 1–28, 2018.

GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. [S.l.]: Chapman and Hall/CRC, 2006.

GOOD, I. J. Probability and the weighing of evidence. 1950.

GREWAL, M.; ANDREWS, A. Kalman filtering: Theory and practice with matlab 4 th edition. John Wiley & Sons, 2015.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. Oxford University Press, 1970.

JEFFREYS, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, The Royal Society London, v. 186, n. 1007, p. 453–461, 1946.

KALMAN, R. E. A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, 1960.

KIM, J.-Y.; KIM, T.-Y. Soccer ball tracking using dynamic kalman filter with velocity control. In: IEEE. *2009 Sixth International Conference on Computer Graphics, Imaging and Visualization*. [S.l.], 2009. p. 367–374.

LINDLEY, D. V. *et al.* On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 27, n. 4, p. 986–1005, 1956.

MCGRAYNE, S. B. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. [S.l.]: Yale University Press, 2011.

NAYLOR, T. H.; SEAKS, T. G.; WICHERN, D. W. Box-jenkins methods: An alternative to econometric models. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, p. 123–137, 1972.

O'HAGAN, A.; FORSTER, J. Bayesian inference, volume 2b of kendall's advanced theory of statistics. 1994.

PAULINO, C. D. M.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística bayesiana*. [S.l.]: Fundação Calouste Gulbenkian, 2003.

PETRIS, G.; PETRONE, S. State space models in r. *Journal of Statistical Software*, jstatsoft, v. 41, n. 4, 2011.

RAMSEY, F. P. Truth and probability. *Studies in subjective probability*, 01 1926.

SAVAGE, L. J.; BARNARD, G.; CORNFIELD, J.; BROSS, I.; GOOD, I.; LINDLEY, D.; CLUNIES-ROSS, C.; PRATT, J. W.; LEVENE, H.; GOLDMAN, T. *et al.* On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, JSTOR, v. 57, n. 298, p. 307–326, 1962.

SHRIDHAR, K.; LAUMANN, F.; LIWICKI, M. A comprehensive guide to bayesian convolutional neural network with variational inference. *arXiv preprint arXiv:1901.02731*, 2019.

WEST, M.; HARRISON, J. *Bayesian forecasting and dynamic models*. [S.l.]: Springer Science & Business Media, 2006.

XIE, X.; SUDHAKAR, R.; ZHUANG, H. Real-time eye feature tracking from a video image sequence using kalman filter. *IEEE Transactions on systems, man, and cybernetics*, IEEE, v. 25, n. 12, p. 1568–1577, 1995.

Anexo

Inspirado no texto *Bayesian state space estimation in Python via Metropolis-Hastings* disponível em <https://github.com/ChadFulton/tsa-notebooks/blob/master/state_space_mh.ipynb>, acessado em 20/05/2019.

```
###################################
## IMA(1,1) via Espaço de Estado ##
###################################
\%matplotlib inline
import numpy as np
import pandas as pd
import pymc as mc
from scipy import signal
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sn
np.set_printoptions(precision=4, suppress=True, linewidth=120)

# True values
T = 1000
sigma2_eps0 = 3
sigma2_eta0 = 10

# Simulate data
np.random.seed(1234)
eps = np.random.normal(scale=sigma2_eps0**0.5, size=T)
eta = np.random.normal(scale=sigma2_eta0**0.5, size=T)
mu = np.cumsum(eta)
y = mu + eps

# Plot the time series
fig, ax = plt.subplots(figsize=(13,2))
ax.plot(y);
ax.set(xlabel='$T$', title='Simulated series');
```

```
##Simulação Posteriori
# Priors
precision = mc.Gamma('precision', 2, 4)
ratio = mc.Gamma('ratio', 2, 1)

# Likelihood calculated using the state-space model
class LocalLevel(sm.tsa.statespace.MLEModel):
    def __init__(self, endog):
        # Initialize the state space model
        super(LocalLevel, self).__init__(endog, k_states=1,
                                         initialization='approximate_diffuse',
                                         loglikelihood_burn=1)

        # Initialize known components of the state space matrices
        self.ssm['design', :] = 1
        self.ssm['transition', :] = 1
        self.ssm['selection', :] = 1

    @property
    def start_params(self):
        return [1. / np.var(self.endog), 1.]

    @property
    def param_names(self):
        return ['h_inv', 'q']

    def update(self, params, transformed=True, **kwargs):
        params = super(LocalLevel, self).update(params, transformed, **kwargs)

        h, q = params
        sigma2_eps = 1. / h
        sigma2_eta = q * sigma2_eps

        self.ssm['obs_cov', 0, 0] = sigma2_eps
        self.ssm['state_cov', 0, 0] = sigma2_eta

# Instantiate the local level model with our simulated data
```

```
ll_mod = LocalLevel(y)
ll_mod.filter(ll_mod.start_params)

# Create the stochastic (observed) component
@mc.stochastic(dtype=LocalLevel, observed=True)
def local_level(value=ll_mod, h=precision, q=ratio):
    return value.loglike([h, q], transformed=True)

# Create the PyMC model
ll_mc = mc.Model((precision, ratio, local_level))

# Create a PyMC sample
ll_sampler = mc.MCMC(ll_mc)
```