FUNDAÇÃO UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
FACULDADE DE ENGENHARIAS, ARQUITETURA E URBANISMO E GEOGRAFIA
PROGRAMA DE PÓS-GRADUAÇÃO EM EFICIÊNCIA ENERGÉTICA E SUSTENTABILIDADE
CURSO DE MESTRADO PROFISSIONAL EM EFICIÊNCIA ENERGÉTICA E SUSTENTABILIDADE

MODELO ESTATÍSTICO PARA PREDIÇÃO DE GERAÇÃO DE ENERGIA SOLAR FOTOVOLTAICA

Mariana Villela Flesch

2024 **CAMPO GRANDE|MS**

FUNDAÇÃO UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

FACULDADE DE ENGENHARIAS, ARQUITETURA E URBANISMO E GEOGRAFIA

MESTRADO PROFISSIONAL EM EFICIÊNCIA ENERGÉTICA E SUSTENTABILIDADE

MODELO ESTATÍSTICO PARA PREDIÇÃO DE GERAÇÃO DE ENERGIA SOLAR FOTOVOLTAICA

MARIANA VILLELA FLESCH

Trabalho de Conclusão de Curso do Mestrado Profissional apresentada na Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia da Universidade Federal de Mato Grosso do Sul, para obtenção do título de Mestre em Eficiência Energética e Sustentabilidade, na área de concentração Eficiência Energética.

Orientador: Prof. Dr. Erlandson Ferreira Saraiva

CAMPO GRANDE

Agosto/2024

FOLHA DE APROVAÇÃO

Mariana Villela Flesch

MODELO ESTATÍSTICO PARA PREDIÇÃO DE GERAÇÃO DE ENERGIA SOLAR FOTOVOLTAICA

Redação final do Trabalho de Conclusão de Curso, aprovada pela Banca Examinadora em 29 de Agosto de 2024, na Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia da Universidade Federal de Mato Grosso do Sul para obtenção do título de Mestre em Eficiência Energética e Sustentabilidade.

Banca examinadora:

Prof. Dr. Erlandson Ferreira Saraiva

Documento assinado digitalmente
gov.br
ERLANDSON FERREIRA SARAIVA
Data: 26/09/2024 18:06:39-0300
Verifique em https://validar.iti.gov.br

Bacharel em Matemática Aplicada e Computacional (UCDB)

Mestre e Doutor em Estatística (DEs/UFSCar)

Pós-doutorado em Estatística (ICMC/USP)

Prof. Dr. Willy Alves de Oliveira Soler

Documento assinado digitalmente
gov.br
WILLY ALVES DE OLIVEIRA SOLER
Data: 26/09/2024 18:35:07-0300
Verifique em https://validar.iti.gov.br

Licenciado em Matemática (UFMS)

Mestre em Matemática (PROFMAT/UFMS)

Doutor em Ciência da Computação e Matemática Computacional (ICMC/USP)

Prof. Dr. Valdemiro Piedade Vigas

Documento assinado digitalmente
gov.br
VALDEMIRO PIEDADE VIGAS
Data: 26/09/2024 21:03:52-0300
Verifique em https://validar.iti.gov.br

Bacharel em Estatística (UFBA)

Mestre em Estatística (UFScar)

Doutor em Ciências - Área de concentração: Estatística e Experimentação Agronômica (ESALQ/USP)

AGRADECIMENTOS

Primeiramente agradeço a Deus, por ter me dado serenidade e sabedoria para chegar até aqui.

A minha família, em especial meu filho, pela paciência nas minhas ausências.

Agradeço ao meu orientador Prof. Dr Erlandson Ferreira Saraiva, por toda a sua dedicação e orientação, foram fundamentais para o desenvolvimento deste trabalho.

Ao Programa de Pós Graduação em Eficiência Energética e Sustentabilidade, onde pude compartilhar os conhecimentos durante o curso.

ABREVIAÇÕES

MEPA – Média do erro percentual absoluto

RQEM – Raiz quadrada do erro quadrático médio

AIC – Critério de informação de Akaike

BIC – Critério de informação bayesiano

EPA – Erro percentual absoluto

f.d.p. – Função densidade de probabilidade

DM – Distância de Mahalanobis

PG – Processo Gaussiano

D – Conjunto de dados reais

MCMC – Markov Chain Monte Carlo

R – Software estatístico

Resumo

Nos últimos 10 anos, houve um aumento significativo da geração de energia através de usinas fotovoltaicas, tanto residenciais quanto de grande porte. Em regiões com incidência solar abundante, como é o caso do Brasil, usinas fotovoltaicas são frequentemente instaladas devido a viabilidade financeira. Com isso, a cada dia mais usinas fotovoltaicas estão sendo ligadas ao sistema de rede elétrica das cidades. No entanto, isto pode causar instabilidade na rede, gerando desafios para as concessionárias de energia, pois estas administram a estrutura de transmissão e distribuição de energia elétrica. Uma ferramenta de auxílio na resolução deste problema é o desenvolvimento de modelos preditivos capazes de informar com alta confiabilidade a quantidade de energia que será gerada e inserida no sistema elétrico por uma usina. Nesta dissertação, propomos uma abordagem Bayesiana para: (i) estimar a curva de crescimento de uma função $f(\cdot)$ que modela a geração de energia solar em n dias; e (ii) predizer a curva de crescimento para o $(n+1)$-ésimo dia usando o histórico dos valores registrados. Para isso, assumimos que $f(\cdot)$ é uma função desconhecida, mas com vetor de valores $\mathbf{f(t)} = (f(1), \ldots, f(k))$ sendo gerado *a priori* de um processo Gaussiano. Uma vantagem dessa abordagem é que podemos estimar a curva das funções $f(\cdot)$ e $f_{n+1}(\cdot)$ por "funções suaves" que são obtidas ligando por linhas os pontos gerados a partir de uma distribuição normal k-variada com vetor de médias e matriz de covariâncias apropriados. Contudo, como a distribuição *a posteriori* conjunta para os parâmetros de interesse não possui uma forma matemática conhecida, descrevemos como implementar um algoritmo *Gibbs sampling* para obter as estimativas para os parâmetros. Ilustramos a performance do modelo proposto utilizando dois estudos de simulação e uma aplicação a um conjunto de dados reais. Como medidas de desempenho, calculamos o erro percentual absoluto, a média do erro percentual absoluto (MEPA) e a raiz quadrada do erro quadrático médio (RQEM). Em todos os casos simulados e na aplicação, os valores de MEPA e REQM são próximos a 0, indicando um desempenho muito satisfatório da

abordagem proposta.

Palavras chave: Energia Solar Fotovoltaica; Distribuição Normal Mutivariada; Processo Gaussiano; Inferência Bayesiana; Algoritmo *Gibbs Sampling*.

Abstract

Over the past 10 years, there has been a significant increase in energy generation through photovoltaic plants, both residential and large-scale. In regions with abundant solar incidence, such as Brazil, photovoltaic plants are often installed due to financial viability. As a result, more and more photovoltaic plants are being connected to the electrical grid system of cities every day. However, this can cause instability in the grid, creating challenges for energy concessionaires, as they manage the structure of transmission and distribution of electrical energy. A tool to help solve this problem is the development of predictive models capable of informing with high reliability the amount of energy that will be generated and inserted into the electrical system by a plant. In this dissertation, we propose a Bayesian approach to estimate the curve of a function $f(\cdot)$ that models the solar power generated at k moments per day for n days and to forecast the curve for the $(n+1)$th day by using the history of recorded values. We assume that $f(\cdot)$ is an unknown function and adopt a Bayesian model with a Gaussian-process prior on the vector of values $\mathbf{f(t)} = (\mathbf{f(1)}, \ldots, \mathbf{f(k)})$. An advantage of this approach is that we may estimate the curves of $f(\cdot)$ and $f_{n+1}(\cdot)$ as "smooth functions" obtained by interpolating between the points generated from a k-variate normal distribution with appropriate mean vector and covariance matrix. Since the joint posterior distribution for the parameters of interest does not have a known mathematical form, we describe how to implement a Gibbs sampling algorithm to obtain estimates for the parameters. The good performance of the proposed approach is illustrated using two simulation studies and an application to a real dataset. As performance measures, we calculate the absolute percentage error, the mean absolute percentage error (MAPE), and the root-mean-square error (RMSE). In all simulated cases and in the application to real-world data, the MAPE and RMSE values were all near 0, indicating the very good performance of the proposed approach.

Palavras chave: Photovoltaic solar power forecasting; statistical modeling; Bayesian inference; Gaussian process; MCMC; Gibbs sampling algorithm.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Energia solar fotovoltaica é a energia gerada através da conversão direta da luz solar em eletricidade (Sampaio e González, 2017). Esta conversão ocorre por meio do efeito fotovoltaico, sendo este um processo físico pelo qual uma célula fotovoltaica converte luz solar em eletricidade. Este efeito foi primeiramente observado pelo físico francês Alexandre-Edmond Becquerel em 1839 (Parida *et al.*, 2011; Sampaio e González, 2017).

O efeito fotovoltaico ocorre em materiais denominados de semicondutores, *i.e.*, materiais com propriedades de condução elétrica intermediarias entre isolantes e condutores. O material semi-condutor mais utilizado na construção de painéis solares é o silício, que de acordo com Sampaio e González (2017) é o segundo elemento mais abundante na terra.

Além de ser um tipo de energia limpa e renovável, Sampaio e González (2017) e Silveira *et al.* (2013) citam as seguinte vantagens do sistema de geração de energia solar fotovoltaica: sistema de geração de energia confiável, baixo custo de operação e manutenção, a geração pode ficar mais próxima do consumidor, baixo impacto ambiental, potencial para mitigar emissões de gases de efeito estufa e é um sistema silencioso.

Como desvantagens, podemos citar: alto custo inicial, necessita de uma área relativamente grande para a instalação das placas solares, alta dependência de desenvolvimento tecnológico e condições geográficas (Sampaio e González, 2017; Silveira *et al.*, 2013).

Nos últimos anos, com o aumento do interesse da sociedade e dos governos por geração de energias limpas, renováveis e que reduzam a emissão de CO_2 na atmosfera, juntamente com desenvolvimento tecnológico e redução dos valores dos painéis solares, houve um aumento significativo da geração de energia solar fotovoltaica, tanto por centrais e/ou usinas fotovoltaicas como por residências com painéis fotovoltaicos instalados nos telhados.

Em regiões com incidência de sol abundante, como é o caso do Brasil, a cada dia mais centrais e/ou usinas fotovoltaicas estão sendo ligadas ao sistema de rede elétrica das cidades. No entanto, de acordo com Alkandri e Ahmad (2020) e Sharadga *et al.* (2020), isso pode causar instabilidade na rede, constituindo-se num desafio para as concessionárias de energia. Isso acontece porque os operadores elétricos precisam saber o quanto de energia será introduzida no sistema para equilibrar com o consumo e garantir que o sistema seja capaz de cobrir a demanda dos consumidores. Para Sharadga *et al.* (2020), ser capaz de predizer a quantidade de energia que será inserida na rede é muito importante para garantir a eficiência e o gerenciamento da rede elétrica.

Uma solução para esse problema é modelar a geração de energia solar fotovoltaica como função do tempo e desenvolver modelos preditivos capazes de informar com alta confiabilidade a quantidade de energia que será gerada por uma usina e a quantidade que será inserida no sistema elétrico. Esta capacidade de predição também possibilita ao administrador da usina um melhor gerenciamento, desenvolvimento e/ou adaptação da rede que levará a energia gerada até o sistema elétrico de forma eficiente e consequentemente com redução de custos. Além disso, um modelo de predição eficiente permite a identificação de possíveis avarias no sistema de geração e/ou perdas devido a sujidade presente na superfície dos painéis solares.

Sob este cenário, os modelos estatísticos surgem como uma ferramenta importante para modelagem e predição da geração de energia solar fotovoltaica. Algumas abordagens estatísticas utilizadas para este propósito incluem: modelos de regressão linear (Ibrahim *et al.*, 2012; Asri *et al.*, 2021; Erten e Aydilek, 2022; El-Aal *et al.*, 2023), modelos autoregressivos (Bimenyimana *et al.*, 2017; Rogier e Mohamudally, 2019; Pamanin *et al.*, 2022), modelos autoregressivos de médias móveis (Singh e Pozo, 2019), modelos de redes neurais artificiais (Mellit *et al.*, 2013; Abdelhak *et al.*, 2023; Amer *et al.*, 2023), modelos de crescimento com efeitos mistos (Souza *et al.*, 2022), entre outros. Isto é, os modelos paramétricos ainda são comumente empregados devido à sua facilidade de uso.

No entanto, de acordo com Saraiva *et al.* (2023), os modelos paramétricos podem apresentar pelo menos três limitações. São elas:

(i) a análise fica limitada à função (ou funções) previamente escolhida pelo analista;

(ii) a complexidade e/ou flexibilidade das funções consideradas é limitada pelo número de parâmetros presentes nas funções; e

(iii) pode haver diversas funções paramétricas que modelam igualmente bem os valores registrados.

Uma solução usual adotada em muitos artigos de modelagem e/ou análise de dados é ajustar um conjunto de modelos candidatos, previamente escolhidos pelo analista, e então selecionar o melhor modelo usando algum critério de seleção de modelos, como o AIC (Akaike, 1974) ou BIC (Schwarz, 1978). No entanto, as questões (ii) e (iii) descritas acima ainda permanecem.

Nesta dissertação, propomos uma abordagem Bayesiana para modelar a geração de energia solar fotovoltaica por usinas e também fazer predições. Como a quantidade de energia solar gerada em um dia apresenta um comportamento não-linear instável, optamos por modelar o logaritmo das quantidades acumulada, pois estas medidas apresentam um comportamento estável e padrão. Então, assumimos que a energia solar fotovoltaica gerada e registrada em k instantes de tempo de um dia, de forma acumulada, é modelada por um modelo aditivo composto pelos valores de uma função de crescimento não-linear $f(\cdot)$, avaliada em k instantes de tempo, acrescida de um erro aleatório $\boldsymbol{\varepsilon}$ gerado de acordo com uma distribuição Gaussiana.

No entanto, com o objetivo de flexibilizar a modelagem e não ficar restrito a uma função $f(\cdot)$ escolhida previamente pelo analista, assumimos que $f(\cdot)$ é uma função desconhecida, porém, com vetor de valores $\mathbf{f}(\mathbf{t}) = (f(1), \ldots, f(k))$ sendo parâmetros que precisam ser estimados a partir do conjunto de dados registrados, para $\mathbf{t} = (1, \ldots, k)$. Para isso, desenvolvemos uma modelagem Bayesiana considerando que os valores $\mathbf{f}(\mathbf{t})$ são gerados *a priori* de um processo Gaussiano. Uma vantagem dessa abordagem é que podemos estimar a curva da função $f(\cdot)$ por "funções suaves" que são obtidas ligando por linhas, pontos gerados de uma distribuição normal de k-variada com apropriado vetor médias e matriz de covariâncias.

Além disso, apresentamos um procedimento de predição para a curva de crescimento do $(n+1)$-ésimo dia, dado os valores registrados nos primeiros n dias. Porém, como a distribuição *a posteriori* conjunta para os parâmetros de interesse e a distribuição preditiva do modelo proposto não possuem formas matemáticas conhecidas, que permitam obter os estimadores para os parâmetros de maneira analítica, descrevemos como implementar um algoritmo *Gibbs sampling* (Geman e Geman, 1984; Gelfand e Smith, 1990; Gelman e Rubin, 1992) para gerar valores aleatórios dessas distribuições de probabilidades e obter

as estimativas para os parâmetros. Este algoritmo gera valores das distribuições de probabilidades de interesse, de forma indireta, usando as distribuições *a posteriori* condicionais, desde que sejam conhecidas, como é o caso do modelo proposto.

Para ilustrar a performance do modelo, desenvolvemos dois estudos de simulação. No primeiro, apresentamos o desempenho do modelo na estimação da curva da função $f(\cdot)$ utilizando dados simulados. Como medidas de desempenho, calculamos o erro percentual absoluto (EPA) e a média do erro percentual absoluto (MEPA). Em todos os casos simulados, a abordagem proposta apresentou valores de MEPA próximos de 0, indicando que os valores estimados estão próximos dos valores reais. No segundo estudo de simulação, avaliamos a performance da abordagem proposta em relação à predição da curva para o $(n+1)$-ésimo dia. Analogamente ao estudo de simulação 1, o modelo apresentou um desempenho satisfatório, uma vez que os valores de MEPA estão próximos de zero. Além disso, também avaliamos a performance das predições utilizando a raiz quadrada do erro quadrático médio (RQEM). Assim como, os valores MEPA ou valores RQEM são próximos de zero.

Aplicamos a abordagem proposta a um conjunto de dados reais, obtido de um experimento realizado em uma mini-usina fotovoltaica instalada no campus de Campo Grande da Universidade Federal de Mato Grosso do Sul. Este conjunto de dados está disponível no *website* `https://github.com/lscad-facom-ufms/Solar2` e maiores detalhes sobre o experimento estão descritos no artigo de Souza *et al.* (2022). Semelhante aos estudos de simulação 1 e 2, os resultados obtidos foram satisfatoriamente precisos com valores de MEPA próximos de zero.

O restante desta dissertação está organizada da seguinte maneira. Como o modelo é baseado no uso do processo Gaussiano, no Capítulo 2, apresentamos a definição de processo Gaussiano e como este processo pode ser utilizado para estimar funções "suaves" apenas gerando valores de uma distribuição normal multivariada. No Capítulo 3, apresentamos os dados que motivaram o desenvolvimento desta dissertação e o modelo proposto. Além disso, apresentamos o procedimento de estimação dos parâmetros de interesse e de predição. No Capítulo 4, apresentamos os resultados obtidos com a aplicação da modelagem proposta aos dados reais. No Capítulo 5, apresentamos as considerações finais e descrevemos algumas propostas a serem desenvolvidas em trabalhos futuros.

Capítulo 2

Processo Gaussiano

Neste Capítulo, apresentamos a definição do processo Gaussiano e como podemos utilizá-lo em um modelo Bayesiano como distribuição *a priori* sobre um vetor $\mathbf{f}(\mathbf{t}) = (f(1), \ldots, f(k))$ composto por k valores de uma função $f(\cdot)$ considerada desconhecida. Como o processo Gaussiano é uma extensão da distribuição normal multivariada para espaços de dimensão infinita, iniciamos relembrando a definição da distribuição normal univariada, bivariada e multivariada.

2.1 Distribuição normal

No século XIX, os assistentes de Carl Friedrich Gauss, grande matemático e astrônomo da época, realizavam experimentos e medições astronômicas continuamente, porém todas as vezes que se repetiam as medições, nunca obtinham os mesmos resultados. Gauss, com seu vasto conhecimento e observando o problema, se propôs a resolvê-lo. Gauss, então, construiu um histograma dos resultados e notou que os valores observados formavam uma curva, que é denominada de curva gaussiana ou curva em forma de sino (Patel e Campbell, 1996).

De acordo com Wlodzimierz (1995) e Patel e Campbell (1996), o que Gauss criou foi um método para minimizar os erros de medições repetidas, que é obtido através de um valor estimado somado a uma incerteza associada ao resultados, que chamamos de erro aleatório. Esse método também é conhecido como a lei gaussiana dos erros e a curva em forma de sino é denominada de distribuição Gaussiana (ou normal).

De acorco com Casella e Berger (2002), temos a seguinte definição para a distribuição normal univariada.

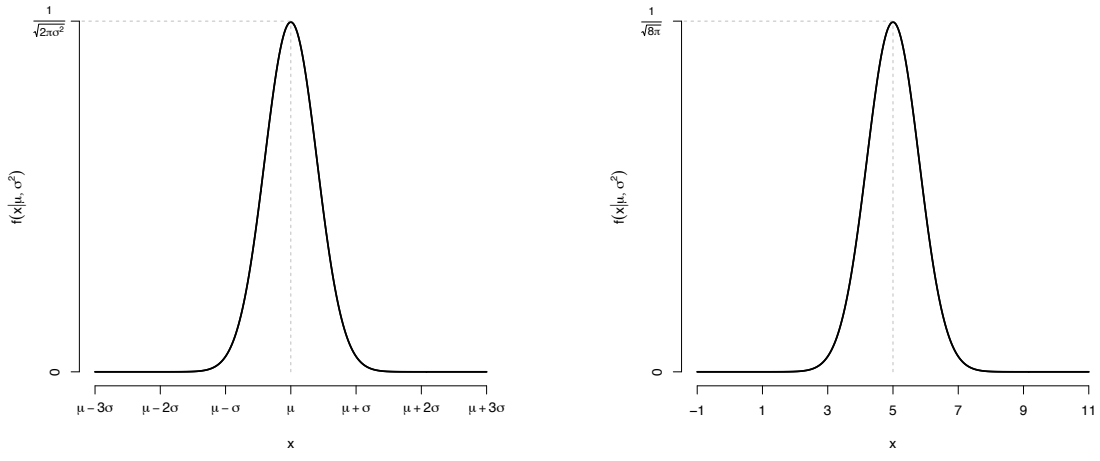
Definição 1. *Seja X uma variável aleatória contínua definida em $\mathbb{R}$. Dizemos que X segue o modelo Normal de parâmetros μ e σ^2, se sua função densidade de probabilidade é dada por*

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

para $\mu \in \mathbb{R}$ e $\sigma^2 \in \mathbb{R}^+$.

Utilizamos a notação $X \sim \mathcal{N}(\mu, \sigma^2)$ para indicar que a variável aleatória X segue o modelo Normal de parâmetros μ e σ^2, sendo $\mu = \mathbb{E}(X)$ e $\sigma^2 = \mathbb{V}ar(X)$, *i.e.*, o valor esperado e a variância de X, respectivamente.

Fixado os valores de μ e σ^2, a distribuição normal fica completamente definida e seu gráfico é uma curva em forma de sino. A Figura 2.1(a), apresenta uma ilustração do gráfico de uma distribuição normal de parâmetros μ e σ^2 (caso geral); e a Figura 2.1(b) ilustra o gráfico da distribuição normal para um caso específico com $\mu = 5$ e $\sigma^2 = 4$ (*i.e.*, $\sigma = 2$). Note que, o pico da distribuição (valor máximo) se dá para o valor $x = \mu$ e a escala é dada pelo valor do desvio-padrão σ. Além disso, note que, a curva em forma de sino é assintota ao eixo horizontal em ambas as direções; e como $f(x|\mu,\sigma^2)$ é uma função densidade de probabilidade, então a área total abaixo da curva e acima do eixo das abcissas é igual a 1, sendo a distribuição, simétrica em relação a $x = \mu$, *i.e.*, a mediana é o valor $x = \mu$.



(a) Parâmetros μ e σ^2

(b) Parâmetros $\mu = 5$ e $\sigma^2 = 4$

Figura 2.1: Gráfico de $f(x|\mu,\sigma^2)$.

2.2 Distribuição normal bivariada

Para definirmos a distribuição normal bivarida, considere Z_1 e Z_2 duas variáveis aleatórias independentes com distribuição normal padrão, *i.e.*,

$$Z_1 \sim \mathcal{N}(0,1) \qquad \text{e} \qquad Z_2 \sim \mathcal{N}(0,1).$$

As funções densidades de probabilidades de Z_1 e Z_2 são dadas por

$$f(z_1) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{z_1^2}{2}\right\} \qquad \text{e} \qquad f(z_2) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{z_2^2}{2}\right\},$$

para $z_1, z_2 \in \mathbb{R}$.

Como Z_1 e Z_2 são independentes, então a função densidade de probabilidade conjunta do vetor aleatório $\mathbf{Z} = (Z_1, Z_2)$ é dada por

$$f_{\mathbf{Z}}(z_1, z_2) = f(z_1) f(z_2) = \frac{1}{2\pi} exp\left\{-\frac{1}{2}\left(z_1^2 + z_2^2\right)\right\}.$$

Considere agora, a seguinte transformação

$$X_1 = \sigma_1 Z_1 + \mu_1 \qquad \text{e} \qquad X_2 = \sigma_2\left[\rho Z_1 + \left(1-\rho^2\right)^{1/2} Z_2\right] + \mu_2$$

onde μ_1, μ_2, σ_1, σ_2 são constantes, tais que, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in \mathbb{R}^+$ e $-1 < \rho < 1$. Assuma que temos interesse em determinar a função densidade de probabilidade conjunta do vetor aleatório $\mathbf{X} = (X_1, X_2)$.

Como a transformação (Z_1, Z_2) em (X_1, X_2) é uma transformação linear, então o Jacobiano da transformação é dado por

$$J = \begin{vmatrix} \frac{dX_1}{dZ_1} & \frac{dX_1}{dZ_2} \\ \frac{dX_2}{dZ_1} & \frac{dX_2}{dZ_2} \end{vmatrix} = \begin{vmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\left(1-\rho^2\right)^{1/2} \end{vmatrix} = \left(1-\rho^2\right)^{1/2} \sigma_1\sigma_2.$$

Portanto, a função densidade de probabilidade de $\mathbf{X} = (X_1, X_2)$ é dada por

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2|\boldsymbol{\theta}) &= f_{\mathbf{Z}}(x_1, x_2)\left|J^{-1}\right| \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right.\right. \\ &\quad \left.\left. -2\rho\frac{(X_1-\mu_1)(X_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}, \end{aligned} \tag{2.1}$$

para $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Detalhes sobre a obtenção desta função densidade de probabilidade é dada no Apêndice 1.

O termo ρ presente na Equação em (2.1) é denominado de coeficiente de correlação linear entre as variáveis X_1 e X_2. O valor de ρ é definido no intervalo $[-1, 1]$. Quanto mais próximo de 1 é o valor de ρ, mais forte é a relação linear (positiva) entre as variáveis. Quanto mais próximo de -1 é o valor de ρ, mais forte é a relação linear (negativa) entre as variáveis. Se $\rho = 1$ ou $\rho = -1$, a relação entre as variáveis é linearmente perfeita, positiva ou negativa, respectivamente. Se $\rho = 0$, não há relação linear entre as variáveis, *i.e.*, as variáveis aleatórias X_1 e X_2 são ditas não correlacionadas. Neste caso, sua função densidade de probabilidade (f.d.p.) é obtida substituindo ρ por 0 na expressão em (2.1). Ou seja,

$$f(x_1, x_2|\boldsymbol{\theta}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} - \frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right\},$$

para $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

2.2.1 Notação matricial

Uma alternativa de apresentar a distribuição normal bivariada é utilizando a notação matricial. Para isto, considere Σ uma matriz de dimensão 2×2 dada por

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

onde σ_1^2 é a variância da variavel aleatória X_1, σ_2^2 é a variância da variável aleatória X_2 e $\sigma_{12} = Cov(X_1, X_2) = \rho\sigma_1\sigma_2$ é a covariância de (X_1, X_2). A matriz Σ é denominada de matriz de covariâncias.

O determinante de Σ é dado por

$$|\Sigma| = \sigma_1^2\sigma_2^2 \left(1 - \rho^2\right).$$

Portanto, o termo $\sqrt{\sigma_1^2\sigma_2^2 \left(1 - \rho^2\right)}$ presente na Equação em (2.1) pode ser escrito como $|\Sigma|^{1/2}$. Além disso, temos que a inversa de Σ, denotada por Σ^{-1} é dada por

$$\Sigma^{-1} = \frac{1}{(1 - \rho^2)} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}.$$

Detalhes sobre a obtenção de Σ^{-1} é dada no Apêndice 2.

Definindo o vetor coluna $(\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$, temos que, a distância de Mahalanobis $\mathbb{DM} = (\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$ é dada por:

$$
\begin{aligned}
\mathbb{DM} &= \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1^2(1-\rho^2)} & -\frac{\rho}{\sigma_1\sigma_2(1-\rho^2)} \\ -\frac{\rho}{\sigma_1\sigma_2(1-\rho^2)} & \frac{1}{\sigma_2^2(1-\rho^2)} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{x_1-\mu_1}{\sigma_1^2(1-\rho^2)} - \frac{\rho(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} & -\frac{\rho(x_1-\mu_1)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(x_2-\mu_2)}{\sigma_2^2(1-\rho^2)} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \frac{(x_1-\mu_1)^2}{\sigma_1^2(1-\rho^2)} - \frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} - \frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(x_2-\mu_2)^2}{\sigma_2^2(1-\rho^2)} \\
&= \frac{(x_1-\mu_1)^2}{\sigma_1^2(1-\rho^2)} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(x_2-\mu_2)^2}{\sigma_2^2(1-\rho^2)} \\
&= \frac{1}{(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]. \qquad (2.2)
\end{aligned}
$$

Ou seja, a parte interna de $exp(\cdot)$ da expressão em (2.1) pode ser escrita como

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

Portanto, a função densidade de probabilidade da distribuição normal bivariada na forma matricial é dada por

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Como notação matemática, escrevemos $\mathbf{X} = (X_1, X_2) \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$, onde $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ é o vetor de média e $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ é a matriz de covarâncias.

2.3 Distribuição normal multivariada

Considere $\mathbf{X} = (X_1, \ldots, X_k)$ um vetor de dimensão k, para $k > 1$ e $\mathbf{X} \in \mathbb{R}^k$. Seja $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\top$ um vetor de médias de dimensão $k \times 1$ e Σ uma matriz de dimensão $k \times k$, dada por

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2k} \\ \vdots & \ddots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{kk} \end{pmatrix}$$

onde $\sigma_{ij} = \mathbb{C}ov(X_i, X_j)$ e Σ é uma matriz simétrica e positiva definida, para $i, j = 1, \ldots, k$.

Análogo ao caso bivariado, distância de Mahalanobis é dada por $(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$. A função densidade de probabilidade de uma distribuição normal multivariada é dada por

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = (2\pi)^{k/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \qquad (2.3)$$

Utilizamos a notação $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$, para "dizer" que o vetor $\mathbf{X}$ têm distribuição Normal multivariada de dimensão k, para $k > 1$.

De acordo com Patel e Campbell (1996) e Wlodzimierz (1995), se $\mathbf{X}$ é um vetor aleatório com distribuição normal multivariada, temos as seguintes propriedades:

(1) Se $\mathbf{a} = (a_1, \ldots, a_k)$ é um vetor real de dimensão $k \times 1$, então a combinação linear $\mathbf{a}^\top \mathbf{X} = a_1 X_1 + a_2 X_2 + ... + a_k X_k$ tem distribuição $\mathcal{N}(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^T \Sigma \mathbf{a})$. Note que:

$$\mathbb{E}\left(\mathbf{a}^\top \mathbf{X}\right) = \mathbf{a}^\top \mathbb{E}\left(\mathbf{X}\right) = \mathbf{a}^\top \boldsymbol{\mu};$$

e

$$\begin{aligned} Var(\mathbf{a}^\top \mathbf{X}) &= \mathbb{E}\left[(\mathbf{a}^T \mathbf{X} - \mathbf{a}^T \boldsymbol{\mu})(\mathbf{a}^T \mathbf{X} - \mathbf{a}^T \boldsymbol{\mu})^T\right] \\ &= \mathbb{E}\left[\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{a}\right] \\ &= \mathbf{a}^T \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] \mathbf{a} \\ &= \mathbf{a}^T \Sigma \mathbf{a}. \end{aligned}$$

(2) Subconjuntos de $\mathbf{X}$ têm distribuição normal multivariada. Por exemplo, fazendo $a_i = 0$ para $i = 3, \ldots, k$, então pela propriedade 1, temos que

$$\mathbf{a}^T \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} = (X_1, X_2)$$

com $\mathbb{E}(X_1, X_2) = \boldsymbol{\mu}_{12} = (\mu_1, \mu_2)^\top$ e $\Sigma_{12} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$. Logo, $(X_1, X_2) \sim \mathcal{N}_2(\boldsymbol{\mu}_{12}, \Sigma_{12})$.

(3) Se os componentes de covariância forem iguais a zero entre dois subconjuntos de $\mathbf{X}$, isto implica que eles são independentes (Casella e Berger, 2002). Esta propriedade só é valida se $\mathbf{X}$ tiver distribuição normal multivariada.

(4) A distribuição condicional de subconjuntos de $\mathbf{X}$ é normal multivariada. Por exemplo, suponha que $\mathbf{X}$ é particionado em $\mathbf{X}_1$ e $\mathbf{X}_2$ com dimensões $q \times 1$ e $(k - q) \times 1$ respectivamente. Da propriedade (1), temos que

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{e} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Com Σ_{11} e Σ_{12} tendo dimensões $q \times q$ e $(k-q) \times (k-q)$ respectivamente; e Σ_{12} e Σ_{21} tendo dimensões $q \times (k-q)$. Então as distribuições marginais são dadas por

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \qquad \text{e} \qquad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22});$$

e a distribuição condicional de $X_1|X_2 = x_2$ é dada por

$$X_1|X_2 = x_2 \sim \mathcal{N}_q\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

2.4 Processo Gaussiano

De acordo com Rasmussen e Williams (2006) o processo Gaussiano é uma extensão da distribuição normal multivariada para espaços de dimensão infinita e têm a seguinte definição.

Definição 2. *Uma coleção de varáveis aleatórias* $\{\mathbf{Y}(\mathbf{t}), t \in \mathrm{T}\}$ *indexadas por um conjunto de índices* T *é denominado de processo Gaussiano se para todo conjunto finito de índices* $\mathbf{t} = \{t_1, \ldots, t_k\} \subset \mathrm{T}$ *o vetor aleatório* $\mathbf{Y}(\mathbf{t}) = (Y_{t_1}, \ldots, Y_{t_k})$ *tem distribuição normal multivariada. Isto é, as distribuições finito-dimensionais do processo são normais multivariada. Neste caso, a distribuição do processo é unicamente determinada por sua função média* $\boldsymbol{\mu}_{\mathbf{y}} = (\mu_{t_1}, \ldots, \mu_{t_k}) : \mathrm{T} \to \mathbb{R}$ *e sua função de covariâncias* $\Sigma_{\mathbf{y}} : \mathrm{T} \times \mathsf{T} \to \mathbb{R}$ *composta por elementos* $\kappa_{t,t'}$*, definidas como*

$$\begin{aligned} \mu_t &= \mathbb{E}(Y_t), \\ \kappa(t,t') &= \mathbb{E}\left[(Y_t - \mu_t)(Y_{t'} - \mu_{t'})\right], \end{aligned}$$

para $t, t' \in \mathrm{T}$.

Usamos a notação $\mathbf{Y}(\mathbf{t}) \sim \mathcal{PG}\left(\boldsymbol{\mu}_{\mathbf{y}}, \Sigma_{\mathbf{y}}\right)$ para "dizer" que $\mathbf{Y}(\mathbf{t})$ é gerado de acordo com um processo Gaussiano, significando que $\mathbf{Y}(\mathbf{t}) \sim \mathcal{N}_k\left(\boldsymbol{\mu}_{\mathbf{y}}, \Sigma_{\mathbf{y}}\right)$, para $k > 1$.

2.4.1 Regressão por processo Gaussiano

Considere um cenário de regressão em que os dados são gerados de acordo com o seguinte modelo aditivo:

$$Y = f(x|\boldsymbol{\theta}) + \varepsilon, \tag{2.4}$$

onde $f(x|\boldsymbol{\theta})$ é um função indexada pelo parâmetro $\boldsymbol{\theta}$ (escalar ou vetor), com $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Dado uma amostra aleatória $\mathbf{Y} = (Y_1, \ldots, Y_k)$ do modelo em (2.4) para um vetor de

entradas $\mathbf{x} = (x_1, \ldots, x_k)$, o objetivo é obter estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$ para os parâmetros $\boldsymbol{\theta}$ e σ^2, repectivamente.

Sem perda de generalidade, considere $\boldsymbol{\theta} = (\beta_0, \beta_1)$ e $f(x|\boldsymbol{\theta}) = \beta_0 + \beta_1 x$, *i.e.*, um modelo de regressão linear simples, com $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, com $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, para $i = 1, \ldots, n$. O objetivo é obter estimativas $\hat{\boldsymbol{\theta}} = \left(\hat{\beta}_0, \hat{\beta}_1\right)$ e $\hat{\sigma}^2$ para os parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1)$ e σ^2, repectivamente. Neste ponto, pode-se obter as estimativas utilizando o método dos mínimos quadrados (Aldrich , 1997) ou o método da máxima verossimilhança (Casella e Berger, 2002).

Considerando a abordagem Bayesiana, as inferências sobre $\boldsymbol{\theta}$ são feitas utilizando a distribuição *a posteriori* para $\boldsymbol{\theta}$ (Gelman *et al.*, 2004). Para isto, primeiramente, precisamos especificar uma distribuição *a priori* para $\boldsymbol{\theta}$. Considerando $(\beta_0, \beta_1) \sim \mathcal{N}_2(\mathbf{0}, \Sigma_\beta)$, a distribuição *a posteriori* para (β_0, β_1) é dada por

$$\pi(\beta_0, \beta_1|\mathbf{y}, \mathbf{x}) \sim \mathcal{N}_2\left(\frac{\left(\Sigma^{-1} + \sigma^2\mathbf{x}\mathbf{x}^\top\right)\mathbf{x}\mathbf{y}}{\sigma^2}, \left(\Sigma^{-1} + \sigma^2\mathbf{x}\mathbf{x}^\top\right)^{-1}\right).$$

O desenvolvimento matemático para obtenção desta distribuição *a posteriori* está descrito em detalhes no artigo de Wanhg (2021). Como as inferências são feitas sobre $\boldsymbol{\theta} = (\beta_0, \beta_1) \in \Theta = \mathbb{R} \times \mathbb{R}$, dizemos que temos uma "regressão sobre o espaço paramétrico Θ". Como cada vetor de valores $(\beta_0, \beta_1) \in \Theta$ implica em uma função particular $f(x|\boldsymbol{\theta}) = \beta_0 + \beta_1 x$, então, uma distribuição de probabilidades sobre Θ é de maneira implicita uma distribuição de probabilidades sobre funções.

Na regressão por processo Gaussiano, a distribuição de probabilidades é dada de forma direta sobre o espaço $\mathcal{H}$ de todas as possíveis funções $f(x|\boldsymbol{\theta})$, com $\mathcal{H}$ podendo ter dimensão infinita. Assim, para entendermos como podemos parametrizar uma distribuição de probabilidades sobre um espaço de funções, considere $\mathbf{x} = (x_1, \ldots, x_k)$ um vetor de valores reais e $\mathbf{f}(\mathbf{x}) = (f(x_1), \ldots, f(x_k))$ o vetor de valores de uma função $f(\cdot) \in \mathcal{H}$ avaliada em $\mathbf{x}$. Ou seja, como o domínio $\mathbf{x}$ de qualquer função $f(\cdot) \in \mathcal{H}$ é composto de k valores, então, os k valores da função $f(\cdot)$ avaliada em $\mathbf{x}$ é dado pelo vetor k-dimensional $\mathbf{f}(\mathbf{x})$.

Assim, para especificar uma distribuição de probabilidades sobre as funções $f(\cdot) \in \mathcal{H}$, precisamos associar alguma função densidade de probabilidade a cada função de $\mathcal{H}$. Uma maneira simples de se fazer isto, é utilizando a relação um-a-um que existe entre toda $f(\cdot) \in \mathcal{H}$ e sua representação vetorial $\mathbf{f}(\mathbf{x}) = (f(x_1), \ldots, f(x_k))$. Assim, se considerarmos

que $\mathbf{f}(\mathbf{x}) \sim \mathcal{N}_k(\mathbf{m}, \Sigma_\mathbf{f})$, isto implica que a função densidade de probabilidade é dada por:

$$g[\mathbf{f}(\mathbf{x})|\mathbf{m}, \Sigma_\mathbf{f}] = (2\pi)^{k/2}|\Sigma_\mathbf{f}|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{f}(\mathbf{x}) - \mathbf{m})^\top \Sigma_\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{m})\right\}.$$

Neste caso, dizemos que o vetor $\mathbf{f}(\mathbf{x})$ é gerado de acordo com um processo Gaussiano de parâmetros $\mathbf{m}$ e $\Sigma_\mathbf{f}$; e escrevemos $\mathbf{f}(\mathbf{x})|\mathbf{m}, \Sigma_\mathbf{f} \sim PG(\mathbf{m}, \Sigma_\mathbf{f})$.

Ou seja, em um PG, assumimos que o vetor de valores $\mathbf{f}(\mathbf{x})$ é gerado de uma distribuição normal k-variada com vetor de médias $\mathbf{m}$ e matriz de covariâncias $\Sigma_\mathbf{f}$, independentemente de quem seja a função $f(\cdot) \in \mathcal{H}$. Utilizando a notação matricial, temos que

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_k) \end{bmatrix} \sim \mathcal{N}_k \left(\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix}, \begin{bmatrix} \sigma_\mathbf{f}(x_1, x_1) & \sigma_\mathbf{f}(x_1, x_2) & \dots & \sigma_\mathbf{f}(x_1, x_k) \\ \sigma_\mathbf{f}(x_2, x_1) & \sigma_\mathbf{f}(x_2, x_2) & \dots & \sigma_\mathbf{f}(x_2, x_k) \\ \vdots & \vdots & \dots & \vdots \\ \sigma_\mathbf{f}(x_k, x_1) & \sigma_\mathbf{f}(x_k, x_2) & \dots & \sigma_\mathbf{f}(x_k, x_k) \end{bmatrix} \right)$$

em que, o vetor de médias $\mathbf{m} = (m_1, \dots, m_k)^\top$ reflete o valor esperado da função $f(\cdot)$ para os valores $\mathbf{x}$, *i.e.*, $\mathbf{m}(\mathbf{x}) = \mathbb{E}[\mathbf{f}(\mathbf{x})]$; e $\Sigma_\mathbf{f}$ é a matriz de covariâncias composta pelos elementos $\sigma_\mathbf{f}(x, x') = \mathbb{C}ov\,(f(x), f(x'))$, para $x, x' \in \mathbf{x}$.

Neste ponto da modelagem, é usual fixar $\mathbf{m} = (m_1, \dots, m_k) = (0, \dots, 0) = \mathbf{0}$ para representar um conhecimento *a priori* não informativo sobre o valor esperado de $\mathbf{f}(\mathbf{t})$ e também para simplificar os cálculos para obtenção da distribuição *a posteriori* de $\mathbf{f}(\mathbf{x})$. Sob este cenário, as inferências são feitas com base na escolha da matriz de covariâncias $\Sigma_\mathbf{f}$; que também é denominada de função *kernel* do PG Rasmussen e Williams (2006).

De acordo com Rasmussen e Williams (2006), a escolha da função *kernel* $\Sigma_\mathbf{f}$ tem grande influência sobre os resultados obtidos. Em geral, a escolha de uma função *kernel* $\Sigma_\mathbf{f}$ apropriada é feita com base no que se espera em relação a padrão e suavidade da curva da função geradora dos dados. Uma suposição bastante sensata e utilizada para se definir um *kernel* $\Sigma_\mathbf{f}$ é que a correlação entre os valores $f(x)$ e $f(x')$ decresce a medida que a distâncias dos valores x e x' aumenta.

Uma função *kernel* muito utilizada e que cumpre essa suposição é o *squared exponential kernel*, em que, cada elemento é calculado de acordo com a seguinte expressão:

$$\sigma_\mathbf{f}(x, x') = \mathbb{C}ov\,(f(x), f(x')) = \eta^2 \exp\left\{-\frac{(x - x')^2}{2\nu^2}\right\},$$

para $\eta, \nu > 0$. O parâmetro η controla o quão distante os valores gerados estão da sua média. Assim, valores pequenos para η caracterizam funções que ficam próximas de

seu valor médio; enquanto que, valores maiores permitem maior variação. O parâmetro ν controla o quão suave é a função obtida ligando os pontos. Valores pequenos de ν significa que os valores das funções podem mudar rapidamente; enquanto que, valores grandes caracterizam funções que mudam de forma mais lenta (mais suaves).

Contudo, na literatura, temos a descrição de diversas outra funções de covariância que podemos utilizar para definir $\Sigma_{\mathbf{f}}$. Entre elas: o *rational quadratic kernel*, o *polynomial kernel*, o *Matérn kernel*, entre outros. Para maiores detalhes sobre funções de covariância ver Rasmussen e Williams (2006). No restante desta dissertação, optamos por usar o *squared exponential kernel*, devido a sua simplicidade de uso e eficiência, como discutido por Schulz *et al.* (2018) e Saraiva *et al.* (2023). No entanto, o procedimento de estimação descrito nas próximas seções podem ser facilmente adaptados para outros tipos de funções de covariâncias.

Definido o vetor de médias $\mathbf{m}$ e a função *kernel* $\Sigma_{\mathbf{f}}$, podemos gerar valores aleatórios para o vetor $\mathbf{f}(\mathbf{x})$ tanto da distribuição *a priori* quanto da distribuição *a posteriori*. Gerado os valores, podemos estimar a curva da função $f(\cdot)$, construindo o gráfico dos pontos $(x_i, f^{(g)}(x_i))$ conectados por linhas, onde $f^{(g)}(x_i)$ é o valor gerado, para $i = 1, \ldots, k$. No próximo Capítulo utilizaremos o processo Gaussiano na modelagem da função de crescimento assiada a geração de energia solar ao longo de um dia e apresentamos um algoritmo para gerar valores aleatórios da distribuição *a posteriori* de $\mathbf{f}(\mathbf{x})$.

Capítulo 3

Dados e modelagem

Neste capítulo apresentamos os dados que motivaram o desenvolvimento da modelagem proposta. O interesse na modelagem desse tipo de dados se deve basicamente a três fatos:

(i) ser capaz de informar ao operador do sistema elétrico uma predição do quanto de energia será inserida no sistema;

(ii) identificar possíveis problemas nos módulos fotovoltaicos;

(iii) quantificar perdas devido a sujidade presente nas placas solares.

3.1 Dados

Em qualquer análise e/ou modelagem estatística, um ponto importante é o conjunto de dados que será usado para se fazer inferências sobre os parâmetros de interesse. Nesta dissertação, utilizamos um conjunto de dados referente a geração de energia solar fotovoltaica obtido de um experimento realizado em uma mini-usina fotovoltaica instalada no campus de Campo Grande da Universidade Federal de Mato Grosso do Sul. Esse conjunto de dados está disponível no *website* `https://github.com/lscad-facom-ufms/Solar2` e detalhes sobre o experimento estão descritos no artigo de Souza *et al.* (2022).

O conjunto de dados $\mathbb{D}$ usado para se fazer inferências sobre os parâmetros do modelo Bayesiano hierárquico proposto contém o valor da potência solar gerada e registrada em $k = 74$ instantes de tempo de um dia, durante um período de $N = 19$ dias.

A Figura 3.1, apresenta um "recorte" da planilha de dados $\mathbb{D}$. Na primeira coluna, temos a indicação do dia. Na segunda coluna, temos o instante de tempo do dia que foi registrado os valores de interesse. Para cada dia i, temos o registro de $k = 74$ valores, para $i = 1, \ldots, N$. Na terceira coluna, temos os valores registrados para a potência gerada em kWh.

	A	B	C
1	Dia	Tempo	Potência gerada
2	1	1	63.15
3	1	2	140.31
4	1	3	273.65
5	1	4	435.1
6	1	5	620.06
7	1	6	905.04
8	1	7	1276.98
9	1	8	1675.41
10	1	9	2080.53

Figura 3.1: Recorte do conjunto de dados $\mathbb{D}$.

A Figura 3.2 ilustra um gráfico em barras com o total de energia gerada em cada um dos $N = 19$ dias; e a Tabela 3.1 apresenta um resumo desses valores. O valor mínimo foi de 136.955 kWh, a média foi de 289.749 kWh e o máximo foi de 357.189 kWh.

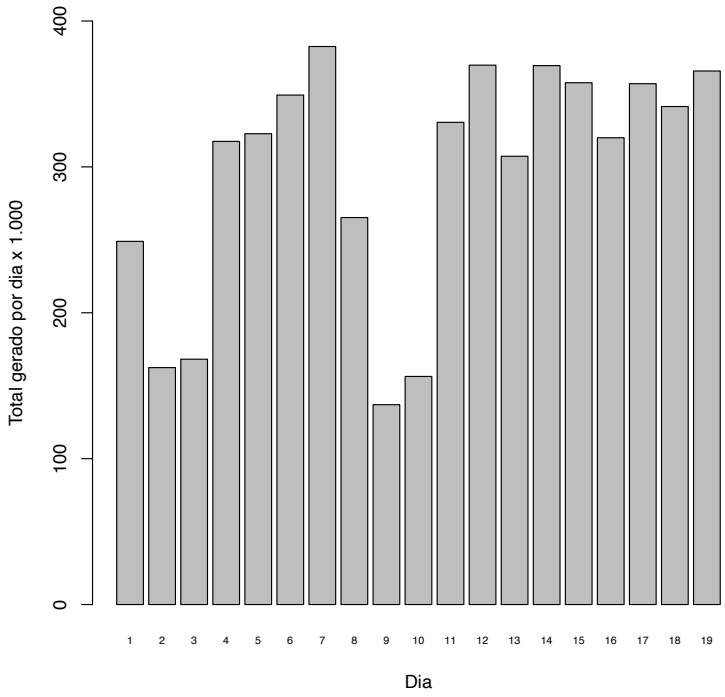


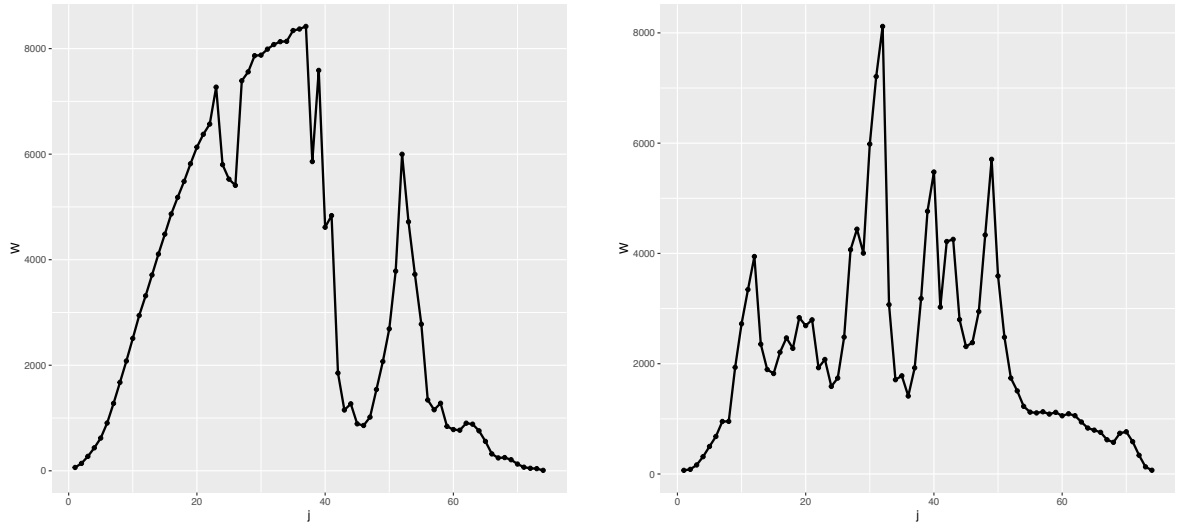
Figura 3.2: Gráfico do total de energia gerada por dia.

Tabela 3.1: Valores descritivos.

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
136,995	228,764	321,368	289,749	357,189	357,189

Como o experimento foi realizado por N dias, denote por W_{it} a potência registrada no t-ésimo instante de tempo do dia i, para $i = 1, \ldots, N$ e $t = 1, \ldots, k$. Assim, o vetor $\mathbf{W}_i = (W_{i1}, \ldots, W_{ik})$ são as medições registradas no dia i, para $i = 1, \ldots, N$.

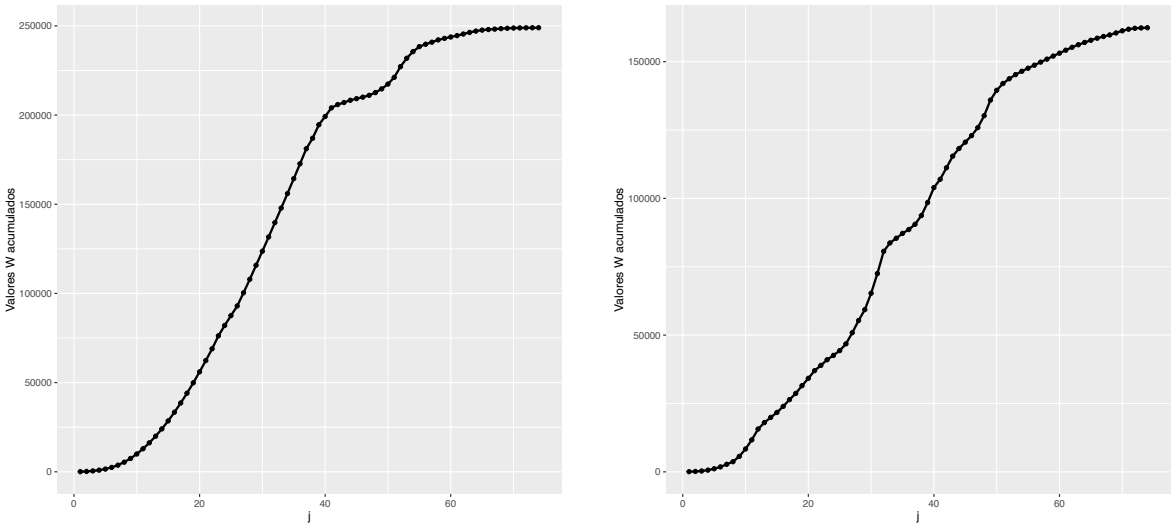
Como ilustração dos valores registrados, a Figura 3.3, mostra os valores de potência gerada nos dias 1 e 2 ao longo do tempo. Note que, os valores registrados apresentam alta variabilidade, o que dificulta o processo de modelagem. Para os outros dias, os valores registrados apresentam comportamento similar. Devido a isto, optamos por modelar os valores acumulados, pois estes apresentam um comportamento mais estável.



(a) Dia 1 (b) Dia 2

Figura 3.3: Potência gerada nos dias 1 e 2.

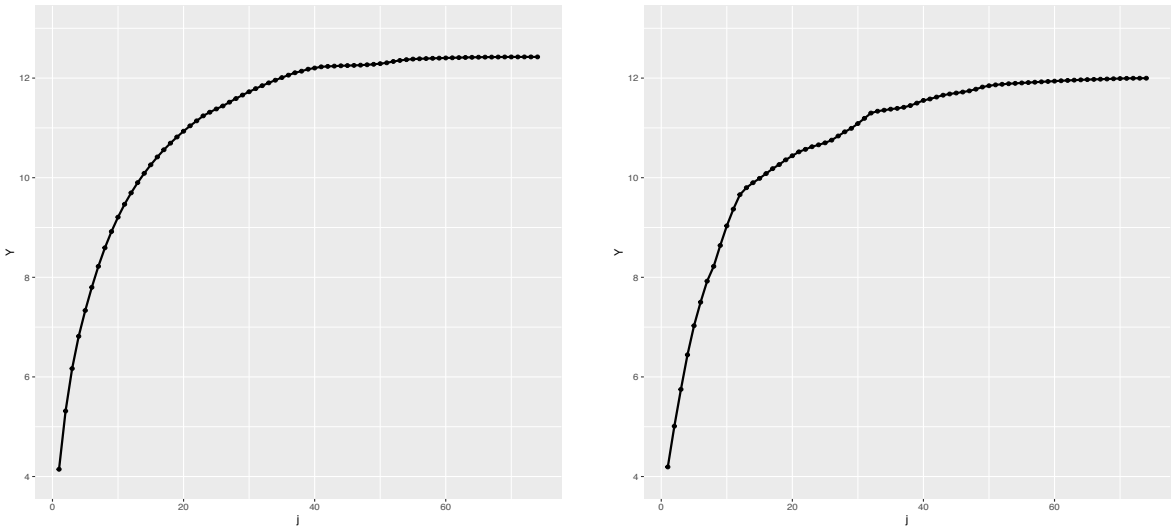
Assim, considere $W_{it}^{ac} = \sum_{t'=1}^{t} W_{it}$ os valores acumulados da energia solar gerada até o t-ésimo instante de tempo do i-ésimo dia e $\mathbf{W}_i^{ac} = (W_{i1}^{ac}, \ldots, W_{ik}^{ac})$ o vetor de valores acumulados, para $i = 1, \ldots, N$ e $t = 1, \ldots, k$. Note que, $W_{ik}^{ac} = \sum_{t=1}^{k} W_{it'}$ é a quantidade total gerada no dia i, para $i = 1 \ldots, N$. A Figura 3.4, apresenta os valores W^{ac} registrados nos dias 1 e 2. Note que, temos um comportamento mais estável e padronizado. Contudo, como muitos dos valores de W^{ac} estão na escala de 100.000, optamos por tomar o logaritmo dos valores acumulados para evitar problemas computacionais.



(a) Dia 1 (b) Dia 2

Figura 3.4: Potência gerada nos dias 1 e 2 de forma acumulada.

Assim, considere $Y_{it} = log(W_{it}^{ac})$ e $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik})$ o vetor de valores registrados no i-ésimo dia, para $i = 1, \ldots, N$ e $t = 1, \ldots, k$. A Figura 3.5 apresenta os gráficos da Figura 3.4 na escala logaritmica. Note que, temos valores com comportamento estável, padronizado e escala de valores mais simples de se trabalhar computacionalmente.

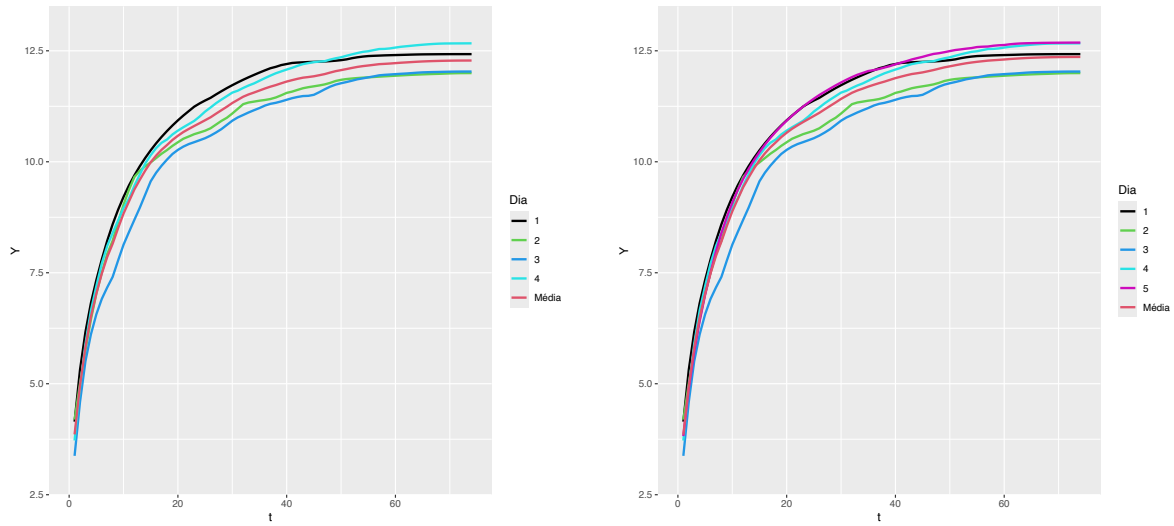


(a) Dia 1 (b) Dia 2

Figura 3.5: Potência gerada nos dias 1 e 2 de forma acumulada na escala logaritmica.

A partir deste ponto da modelagem e sem perda de generalidade, considere $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ os dados registrados nos primeiros n dias de realização do experimento, para $n < N$. Ou seja, $\mathbf{y}$ é uma matriz de dimensão $n \times k$, em que, cada linha contém os

valores y's registrados no dia i, para $i = 1, \ldots, n$. Por exemplo, fixando $n = 4$, a Figura 3.6(a) mostra o gráfico em linhas dos valores Y's registrados nos quatro primeiros dias, *i.e.*, os valores $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4)$, em que, a linha em vermelho representa a média dos valores em cada instante de tempo t, para $t = 1, \ldots, k$. Similarmente, a Figura 3.6(b), apresenta o gráfico em linhas dos valores $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5)$, *i.e.*, para $n = 5$. Nosso principal interesse é no desenvolvimento de um método de predição dos valores Y's que serão gerados no dia $(n + 1)$.



(a) $n = 4$ (b) $n = 5$

Figura 3.6: Valores $\mathbf{y}$ para $n = 4$ e $n = 5$.

3.2 Modelagem

Tomando como base a curva média dos gráficos da Figura 3.6 (linha em vermelho), considere $f(\cdot)$ um função de crescimento não-linear e $\mathbf{f}(\mathbf{t}) = (f(1), \ldots, f(k))$ um vetor composto pelos valores da função $f(\cdot)$ avaliada em k instantes de tempo, para $\mathbf{t} = (1, \ldots, k)$. Além disso, assuma que, a função de crescimento para o dia i é proporcional a função $f(\cdot)$, *i.e.*, $f_i(\cdot) = C_i f(\cdot)$ para $C_i > 0$ e $i = 1, \ldots, n$. Em outras palavras, estamos assumindo que existe uma função de crescimento $f(\cdot)$ cujo gráfico é uma curva que representa a curva média de n curvas; sendo a curva do dia i proporcional a curva média, para $i = 1, \ldots, n$.

Considere que os valores registrados no i-ésimo dia são gerados de acordo com o seguinte modelo aditivo:

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik}) \sim C_i \cdot \mathbf{f}(\mathbf{t}) + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

para $C_i > 0$, onde, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ik})$ é um vetor de erros aleatórios, para $i = 1, \ldots, n$. Em notação matricial, temos

$$\begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{ik} \end{bmatrix} = \begin{bmatrix} C_i \cdot f(1) \\ \vdots \\ C_i \cdot f(k) \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ik} \end{bmatrix},$$

para $i = 1, \ldots, n$. Como os valores $Y's$ são resgistrados ao longo do tempo, então, apresentam algum tipo de correlação. Assim considere que o vetor de erros aleatórios $\boldsymbol{\varepsilon}$ é gerado de acordo com uma distribuição normal multivariada de dimensão k, com vetor de médias $\mathbf{0} = (0, \ldots, 0)$ e matriz de covariâncias Σ (de dimensão $k \times k$) dada por

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1k} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_{kk} \end{bmatrix} \tag{3.2}$$

onde $\sigma_{tt'} = \sigma_{\boldsymbol{\varepsilon}}(t, t') = \mathbb{C}ov\,(\varepsilon_{it}, \varepsilon_{it'})$, para $t, t' = 1, \ldots, k$ e $i = n - d + 1, \ldots, n$.

Para completar o modelo em (3.1), poderíamos fixar $f(\cdot)$ como sendo uma função matemática conhecida, como, por exemplo, as funções de crescimento logística (Cramer, 2004) ou de Gompertz (Gompertz, 1825), entre outras. Três pontos importantes sobre esta abordagem paramétrica são:

(i) a análise fica limitada a função $f(\cdot)$ (ou funções) previamente escolhida pelo analista;

(ii) a complexidade e/ou flexibilidade das funções $f(\cdot)$ consideradas é limitada pelo número de parâmetros presentes nestas funções; e

(iii) podem existir diferentes funções que se ajustam igualmente bem aos valores registrados.

Uma solução usual, consiste em ajustar um conjunto de funções candidatas; e em seguida escolher o melhor modelo utilizando um critério de seleção de modelos, tal como, o AIC ou o BIC. No entanto, as questões (ii) e (iii) ainda permanecem.

Assim, para evitar que a modelagem proposta fique restrita a uma função paramétrica $f(\cdot)$ previamante especificada por um analista, assumimos que $f(\cdot)$ é uma função desconhecida, com o vetor de valores $\mathbf{f(t)} = (f(1), \ldots, f(k))$ sendo parâmetros do modelo, que precisam ser estimados a partir do conjunto de dados observados. Assim, do modelo em (3.1), os parâmetros de interesse são $\boldsymbol{\theta} = (\mathbf{f(t)}, \Sigma, \mathbf{C})$, onde $\mathbf{f(t)} = (f(1), \ldots, f(k))$, Σ é a matriz de covariâncias dada em (3.2) e $\mathbf{C} = (C_1, \ldots, C_n)$.

3.2.1 Estimação dos parâmetros

Para estimar $\boldsymbol{\theta}$ utilizando os dados $\mathbf{y}$, assumimos uma abordagem Bayesiana com um processo Gaussiano *a priori* sobre $\mathbf{f(t)}$, denotado por $\mathbf{f(t)}|\mathbf{m}, \Sigma_\mathbf{f} \sim \mathcal{PG}(\mathbf{m}, \Sigma_\mathbf{f})$. Isto significa que estamos considerando $f(\cdot)$ como sendo uma função desconhecida; mas, com o vetor de valores $\mathbf{f(t)} = (f(1), \ldots, f(k))$ sendo gerado de uma distribuição normal k-variada com vetor de médias $\mathbf{m}$ e matriz de covariâncias $\Sigma_\mathbf{f}$ composta pelos elementos $\sigma_\mathbf{f}(t, t') = \mathbb{C}ov(f(t), f(t'))$, para $t, t' = 1, \ldots, k$. Para Σ, assumimos uma distribuição *a priori* conjugada dada pela distribuição inversa-Wishart com vetor de parâmetros $(\delta, \mathbb{V})$, sendo $\delta > 0$ e $\mathbb{V}$ uma matrix de dimensão $k \times k$; e para C_i assumimos uma distribuição *a priori* dada pela distribuição normal trucada a esquerda de 0 e com parâmetros $\mu_\mathbf{c}$ e $\sigma^2_\mathbf{c}$, para $i = 1, \ldots, n$. Utilizando uma notação hierárquica, o modelo Bayesiano proposto é dado por

$$\begin{aligned} \mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik}) \,|\mathbf{f(t)}, \Sigma, \mathbf{C} &\sim \mathcal{N}_k(C_i \cdot \mathbf{f(t)}, \Sigma) \\ \mathbf{f(t)}|\mathbf{m}, \Sigma_\mathbf{f} &\sim \mathcal{PG}(\mathbf{m}, \Sigma_\mathbf{f}) \\ \Sigma|\delta, \mathbb{V} &\sim \mathcal{IW}(\delta, \mathbb{V}) \\ C_i|\mu_\mathbf{c}, \sigma_c^2 &\sim \mathcal{N}trunc\left(0; \mu_\mathbf{c}, \sigma^2_\mathbf{c}\right), \end{aligned} \tag{3.3}$$

onde $\mathcal{N}_k(\cdot)$, $\mathcal{PG}(\cdot)$, $\mathcal{IW}(\cdot)$ e $\mathcal{N}trunc(0; \cdot)$ representam a distribuição normal k-variada, o processo Gaussiano, a distribuição inversa-Wishart e a distribuição normal truncada a esquerda de 0, respetivamente; e $\mathbf{m}$, $\Sigma_\mathbf{f}$, δ, $\mathbb{V}$, $\mu_\mathbf{c}$ e σ_c^2 são hiperparâmetros conhecidos, para $i = 1, \ldots, n$.

Para completar o modelo em (3.3), fixamos:

(i) $\mathbf{m} = \mathbf{0}$ para representar o nosso conhecimento prévio não informativo sobre o valor esperado de $\mathbf{f(t)}$;

(ii) $\Sigma_\mathbf{f} = \lambda\mathbb{W}$, para $\lambda > 0$ e $\mathbb{W}$ é uma matriz de dimensão $k \times k$ composta pelos elementos $\kappa(t, t')$ calculados de acordo com o núcleo exponencial quadrático, *i.e.*,

$$\kappa(t, t') = \eta^2 \exp\left\{-\frac{(t - t')^2}{2\nu^2}\right\}, \tag{3.4}$$

para $\eta, \nu > 0$. Nesta expressão, o parâmetro η controla a distância entre os valores gerados e a sua média. Ou seja, valores "pequenos" de η caracterizam funções que estão próximas do seu valor médio; enquanto que valores maiores permitem uma

maior variação. O parâmetro ν controla a suavidade da função obtida pela ligação dos pontos. Ou seja, valores pequenos de ν significam que os valores da função podem mudar rapidamente; enquanto que valores grandes caracterizam funções que mudam mais lentamente (curvas mais suaves). Definimos $\lambda = 100$, $\eta = 1$ e $\nu = 1$ para obter uma distribuição *a priori* pouco informativa.

(iii) Finalizamos a especificação do modelo fixando $\delta = k$, $\mathbb{V} = 0,01 \cdot \mathbb{I}_k$, onde $\mathbb{I}_k$ é a matriz identidade de dimensão $k \times k$ e $\mu_{\mathbf{c}} = \sigma^2_{\mathbf{c}} = 1$.

Aplicando o teorema de Bayes, a distribuição *a posteriori* conjunta para $\boldsymbol{\theta} = (\mathbf{f}(\mathbf{t}), \Sigma, \mathbf{C})$ é dada por

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}) \propto \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}) \pi\left(\mathbf{f}(\mathbf{t})|\mathbf{m}, \Sigma_{\mathbf{f}}\right) \pi\left(\Sigma|\delta, \mathbb{V}\right) \pi\left(\mathbf{C}|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}}\right), \tag{3.5}$$

onde $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t})$ é a função de verosimilhança de uma distribuição normal k-variada com parâmetros $\mathbf{f}(\mathbf{t})$ e Σ, dada por

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}) = (2\pi)^{-nk/2} |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{y}_i - C_i\mathbf{f}(\mathbf{t})\right)^\top \Sigma^{-1} \left(\mathbf{y}_i - C_i\mathbf{f}(\mathbf{t})\right)\right\};$$

e $\pi(\cdot)$ representa uma função de densidade de probabilidade das distribuições *a priori*, dadas por

$$\begin{aligned}
\pi\left(\mathbf{f}(\mathbf{t})|\mathbf{m}, \Sigma_{\mathbf{f}}\right) &= (2\pi)^{-1/2} |\Sigma_{\mathbf{f}}|^{-1/2} \exp\left\{-\frac{1}{2} \left(\mathbf{f}(\mathbf{t}) - \mathbf{m}\right)^\top \Sigma_{\mathbf{f}}^{-1} \left(\mathbf{f}(\mathbf{t}) - \mathbf{m}\right)\right\}; \\
\pi\left(\Sigma|\delta, \mathbb{V}\right) &= \frac{|\mathbb{V}|^{\delta/2}}{2^{\delta k/2} \Gamma_k(\delta)} |\Sigma|^{-(\delta+k+1)/2} \exp\left\{-\frac{1}{2} tr\left(\mathbb{V}\Sigma^{-1}\right)\right\}; \\
\pi\left(\mathbf{C}|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}}\right) &= \prod_{i=1}^{n} f_{trunc}(C_i|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}});
\end{aligned}$$

onde $f_{trunc}(C_i|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}}) = \frac{f(C_i|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}})}{\int_0^{+\infty} f(C_i|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}})}$, para $f(C_i|\mu_{\mathbf{c}}, \sigma^2_{\mathbf{c}}) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathbf{c}}}} \exp\left\{-\frac{1}{2\sigma^2_{\mathbf{c}}} \left(C_i - \mu_{\mathbf{c}}\right)^2\right\}$, para $i = 1, \ldots, n$.

No entanto, a distribuição *a posteriori* conjunta em (3.5) não tem uma forma matemática conhecida que nos permita gerar diretamente valores aleatórios desta distribuição de probabilidades. Dessa forma, precisamos utilizar um algoritmo que gere números aleatórios desta distribuição de maneira indireta. Neste trabalho, optamos por gerar valores de $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t})$ utilizando o algoritmo *Gibbs sampling*, devido a sua simplicidade de implementação e eficiência. Detalhes sobre o desenvolvimento matemático deste algoritmo estão descritos nos artigos dos autores Geman e Geman (1984) e Gelfand e Smith (1990). De

maneira resumida, esse algoritmo gera valores da distribuição *a posteriori* conjunta, de forma indireta, utilizando as distribuições *a posteriori* condicionais, desde que estas sejam conhecidas.

Para o modelo Bayesiano hierárquico proposto em (3.3), as distribuições *a posteriori* condicionais são conhecidas e são dadas por (Ver Apêndices 3, 4 e 5):

$$\mathbf{f}(\mathbf{t})|\bullet \sim \mathcal{PG}\left(\Sigma^{-1}\left(\sum_{i=1}^{n} C_i\Sigma^{-1} + \Sigma_{\mathbf{f}}^{-1}\right)^{-1} n\overline{\mathbf{y}}, \left(\sum_{i=1}^{n} C_i\Sigma^{-1} + \Sigma_{\mathbf{f}}^{-1}\right)^{-1}\right) \quad (3.6)$$

$$\Sigma|\mathbf{y},\mathbf{t},\bullet \sim \mathcal{IW}\left(\delta + k + n, \mathbb{V} + \sum_{i=1}^{n}\left(\mathbf{y}_i - C_i\mathbf{f}(\mathbf{t})\right)^{\top}\left(\mathbf{y}_i - C_i\mathbf{f}(\mathbf{t})\right)\right) \quad (3.7)$$

$$C_i|\mathbf{y},\mathbf{t},\bullet \sim \mathcal{N}trunc\left(0, \frac{\mathbf{f}(\mathbf{t})^{\top}\Sigma^{-1}\mathbf{y}_i + 1)}{\mathbf{f}(\mathbf{t})^{\top}\Sigma^{-1}\mathbf{f}(\mathbf{t}) + 1}, \frac{\sigma_{\mathbf{c}}^2}{\mathbf{f}(\mathbf{t})^{\top}\Sigma^{-1}\mathbf{f}(\mathbf{t}) + \mu_{\mathbf{c}}}\right), \quad (3.8)$$

onde $\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} C_i\mathbf{y}_i$ e o símbolo $\bullet$ representa todos os outros parâmetros.

Utilizando as distribuições *a posteriori* condicionais, implementamos um algoritmo *Gibbs sampling* de acordo com os passos descritos no Algoritmo 1.

Algoritmo 1 : Algoritmo Gibbs sampling

1: Considere uma cadeia de Markov com espaço de estados composto pelo vetor de parâmetros $\boldsymbol{\theta} = (\mathbf{f}(\mathbf{t}), \Sigma, \mathbf{C})$;

2: Inicialize o algoritmo com os valores $\boldsymbol{\theta} = \left(\mathbf{f}(\mathbf{t})^{(0)}, \Sigma^{(0)}, \mathbf{C}^{(0)}\right)$;

3: **Procedimento:** Para a l-ésima iteração do algoritmo, $l = 1, \ldots, L$:

4: Gere $\mathbf{f}^{(l)}$ da distribuição condicional em (3.6), dado os valores $\Sigma^{(l-1)}$ e $\mathbf{C}^{(l-1)}$;

5: Gere $\Sigma^{(l)}$ da distribuição condicional em (3.7), dado no valor $\mathbf{f}(\mathbf{t})^{(l)}$;

6: Gere $C_i^{(l)}$ da distribuição condicional em (3.8), dado os valores $\mathbf{f}^{(l)}$ e $\Sigma^{(l)}$.

Após executar as L iterações do algoritmo *Gibbs sampling*, descartamos as primeiras B iterações como um *burn-in* e consideramos "saltos" de tamanho J, *i.e.*, para cada "bloco" de J iterações do algoritmo, selecionamos os valores gerados na última iteração, formando uma sub-sequência de tamanho $S = [(L - B)/J]$ para fazer inferências sobre $\boldsymbol{\theta}$. As estimativas dos parâmetros de interesse são dadas pela média dos valores gerados, *i.e.*,

$$\hat{\mathbf{f}}(\mathbf{t}) = \frac{1}{S}\sum_{l=1}^{S}\mathbf{f}(\mathbf{t})^{(M(l))}, \quad \hat{\Sigma} = \frac{1}{S}\sum_{l=1}^{S}\Sigma^{(M(l))} \quad \text{and} \quad \hat{C}_i = \frac{1}{S}\sum_{l=1}^{S} C_i^{(M(l))}$$

para $i = 1, \ldots, n$, onde $\mathbf{f}(\mathbf{t})^{(M(l))}$, $\Sigma^{(M(l))}$ e $\mathbf{C}^{(M(l))}$ são os valores gerados para os parâmetros $\mathbf{f}(\mathbf{t})$, Σ e $\mathbf{C}$ na $M(l) = (B+1+(l-1)\cdot J)$-ésima iteração do algoritmo, respetivamente, para $l = 1, \ldots, S$. Um intervalo de credibilidade de 95% para cada um dos parâmetros é dado pelos percentís $2,5\%$ e $97,5\%$ dos valores gerados. A curva estimada de $f(\cdot)$ é obtida fazendo o gráfico dos pontos $\left(t, \hat{f}(t)\right)$ conectados por linhas, para $t = 1, \ldots, k$.

3.2.2 Estudo de simulação 1

Para ilustrar o desempenho da modelagem proposta, desenvolvemos um estudo de simulação. Para gerar um conjunto de dados, consideramos $f(\cdot)$ como sendo a função log-Gompertz de parâmetros α_1, α_2 e α_3 e com a seguinte parametrização:

$$f(t) = log(\alpha_1) - exp\left\{\alpha_2 - \alpha_3 t\right\},$$

para $t > 0$. Fixamos $\alpha_1 = 12$, $\alpha_2 = 2$ e $\alpha_3 = 0,1$. Obtemos a matriz de covariância Σ calculando cada termo de acordo com o núcleo exponencial quadrático dado em 3.2 para $\eta^2 = 0,01$ e $\nu^2 = 10$.

O procedimento para gerar um conjunto de dados artificial é dado pelos seguintes quatro passos:

(i) Fixe o número de dias n e o número de instantes de tempo k.

(ii) Faça $\mathbf{t} = (1, \ldots, k)$ e calcule $\mathbf{f}(\mathbf{t}) = (f(1), \ldots, f(k))$, onde $f(t)$ é dado pela função log-Gompertz descrita acima.

(iii) Fixe os valores de $\mathbf{C} = (C_1, \ldots, C_n)$, para $C_i > 0$ e $i = 1, \ldots, n$.

(iv) Gere $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik}) \sim \mathcal{N}_k\left(\mathbf{f}(\mathbf{t}), \Sigma\right)$, para $i = 1, \ldots, n$.

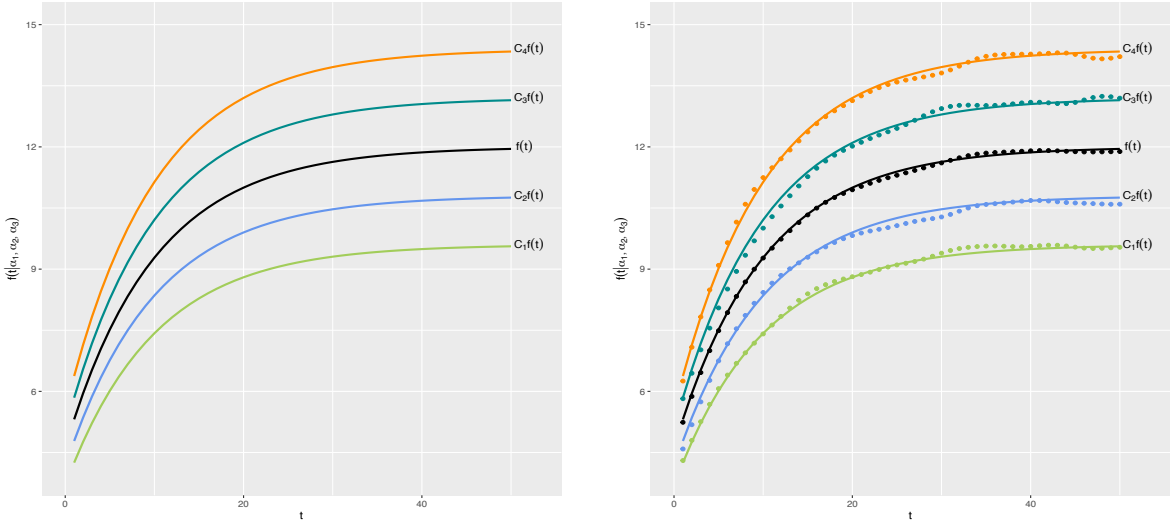
Para simplificar a visualização dos resultados, fixamos $n = 4$ e $k = 50$, *i.e.*, quatro curvas com $\mathbf{C} = (C_1, C_2, C_3, C_4) = (0,8; 0,9; 1,1; 1,2)$. A Figura 3.7(a) apresenta a curva de $f(t)$, a que chamamos "curva média", e as curvas dos dias 1, 2, 3 e 4. A Figura 3.7(b) mostra os valores $\mathbf{y}$'s gerados (símbolos $\bullet$ coloridos) para cada um dos dias, em que, os símbolos $\bullet$ na cor preta são as médias dos valores gerados.

Gerado o conjunto de dados, aplicamos o algoritmo *Gibbs sampling* com $L = 55.000$ iterações, $B = 5.000$ e $J = 10$. Assim, obtemos uma amostra *a posteriori* de tamanho

$S = 5.000$ para se fazer inferências. Para verificar quão próximo os valores estimados $\hat{f}(t)$ são dos valores reais $f(t)$, calculamos o erro percentual absoluto (EPA), dado por

$$EPA(d_t) = \frac{|f(t) - \hat{f}(t)|}{f(t)} \cdot 100,$$

para $t = 1, \ldots, k$.



(a) Curvas

(b) Curvas e valores gerados

Figura 3.7: Curvas e valores gerados.

A Figura 3.8(a) mostra o gráfico da curva de $f(t)$ (linha em negrito) e a curva estimada pelo método proposto (linha em vermelho) com uma banda de credibilidade de 95% (região em vermelho). A Figura 3.8(b) mostra o gráfico dos valores de $EPA(\mathbf{d}) = (EPA(d_1), \ldots, EPA(d_k))$. A Tabela 3.2 mostra as medidas de resumo dos valores do $EPA(\mathbf{d})$. Os valores de $EPA(\mathbf{d})$ variaram de $0,0145$ a $1,4481$ com um erro percentual absoluto médio de $0,4568$. Ou seja, os valores etimados são bem próximos dos valores reais, haja vista que o valores $EPA(\mathbf{d})$ são próximos a zero.

Tabela 3.2: Medidas resumo dos erros percentuais absolutos.

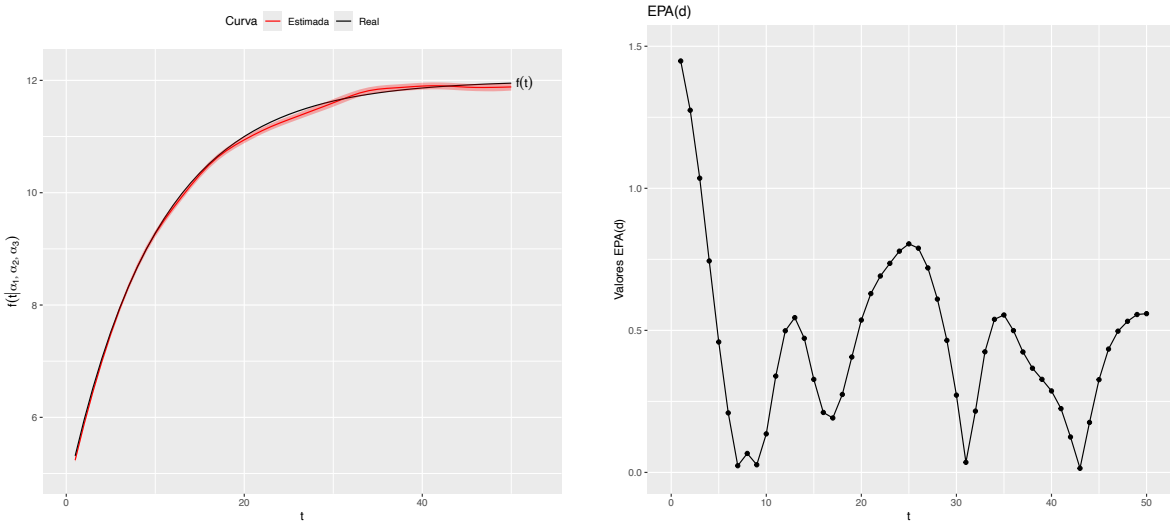
Media	Mínimo	1^o quartil	Mediana	Média	3^o quartil	Máximo
$EPA(\mathbf{d})$	0,0145	0,2364	0,4466	0,44568	0,5580	1,4481

De maneira similar, obtemos a curva estimada para o i-ésimo dia construindo o gráfico dos pontos $(t, \hat{y}_{it})$ conectados por linhas, para $\hat{y}_{it} = \hat{C}_i \hat{f}(t)$, $i = 1, \ldots, n$ e $t = 1, \ldots, k$. Para este caso, calculamos os valores EPA entre o valor real $C_i f(t)$ e o valor estimado

$\hat{y}_{it}$; e os valores EPA entre os valores gerados y_{it} e os valores estimados $\hat{y}_{it}$, dados, respectivamente, por:

$$EPA\left(d_{it}\right) = \frac{|C_i f(t) - \hat{y}_{it}|}{C_i f(t)} \cdot 100 \qquad \text{e} \qquad EPA\left(e_{it}\right) = \frac{|y_{it} - \hat{y}_{it}|}{y_{it}} \cdot 100$$

para $i = 1, \ldots, n$ e $t = 1, \ldots, k$.



(a) Curva real e estimada

(b) Valores $EPA(\mathbf{d})$

Figura 3.8: Curvas e valores gerados.

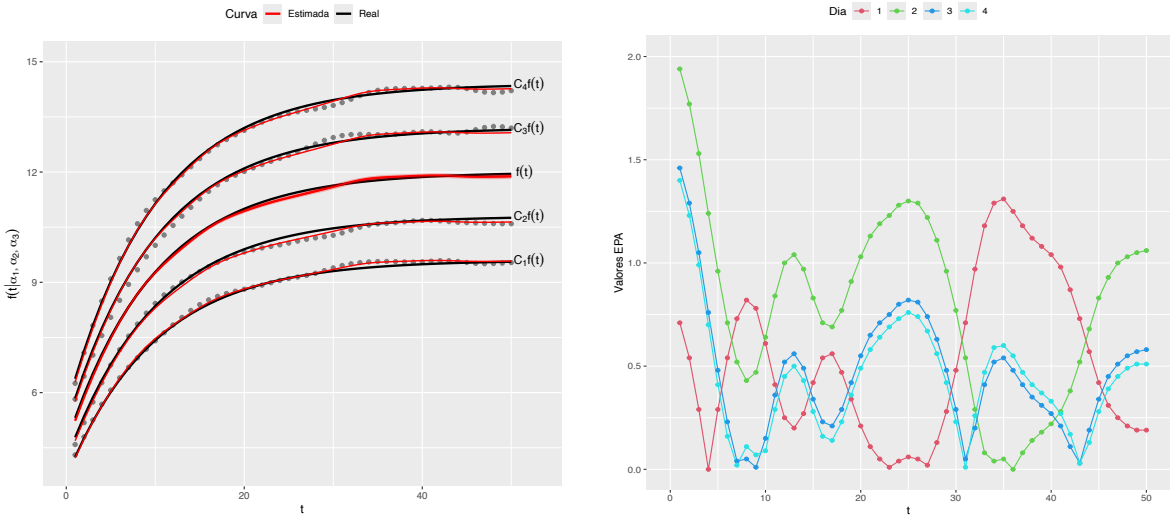
A Tabela 3.3 mostra os valores estimados e os intervalos de credibilidade de 95% para os parâmetros C_i, $i = 1, 2, 3, 4$. Note que, os valores estimados estão próximos aos valores reais. Além disso, os valores reais pertencem aos intervalos de credibilidade.

Tabela 3.3: Estimativas e intervalos de credibilidade (95%) para C_i, $i = 1, 2, 3, 4$.

Parâmetro	Valor real	Estimativa	Intervalo de credibilidade de 95%
C_1	0,8	0,8060	(0,7997, 0,8080)
C_2	0,9	0,8955	(0,8913, 0,9003)
C_3	1,1	1,0998	(1,0939, 1,1057)
C_4	1,2	1,2006	(1,1949, 1,2071)

A Figura 3.9(a) ilustra os gráficos das curvas de $C_i f(t)$ (linhas em negrito) e as curvas estimadas pelo método proposto (linhas em vermelho), em que, os simbolos $\bullet$ representam os valores $\mathbf{y}$'s gerados para cada dia. A Figura 3.9(b), mostra o gráfico dos valores $EPA(\mathbf{d}_i) = (EPA(d_{i1}, \ldots, EPA(d_{ik})))$ e a Tabela 3.4 mostra as medidas resumo dos valores do $EPA(\mathbf{d}_i)$, para $i = 1, \ldots, n$.

Note que, as curvas estimadas para cada um dos quatro dias estão satisfatoriamente próximas as curvas reais, uma vez que os valores EPA são todos inferiores a 2; significando que o erro percentual absoluto entre o valor real e o valor estimado é inferior a 2%.



(a) Curva real e estimada

(b) Valores $EPA(\mathbf{d})$

Figura 3.9: Curvas e valores gerados.

Tabela 3.4: Medidas resumo dos valores $EPA(\mathbf{d}_i)$, para $i = 1, 2, 3, 4$.

Medida	Mínimo	1^o quartil	Mediana	Média	3^o quartil	Máximo
$EPA(\mathbf{d}_1)$	0	0,2100	0,4450	0,5212	0,7675	1,3100
$EPA(\mathbf{d}_2)$	0	0,4825	0,8350	0,7972	1,0575	1,9400
$EPA(\mathbf{d}_3)$	0,0100	0,2400	0,4650	0,4650	0,5775	1,4600
$EPA(\mathbf{d}_4)$	0,0100	0,2375	0,4250	0,4364	1,5750	1,4000

A Figura 3.10, mostra o gráfico dos valores $EPA(\mathbf{e}_i) = (EPA(e_1), \ldots, EPA(e_k))$, e a Tabela 3.5 mostra as medidas resumo dos valores do $EPA(\mathbf{e}_i)$, para $i = 1, \ldots, n$. Note que, todos os valores $EPA(\mathbf{e}_i)$ são inferiores a $2,5$; significando que o erro percentual absoluto entre o valor gerado e o valor estimado é inferior a $2,5\%$. Isto indica que os valores estimados também estão satisfatoriamente próximos aos valores gerados.

Como as inferências para os parâmetros de interesse foram feitas utilizando uma amostra *a posteriori* obtida pela implementação de uma algoritmo do tipo MCMC, então, é importante verificar a convergência dos valores gerados. Uma maneira usual de se fazer esta checagem é através de métodos gráficos, em que, é construído gráficos dos valores gerados ou alguma medida associada aos valores gerados.

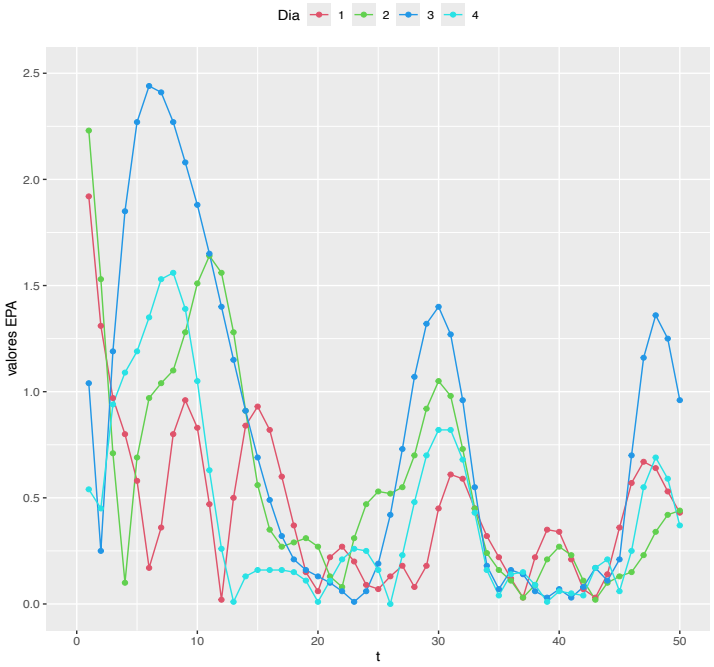


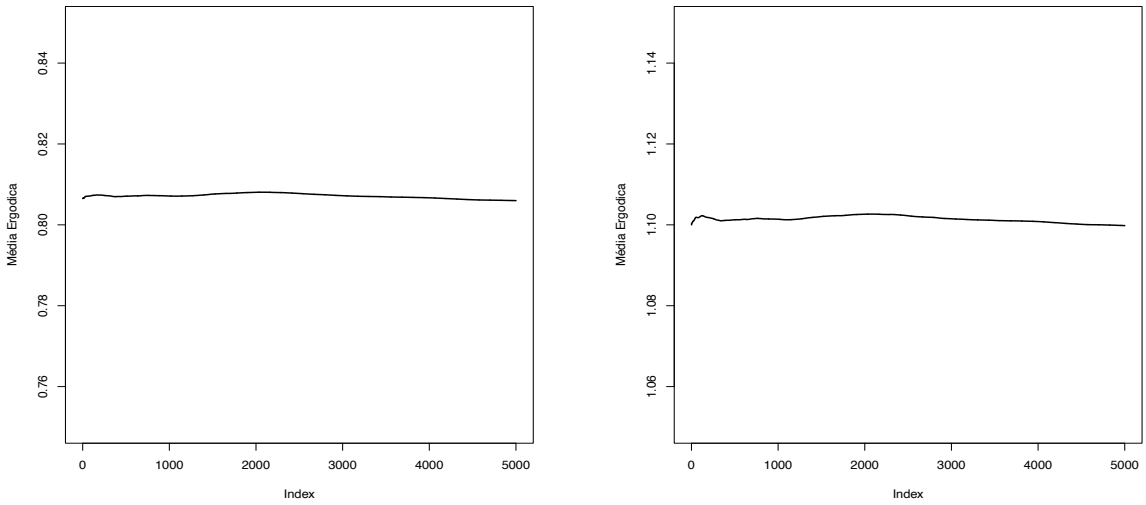
Figura 3.10: Valores $EPA(e_i)$, para $i = 1, 2, 3, 4$.

Tabela 3.5: Medidas resumo dos valores $EPA(e_i)$, para $i = 1, 2, 3, 4$.

Medida	Mínimo	1^o quartil	Mediana	Média	3^o quartil	Máximo
$EPA(\mathbf{e}_1)$	0,0200	0,1775	0,3600	0,4446	0,6075	1,9200
$EPA(\mathbf{e}_2)$	0,0200	0,2150	0,4300	0,5860	0,9175	2,2300
$EPA(\mathbf{e}_3)$	0,0100	0,1450	0,6200	0,7934	1,2650	2,4400
$EPA(\mathbf{e}_4)$	0	0,1325	0,2400	0,4330	0,3675	1,5600

Neste texto, verificamos a convegência dos valores gerados utilizando o gráfico da média ergódica (Gallavotti, 1995). Assim, para ilustrar a convegência do algoritmo *Gibbs sampling* implementado, selecionamos de maneira aleatória dois instantes de tempo t (t_1 e t_2) e construímos o gráfico da média ergódica (ME) dos valores gerados para $f(t_1)$ e $f(t_2)$. Os valores t_1 e t_2 sorteados foram $t_1 = 10$ e $t_2 = 30$.

A Figura 3.11, mostra os gráficos das médias ergódicas para os valores gerados para $f(10)$ e $f(30)$. Note que, não há razões para dúvidar da suposição de convergência, uma vez que, os valores da ME apresentam satisfatória estabilização. Também verificamos a convegência dos valores gerados para $f(1)$, $f(25)$ e $f(40)$, e os resultados são similares aos apresentados. Isto é, os valores da ME apresentam satisfatória estabilização ao longo das iterações do algoritmo *Gibbs sampling*.



(a) $f(10)$ (b) $f(30)$

Figura 3.11: Média ergódica dos valores gerados para $f(10)$ e $f(30)$.

3.3 Predição

Além de obter as estimativas $\hat{\boldsymbol{\theta}} = \left(\hat{\mathbf{f}}(\mathbf{t}), \hat{\Sigma}, \hat{\mathbf{C}}\right)$ para os parâmetros $\boldsymbol{\theta} = (\mathbf{f}(\mathbf{t}), \Sigma, \mathbf{C})$, outro interesse é na predição de $\mathbf{Y}_{n+1} = \left(Y_{1(n+1)}, \ldots, Y_{k(n+1)}\right)$, *i.e.*, os valores que serão registrados no $(n+1)$-ésimo dia. Isso pode ser feito utilizando a distribuição preditiva, dada por:

$$\pi(\mathbf{Y}_{n+1}|\mathbf{y}, \mathbf{t}) = \int \pi\left(\mathbf{Y}_{n+1}|\mathbf{y}, \mathbf{t}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}\right) d\boldsymbol{\theta} \tag{3.9}$$

onde $\pi\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}\right)$ é a distribuição *a posteriori* conjunta para $\boldsymbol{\theta}$, dada na Equação (3.5). No entanto, esta integral não tem uma solução matemática conhecida. Devido a este fato, propomos um algoritmo MCMC para obter uma aproximação para esta integral.

Do modelo aditivo em (3.1), temos que, a distribuição marginal de $\mathbf{Y}_i$ é dada por uma distribuição normal k-variada com vetor de médias $\mathbf{0} = (0, \ldots, 0)^\top$ e matriz de covariância

$$C_i^2 \Sigma_{\mathbf{f}} + \Sigma,$$

para $i = 1, \ldots, n$. Assim,

$$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \,|\Sigma_{\mathbf{y}} \sim \mathcal{N}_{nk}\left(\mathbf{0}, \Sigma_{\mathbf{y}}\right),$$

onde $\mathcal{N}_{nk}(\cdot)$ representa uma distribuição normal nk-variada com vetor de médias $\mathbf{0}$ (de dimensão $nk \times 1$) e matriz de covariâncias $\Sigma_{\mathbf{y}}$ de dimensão $nk \times nk$, dada por:

$$\Sigma_{\mathbf{y}} = \begin{bmatrix} C_1^2\Sigma_{\mathbf{f}} + \Sigma & \Sigma_{12} & \Sigma_{13} & \dots & \Sigma_{1n} \\ \Sigma_{21} & C_2^2\Sigma_{\mathbf{f}} + \Sigma & \Sigma_{23} & \dots & \Sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \Sigma_{n3} & \dots & C_n^2\Sigma_{\mathbf{f}} + \Sigma \end{bmatrix}$$

em que, $\Sigma_{ii'} = C_i C_{i'} \Sigma_{\mathbf{f}}$ são as matrizes de covariâncias (de dimensão $k \times k$) entre as medidas $\mathbf{Y}_i$ e $\mathbf{Y}_{i'}$, para $i, i' = 1, \dots, n$ e $i \neq i'$.

De maneira similar, temos que,

$$\mathbf{Y}_{n+1} \sim \mathcal{N}_k \left(\mathbf{0}, C_{n+1}^2 \Sigma_{\mathbf{f}} + \Sigma \right),$$

para $C_{n+1} > 0$. Portanto, das propriedades da distribuição normal multivariada, temos que a distribuição conjunta para $(\mathbf{Y}, \mathbf{Y}_{n+1})$ é dada por:

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}_{n+1} \end{bmatrix} | \boldsymbol{\theta}, \mathbb{B}, C_{n+1} \sim \mathcal{N}_{(n+1)k} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{y}} & \mathbb{B}^T \\ \mathbb{B} & C_{n+1}^2 \Sigma_{\mathbf{f}} + \Sigma \end{bmatrix} \right),$$

onde $\mathbb{B} = \begin{bmatrix} \Sigma_{(n+1)1} & \Sigma_{(n+1)2} & \dots & \Sigma_{(n+1)(n-1)} \end{bmatrix}$ é um bloco de matrizes, em que $\Sigma_{(n+1)i}$ são as matrizes de covariâncias entre os valores de $\mathbf{Y}_{n+1}$ e $\mathbf{Y}_i$ dadas por $\Sigma_{(n+1)i} = C_{n+1} C_i \Sigma_{\mathbf{f}}$, para $i = 1, \dots, n$.

Novamente, utilizando as propriedades da distribuição normal multivariada, temos que a distribuição *a posteriori* condicional para $\mathbf{Y}_{n+1}$ é dada por

$$\mathbf{Y}_{n+1} | \mathbf{y}, \mathbf{t}, \boldsymbol{\theta}, \mathbb{B}, C_{n+1} \sim \mathcal{N}_k \left(\mathbb{B}\Sigma_{\mathbf{y}}^{-1}\mathbf{y}, \left(C_{n+1}^2 \Sigma_{\mathbf{f}} + \Sigma \right) - \mathbb{B}^\top \left(C_{n+1}^2 \Sigma_{\mathbf{f}} + \Sigma \right)^{-1} \mathbb{B} \right); \quad (3.10)$$

com C_{n+1} sendo gerado da seguinte distribuição:

$$C_{n+1} | \mathbf{C} \sim \mathcal{N}trunc(0, \overline{C}, S_{\mathbf{c}}^2), \quad (3.11)$$

onde $\overline{C}$ e $S_{\mathbf{c}}^2$ são, respectivamente, a média e a variância amostral dos valores do vetor $\mathbf{C} = (C_1, \dots, C_n)$.

Assim, uma amostra *a posteriori* de $(\boldsymbol{\theta}, \mathbf{Y}_{n+1})$ pode ser gerada de acordo com os "passos" descritos no algoritmo 2.

Após executar o algoritmo 2 para as mesmas L iterações, valor de *burn in* B e "salto" J utilizados no algoritmo 1, uma aproximação para a integral em (3.9) é dada por:

$$\tilde{\pi}(\mathbf{Y}_{n+1} | \mathbf{y}) = \frac{1}{S} \sum_{l=1}^{L} \mathbf{Y}_{n+1}^{(M(l))},$$

onde $M(l)$ é a $(B + 1 + l \cdot J)$-ésima iteração do algoritmo, para $l = 1, \dots, S$.

Algoritmo 2 : Predição

1: Considere uma cadeia de Markov com espaço de estados composto pelo vetor de parâmetros $\boldsymbol{\theta} = (\mathbf{f}(\mathbf{t}), \Sigma, \mathbf{C})$ e por $(C_{n+1}, \mathbf{Y}_{n+1})$;

2: Inicialize o algoritmo com os valores $\boldsymbol{\theta}^{(0)} = \left(\mathbf{f}(\mathbf{t})^{(0)}, \Sigma^{(0)}, \mathbf{C}^{(0)}\right)$ e $C_{n+1}^{(0)}$;

3: **Procedimento:** Para a l-ésima iteração do algoritmo, $l = 1, \ldots, L$:

4: Atualize $\boldsymbol{\theta}$ de acordo com Algoritmo 1;

5: Gere $Y_{n+1}^{(l)}$ da distribuição condicional em (3.10), dado $C_{n+1}^{(l-1)}$ e $\boldsymbol{\theta}^{(l)}$.

6: Gere $C_{n+1}^{(l)}$ da distribuição em (3.11), dado $\boldsymbol{\theta}^{(l)}$.

3.3.1 Estudo de simulação 2

Para ilustrar o desempenho do procedimento de predição, desenvolvemos o estudo de simulação 2. Similar ao estudo de simulação 1, fixamos $n = 4$, $k = 50$ e $f(t)$ como sendo a função função log-Gompertz de parâmetros $\alpha_1 = 12, \alpha_2 = 2$ e $\alpha_3 = 0,1$. O objetivo principal é prever a curva para o $(n+1) = 5^o$ dia.

Para obter as curvas dos quatro primeiros dias, sendo a curva de $f(t)$ a curva média, adotamos o seguinte procedimento:

(i) Seja $C_i < n$ e defina $\mathbf{p} = (p_1, p_2, p_3, p_4)$ com $p_i = \frac{C_i}{n}$, para $i = 1, 2, 3, 4$;

(ii) Gere $\mathbf{p} \sim Dirichlet(\boldsymbol{\alpha})$, onde $Dirichlet(\boldsymbol{\alpha})$ é a distribuição de Dirichlet de parâmetro $\boldsymbol{\alpha}$. Fixamos $\boldsymbol{\alpha} = (50, 50, 50, 50, 50)$; e obtenha $C_i = n \cdot p_i$, para $i = 1, 2, 3, 4$.

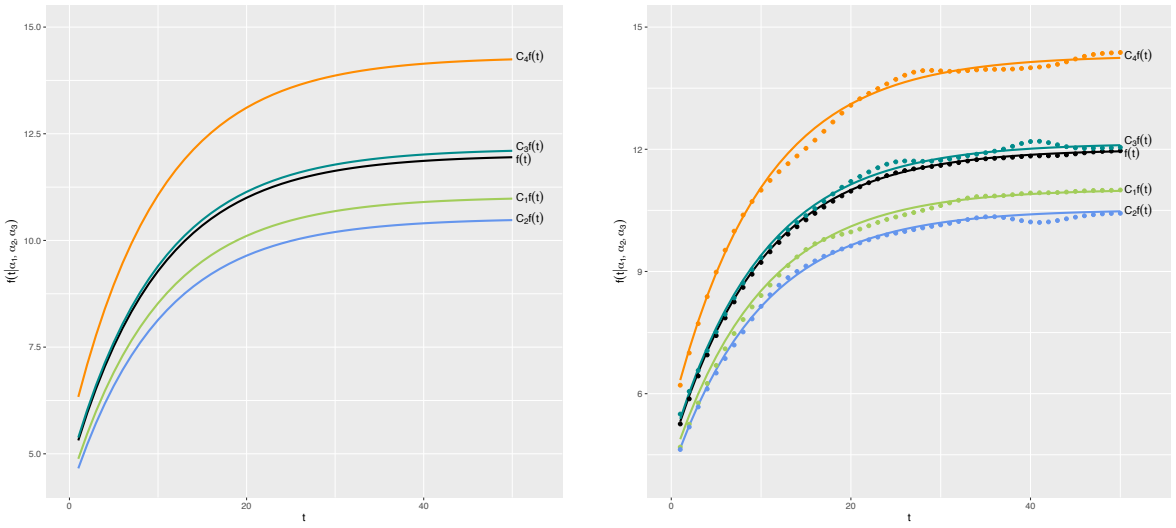
(iii) Gere $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik}) \sim \mathcal{N}_k(C_i\mathbf{f}(\mathbf{t}), \Sigma)$, para $i = 1, \ldots, n+1 = 5$, onde Σ é obtido como descrito no estudo de simulação 1.

Os valores gerados para $\mathbf{C} = (C_1, C_2, C_3, C_4)$ foram $(0,9187; 0,8767; 1,0126; 1,1919)$. A Figura 3.12(a) mostra as curvas reais para os dias 1 a 4, denotadas por $C_i\mathbf{f}(\mathbf{t})$ para $i = 1, 2, 3, 4$, e a Figura 3.12(b) mostra o gráfico da Figura 3.12(a) incluindo os valores gerados para cada dia (símbolos $\bullet$ coloridos), sendo os símbolos $\bullet$ na cor preta os valores médios.

O procedimento utilizado para gerar os dados do $(n+1)$-ésimo dia, é dado pelos seguintes passos:

(i) Gere $C_{n+1} \sim \mathcal{N}trunc\left(0, \bar{C}, \mathbb{S}_c^2\right)$, onde $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$ e $\mathbb{S}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C})^2$.

(ii) Gere $\mathbf{Y}_{n+1} = (Y_{(n+1)1}, \ldots, Y_{(n+1)k}) \sim \mathcal{N}k(C_{n+1}\mathbf{f}(\mathbf{t}), \Sigma)$.



(a) Curvas

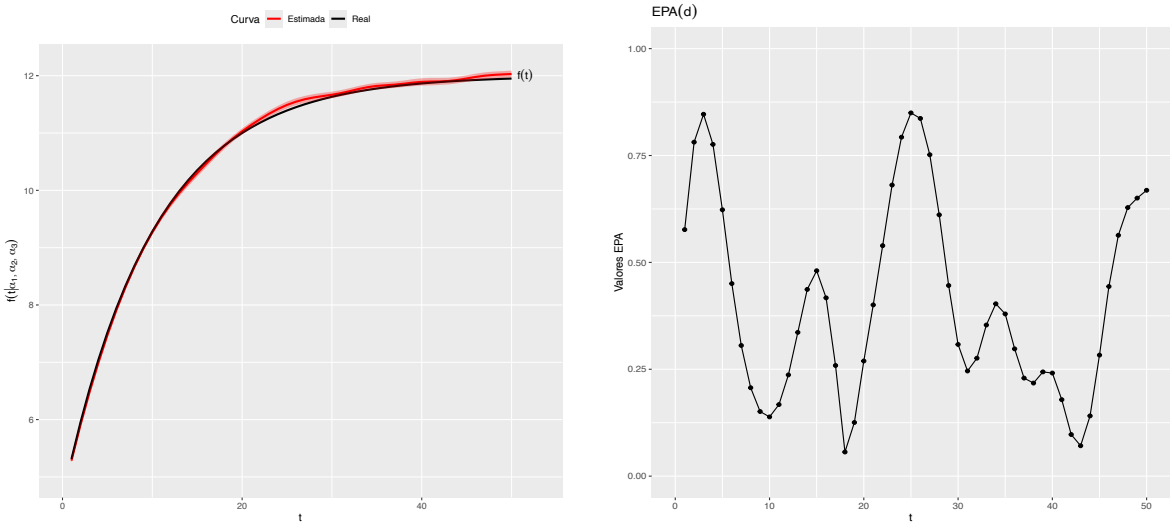
(b) Curvas e valores gerados

Figura 3.12: Curvas e valores gerados.

A partir dos valores gerados de $\mathbf{C} = (0,9187; 0,8767; 1,0126; 1,1919)$, o valor gerado para C_{n+1} foi $0,9823$. Gerado os dados, utilizamos os valores gerados de $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ para desenvolver o procedimento de predição (Algoritmo 2) e obter as estimativas para $\hat{\mathbf{Y}}_{n+1}$. Para isto, executamos o Algoritmo 2 para as mesmas $L = 55.000$ iterações, *burn-in* $B = 5.000$ e "saltos" $J = 10$ do Algoritmo 1. A curva estimada para o $(n+1)$-ésimo dia é obtida construíndo o grafico do pontos $(t, \hat{y}_{(n+1)t})$ ligados por linhas, em que, $\hat{y}_{(n+1)t}$ é o valor previsto de $Y_{(n+1)t}$, para $t = 1, \ldots, k$.

A Figura 3.13(a) mostra o gráfico da curva de $f(t)$ e a curva estimada pelo método proposto. A Figura 3.13(b) mostra o gráfico dos valores de $EPA(\mathbf{d})$. De forma similar do estudo de simulação 1, os resultados mostram um desempenho muito satisfatório do método proposto; com a curva estimada sendo satisfatoriamente próxima da curva real de $f(t)$, uma vez que os valores de $EPA(\mathbf{d})$ foram todos inferiores a 1; significando que os valores dos erros percentuais absolutos entre os valores reais e o valores estimados são todos inferiores a 1%. A média dos valores $EPA(\mathbf{d})$ é $0,4094$, *i.e.*, em média o erro percentual absoluto entre os valores reais e os valores estimados é inferior a $0,5\%$.

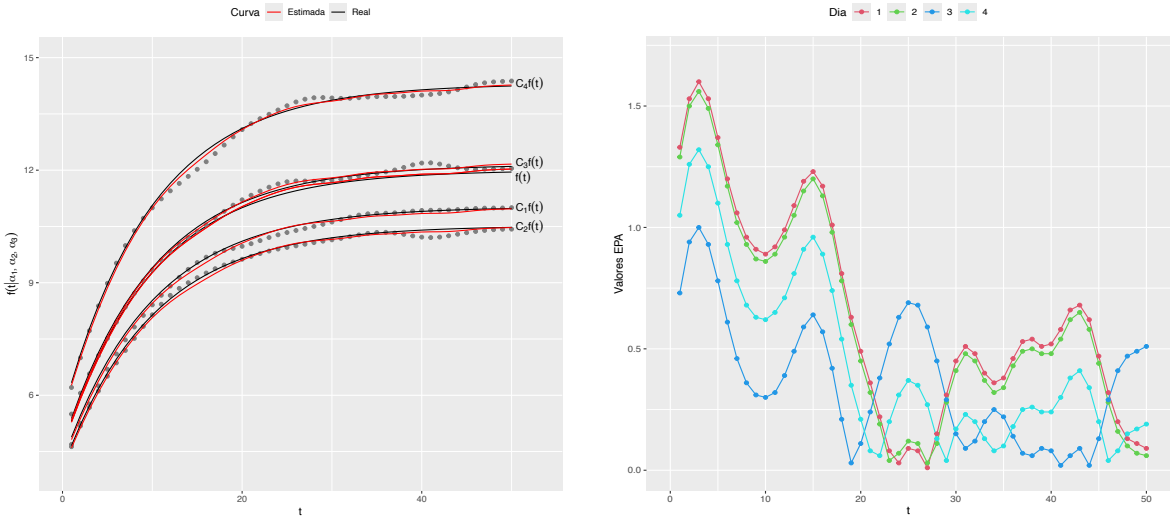
A Figura 3.14(a) mostra o gráfico da curva do dia i (linha preta), denotada por $C_i f(t)$, em que, os símbolos $\bullet$ são os valores gerados o dia i, incluindo as curvas estimadas pelo método proposto (linhas em vermelho), para $i = 1, 2, 3, 4$. A Figura 3.14(b), mostra o gráfico dos valores de $EPA(\mathbf{d}_i)$ e a Tabela 3.6 mostra as medidas resumo dos valores $EPA(\mathbf{d}_i)$, para $i = 1, \ldots, n$.



(a) Curva real e estimada

(b) Valores $EPA(\mathbf{d})$

Figura 3.13: Curva real e estimada e valores $EPA(\mathbf{d})$.



(a) Curva real e estimada

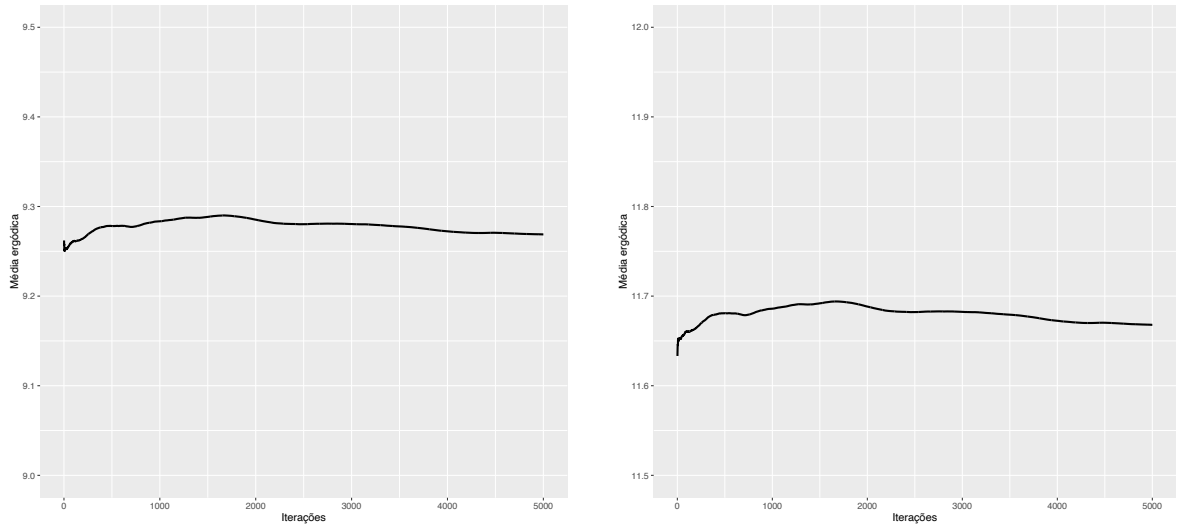
(b) Valores $EPA(\mathbf{d})$

Figura 3.14: Curva real e estimada e valores $EPA(\mathbf{d}_i)$.

Tabela 3.6: Medidas resumo dos valores $EPA(\mathbf{d}_i)$, para $i = 1, 2, 3, 4$.

Medida	Mínimo	1^o quartil	Mediana	Média	3^o quartil	Máximo
$EPA(\mathbf{d}_1)$	0,0100	0,3300	0,5250	0,6448	0,9825	1,6000
$EPA(\mathbf{d}_2)$	0,0300	0,2900	0,4850	0,6148	0,9525	1,5600
$EPA(\mathbf{d}_3)$	0,0200	0,1325	0,3400	0,3724	0,5775	1,0000
$EPA(\mathbf{d}_4)$	0,0400	0,1825	0,3050	0,4508	0,7025	1,3200

Note que, as curvas estimadas para cada um dos quatro dias estão satisfatoriamente próximas das curvas reais, uma vez que os valores de $EPA(\mathbf{d}_i)$ são próximos de zero, para $i = 1, 2, 3, 4$. Também verificamos a convergência dos valores gerados pelo algoritmo *Gibbs sampling*. Similar aos resultados apresentados no estudo de simulação 1, não há razões para duvidar da convergência dos valores gerados, uma vez que os valores da ME apresenta estabilização satisfatória, conforme ilustrado na Figura 3.15.



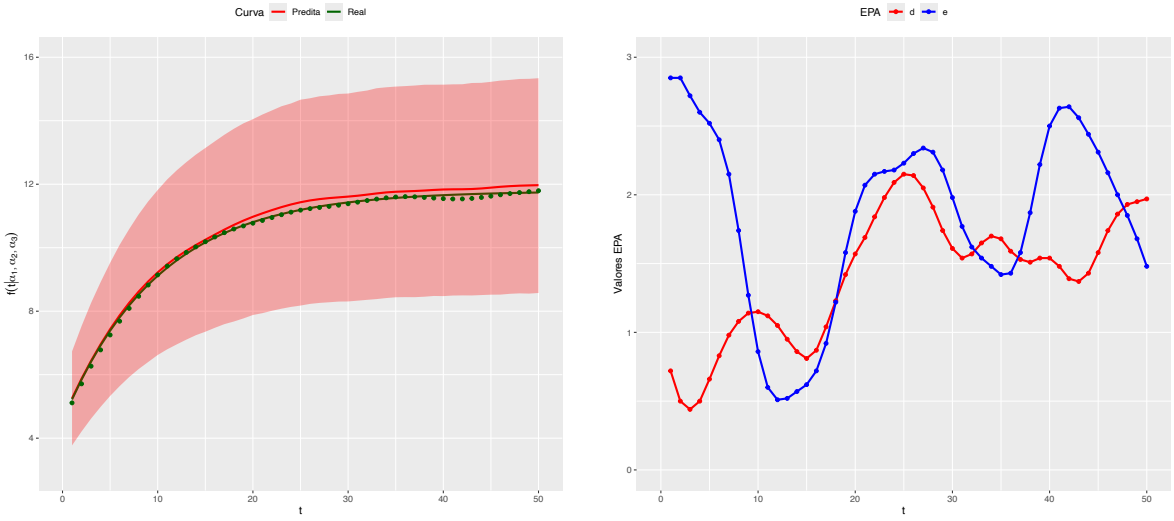
(a) $f(10)$ (b) $f(30)$

Figura 3.15: Média ergódica dos valores gerados para $f(10)$ e $f(30)$.

Com relação a curva gerada para o $(n+1)$-ésimo dia, a estimativa para C_{n+1} é $\hat{C}_{n+1} = 0,9936$ com um intervalo de credibilidade de 95% dado por $(0,7207; 1,2716)$. Ou seja, o valor gerado $C_{n+1} = 0,9823$ pertence ao intervalo de credibilidade.

A Figura 3.16(a) mostra o gráfico da curva gerada para o $(n+1)$-ésimo dia (linha na cor verde), em que, os símbolos $\bullet$ na cor verde são os dados gerados, incluindo a curva pedita (linha na cor vermelha) e uma banda de predição *a posteriori* de 95% (região na cor vermelha), obtida pelo método proposto. Além disso, note que, a curva real está toda dentro da banda de pedição *a posteriori* (95%) obtida pelo método proposto.

A Figura 3.16(b) mostra o gráfico dos valores $EPA(\mathbf{d})$ e $EPA(\mathbf{e})$. Os valores $EPA(\mathbf{d})$ variaram de um valor mínimo de $0,440$ a um valor máximo de $2,150$ e com valor médio de $1,413$. Já os valores $EPA(\mathbf{e})$ variaram de um valor mínimo de $0,510$ a um valor máximo de $2,850$ e com valor médio de $1,844$. Ou seja, em média, o erro percentual absoluto entre o valor predito e o valor real é de $1,413$ e, em média, o erro percentual absoluto entre o valor predito e o valor gerado é de $1,844$.



(a) Curva real e estimada

(b) Valores $EPA(\mathbf{d})$

Figura 3.16: Curva real e estimada e valores $EPA(\mathbf{d}_i)$.

Além dos valores EPA, também calculamos como medida de performance das predições a raiz quadrada do erro quadrático médio (REQM), dada por

$$RMSE = \sqrt{\frac{1}{k}\sum_{t=1}^{k}\left(y_{(n+1)t} - \hat{y}_{(n+1)t}\right)^2},$$

onde $y_{(n+1)t}$ é o valor gerado para o t-ésimo instante de tempo do dia $(n+1)$, e $\hat{y}_{(n+1)t}$ é o respectivo valor predito, para $t = 1, \ldots, k$. O valor REQM obtido foi $0,2021$. Ou seja, de maneira similar aos valores EPA, o valor REQM também indica que os valores preditos estão satisfatoriamete próximos aos valores gerados.

Para não ficarmos limitados aos resultados de apenas um conjunto de dados artificiais, repetimos a simulação 2, $M = 100$ vezes, e calculamos a média dos valores $EPA(\mathbf{d})$ e a porcentagem de vezes que a curva real para o $(n+1)$-ésimo dia está toda dentro da banda de predição de 95% e a média dos valores EPA e REQM. A Figura 3.17, mostra o gráfico das médias dos valores $EPA(\mathbf{d})$ e os valore REQM para as M simulações. Note que, ambos os resultados mostram um desempenho muito satisfatório do método proposto. Sendo todos os valores $MEPA$ inferiores a 5; significando que a média dos valores EPA entre os valores da curva do $(n+1)$-ésimo dia e o valor predito pelo método proposto nas M simulações são todos inferiores a 5%; e os valores REQM menores do que $0,5$. Além disso, no geral, em 96% dos casos simulados, a curva real está toda dentro da banda de predição. Como ilustração dos resultado de predição, a Figura 3.18 mostra o gráfico da

curva prevista com uma banda de predicão *a posteriori* de 95% e a curva real para a 18^a simulação e a 27^a simulação. Estas duas simulações, são os casos com menor (18^a) e maior (27^a) média de $EPA(\mathbf{d})$, dadas respectivamente por: $0,6115$ e $4,0017$. Os valores REQM são: $0,0266$ e $0,4996$.

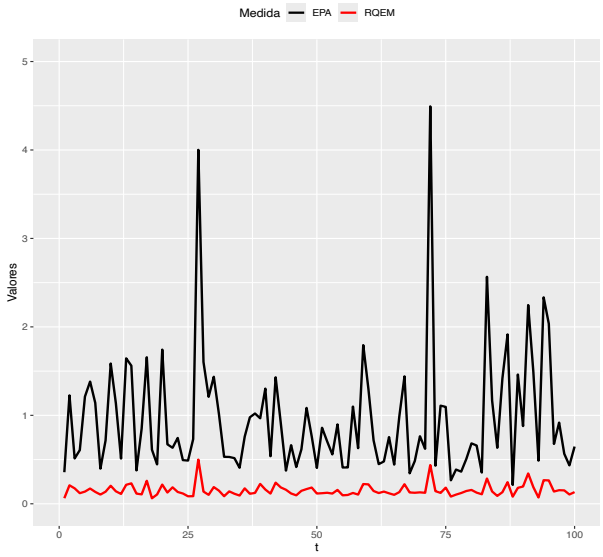
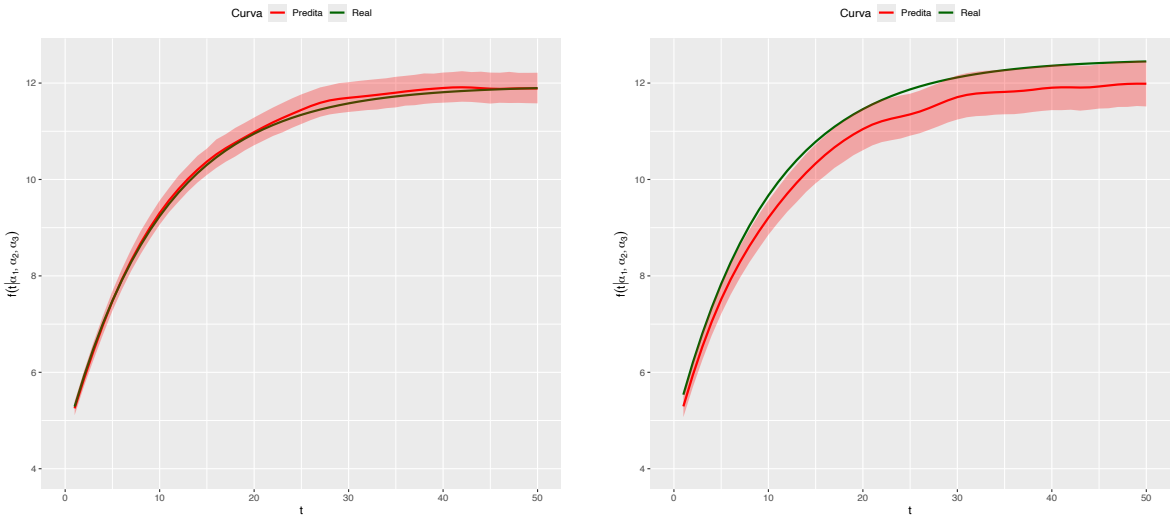


Figura 3.17: Média dos valores $EPA(\mathbf{d})$ e $RQEM$ para as $M = 100$ simulações.



(a) Simulação 88 (b) Simulação 96

Figura 3.18: Curva real e predita para as simulações 88 e 96.

Ou seja, os resultados dos estudos de simulação 1 e 2 mostram que o método proposto é uma alternativa eficiente para se estimar as curvas de crescimento da geração de energia solar para cada um dos dias em estudo e também para predizer a curva de crescimento para o dia $(n+1)$.

Capítulo 4

Aplicação

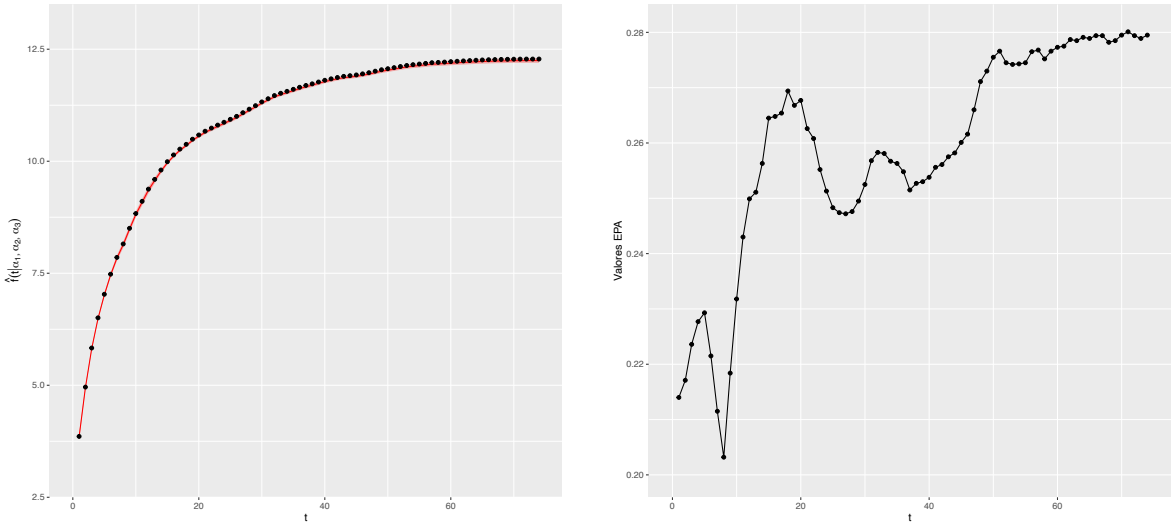
Neste Capítulo, apresentamos os resultados obtidos ao aplicar o método proposto ao conjunto de dados reais descrito na Seção 3.1 do Capítulo 3. Para isto, fixamos os mesmos valores dos hiper-parâmetros para o modelo hierárquico em (3.3) e os mesmos valores de L, B e J utilizados nos estudos de simulação 1 e 2 do Capítulo 3. Além disso, também consideramos $n = 4$, *i.e.*, aplicamos o método para estimar a curva de $f(t)$ utilizando o conjunto de dados de quatro dias, e em seguida predizemos a curva para o dia $(n + 1)$.

Este procedimento foi aplicado para todo o conjunto de dados, sempre utilizando a janela de quatro dias. Ou seja, iniciamos, considerando os dados dos dias 1 a 4 e predizemos a curva do quinto dia. Em seguida, consideramos os dados dos dias 2 a 5 e predizemos a curva para o sexto dia; e procedemos com este processo até a última análise que foi feita utilizando os dados dos dias 15 a 18 e predizemos a curva do dia 19. No geral, foram feitas quinza análises e a predição das curvas para os dias 5 a 19.

4.1 Resultados

A nossa primeira aplicação considera os dados registrados nos primeiros quatro dias de realização do experimento para estimar a curva de $f(t)$, e em seguida predizemos a curva para o quinto dia $(n+1 = 5)$. A Figura 4.1(a), mostra os valores médios registrados nos primeiros quatro dias (símbolos $\bullet$), e a curva estimada de $f(t)$ (linha vermelha) com uma banda de credibilidade de 95% (região na cor vermelha). A Figura 4.1(b) mostra o gráfico dos valores EPA entre $\bar{y} = (\overline{y}_1, \ldots, \overline{y}_k)$ e os valores estimados $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_k)$.

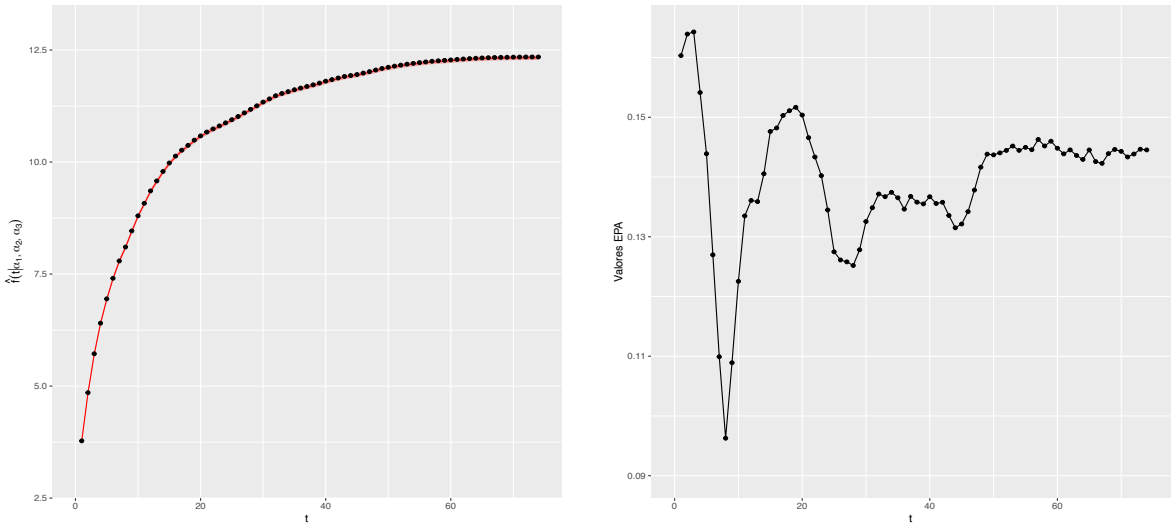
A média dos valores EPA (MEPA) é $0,2590$. Este resultado mostra que os valores estimados estão muito próximos dos valores médios registrados. Em outras palavras, a abordagem proposta apresentou um desempenho muito satisfatório na estimativa dos valores médios registrados nos primeiros quatro dias em que o experimento foi realizado.



(a) Curva estimada (b) Valores EPA

Figura 4.1: Valores médios, curva estimada e valores EPA, análise 1.

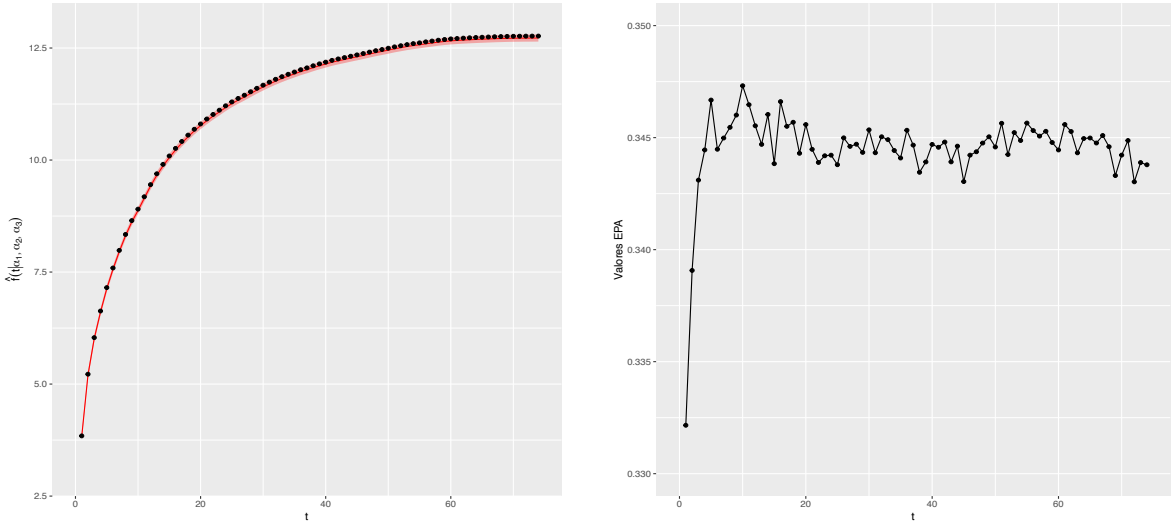
A Figura 4.2, mostra os resultados da 2^a análise, *i.e.*, estimação da curva de $f(t)$ utlizando os dados registrados nos dias 2 a 5. A média dos valores EPA é $0,1395$.



(a) Curva estimada (b) Valores EPA

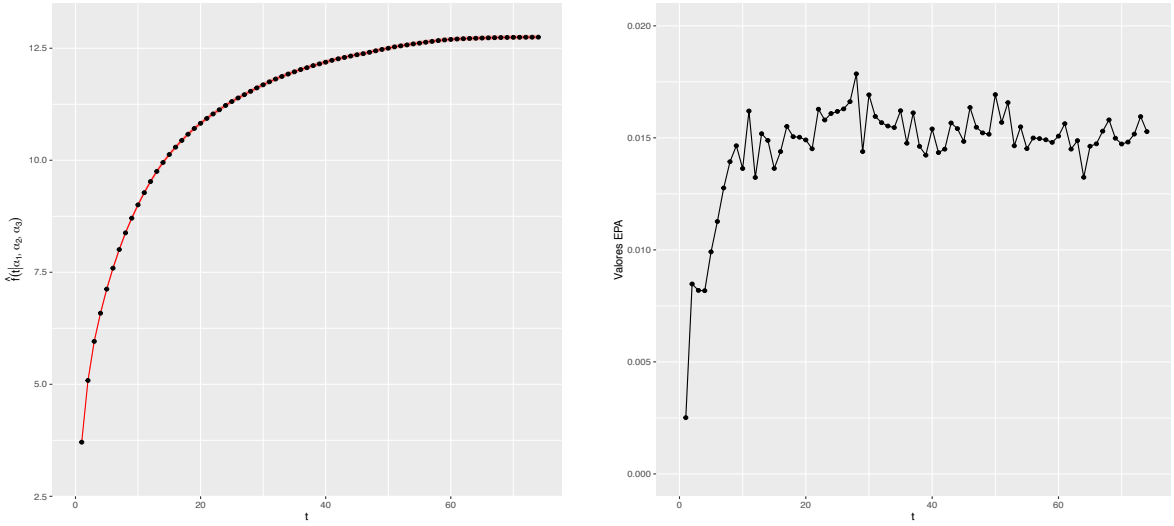
Figura 4.2: Valores médios, curva estimada e valores EPA, análise 2.

As Figuras 4.3 e 4.4, ilustram a performance do método para as duas últimas análises (análises 14 e 15). Na análise 14 foi feita a estimação da curva de $f(t)$ utlizando os dados registrados nos dias 14 a 17; e na análise 15 foi feita a estimação da curva de $f(t)$ utlizando os dados registrados nos dias 15 a 18. Para estas duas análises a média dos valores EPA são: $0,3445$ e $0,0146$.



(a) Curva estimada (b) Valores EPA

Figura 4.3: Valores médios, curva estimada e valores EPA, análise 14.



(a) Curva estimada (b) Valores EPA

Figura 4.4: Valores médios, curva estimada e valores EPA, análise 15.

A Tabela 4.1 apresenta a média dos valores EPA (denotado por MEPA) e os valores REQM para as quinze análises. Note que, tanto os valores de MEPA quanto os valores REQM são próximos de zero; o que indica que os valores estimados $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_k)$ são próximos dos valores médios regitrados, $\bar{\mathbf{y}} = (\overline{y}_1, \ldots, \overline{y}_k)$, para as quinze análises.

Tabela 4.1: Valores MEPA e REQM para as análises 1 a 15.

Análise	$MEPA$	$REQM$	Análise	$MEPA$	$REQM$	Análise	$MEPA$	$REQM$
1	0,2590	0,0292	6	0,2038	0,0228	11	0,5360	0,0612
2	0,1395	0,0155	7	0,3670	0,0393	12	0,1815	0,0208
3	0,0500	0,0061	8	0,2378	0,0273	13	0,0095	0,0011
4	0,0753	0,0085	9	0,2157	0,0243	14	0,3445	0,0394
5	0,3996	0,0450	10	0,4903	0,0542	15	0,0146	0,0017

Ou seja, de maneira similar ao estudo de simulação 1 do Capítulo 3, o método proposto se mostrou muito satisfatório para estimação da curva de crescimento dos valores médios, haja vista que a média dos valores EPA das quinze análises são todas inferiores a $0,55$; significando que, e média, o erro percentual absoluto entre os valores médios registrados e os valores estimados pelo método proposto é inferior a $0,55\%$.

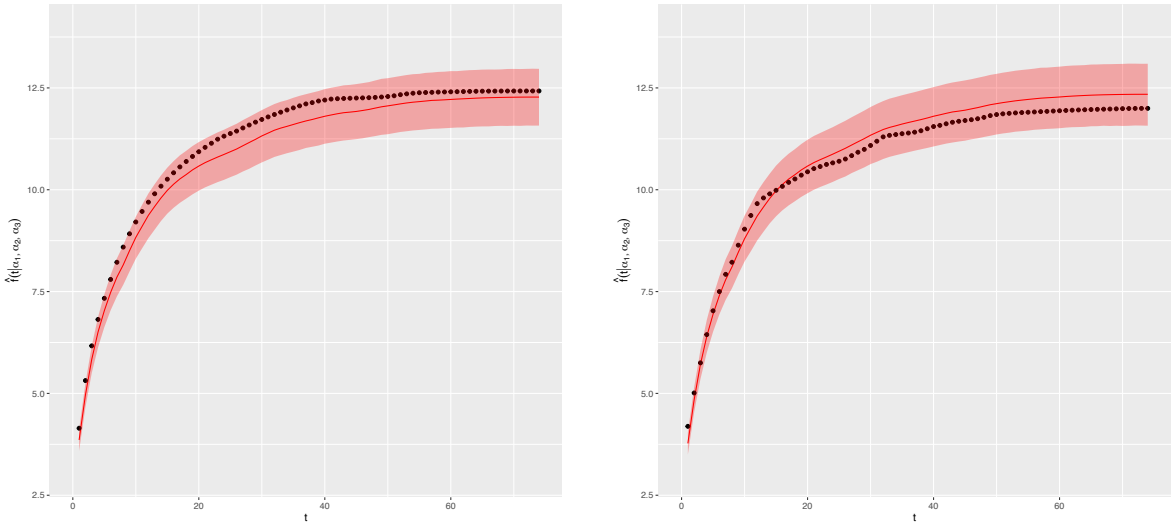
4.2 Predição

Nesta seção, apresentamos os resultados com relação a predição das curvas dos valores gerados para os dias 5 a 19. Os gráficos da Figure 4.5 mostram as curvas preditas para os dias 5 e 6 (linhas em vermelho), obtidas nas análises 1 e 2, respectivamente; em que, os símbolos $\bullet$ são os valores registrados nestes dois dias e a região destacada na cor vermelha é uma banda de credibilidade de 95% obtida pelo método proposto. As médias dos valores EPA para estes dois dias são: $2,8659$ e $2,3021$, respectivamente. E os valores REQM são: $0,3170$ e $0,2628$, repectivamente.

Similarmente, os gráficos da Figura 4.6 mostram as curvas preditas para os dias 18 e 19, obtidas nas análises 14 e 15, respectivamente. As médias dos valores EPA para estes dois dias são: $0,7374$ e $0,3344$, respectivamente. E os valores REQM são: $0,1121$ e $0,0382$, repectivamente. Ou seja, este valores mostram que os valores preditos estão satisfatóriamante próximos aos valores registrados.

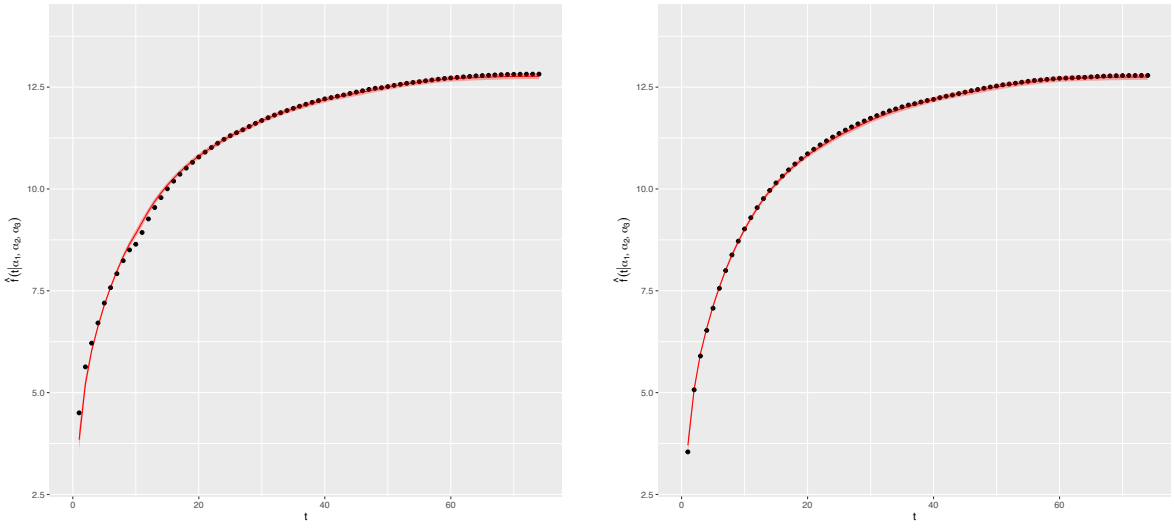
A Tabela 4.2 mostra os valores $MEPA$ para a cuva predita nas quinze análises. Note que, os valores MAPE variaram de um mínimo de $0,3344$ para o dia 19 a um máximo de

7, 1648 para o dia 13, com um valor médio de 2, 5719. E os valores REQM variaram de um mínimo de 0, 0382 para o dia 19 a um máximo de 0, 7410 para o dia 13, com um valor médio de 0, 2895. Ou seja, os valores preditos estão próximos dos valores registrados em todas as 15 análises, uma vez que, os valores MEPA e REQM são próximos a zero.



(a) Dia 5 (b) Dia 6

Figura 4.5: Valores registrados nos dias 5 e 6 e curvas preditas.



(a) Dia 18 (b) Dia 19

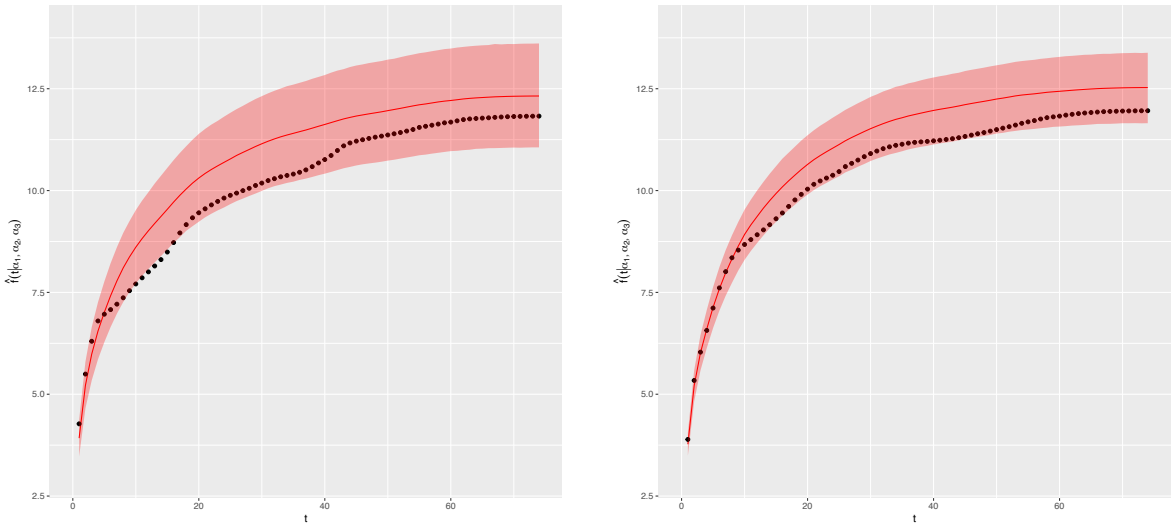
Figura 4.6: Valores registrados nos dias 18 e 19 e curvas preditas.

Os gráficos da Figura 4.7 mostram as curvas preditas para os dias 13 e 14 (análises 9 e 10), que são os dois dias com os maiores de médias de valores EPA. Embora as predições

para estes dois dias apresentem os dois maiores valores de médias de EPA, note quem a maioria dos valores registrados está dentro da banda de credibilidade de predição de 95%.

Tabela 4.2: Valores MEPA e REQM para as predições dos dias 5 a 19.

Dia	*MAPE*	*REQM*	*Dia*	*MAPE*	*REQM*	*Dia*	*MAPE*	*REQM*
5	2,8659	0,3180	10	2,8855	0,3323	15	0,6193	0,0732
6	2,3021	0,2628	11	5,1500	0,5780	16	0,5880	0,0677
7	5,1190	0,5321	12	3,6303	0,4415	17	0,5753	0,0786
8	1,0693	0,1222	13	7,1648	0,7410	18	0,7374	0,1128
9	0,3507	0,0409	14	5,1875	0,6027	19	0,3344	0,0382



(a) Dia 13 (b) Dia 14

Figura 4.7: Valores registrados nos dias 13 e 14 e curvas preditas.

Uma explicação que consideramos plausível para os resultados das análises 9 e 10 é a seguinte: A modelagem proposta considera apenas os valores registrados ao longo do tempo; porém, a energia gerada é influenciada por variáveis ambientais, tais como temperatura e irradiância. Assim, pode ter ocorrido alguma mudança significativa nestas variáveis ambientais, o que fez que os valores registrado fossem diferentes do esperado em relação as medidas registradas nos dias anteriores.

Contudo, mesmo sem levar em consideração a inclusão de covariáveis na modelagem, a abordagem proposta apresentou um desempenho muito satisfatório, pois as médias dos valores EPA são próximas a 0, indicando que os valores preditos são próximos dos valores registrados. Uma abordagem que considere a presença de covariáveis pode ser vista como uma extensão da abordagem proposta e será desenvolvida em trabalhos futuros.

Capítulo 5

Considerações Finais

Nesta dissertação, desenvolvemos um modelo Bayesiano hierárquico para a modelagem e predição da produção de energia solar fotovoltaica ao longo do tempo. Para isso, assumimos que a curva de crescimento da energia gerada ao longo do tempo de um dia é proporcional a uma curva média cuja função associada é denotada por $f(t)$, para $t > 0$.

No entanto, ao invés de assumir uma abordagem paramétrica fixando $f(t)$ como sendo uma função matemática conhecida indexada por parâmetros, assumimos que $f(t)$ é uma função desconhecida, mas com o vetor de valores $\mathbf{f}(\mathbf{t}) = (f(1), \ldots, f(k))$ sendo gerado *a priori* de um processo Gaussiano. Completamos o modelo assumindo uma distribuição *a priori* inversa-Wishart para a matriz de covariâncias Σ dos erros aleatórios e uma distribuição normal truncada a esquerda de zero para as constantes de proporcionalidade C_i, $i = 1, \ldots, n$. Além disso, fixamos os valores para os hiperparâmetros de maneira a obter distribuições *a priori* pouco informativas.

De acordo com a teoria da inferência Bayesiana, definido o modelo, todas as inferências são feitas com base na distribuição *a posteriori* conjunta para os parâmetros de interesse, que no caso do modelo proposto é dado por $\boldsymbol{\theta} = (\mathbf{f}(\mathbf{t}), \Sigma, \mathbf{C})$. Esta distribuição *a posteriori* conjunta é obtida atualizando as distribuições *a priori* pela função de verossimilhança, que para o modelo proposto é dada pela função de verossimilhança de uma distribuição normal k-variada dado uma amostra de tamanho n, para $n \geq 1$. Contudo, como a distribuição *a posteriori* conjunta não possui forma matemática conhecida que nos permita obter os estimadores para os parâmetros de maneira analítica, nossas inferências foram feitas de maneira numérica utilizando o algoritmo *Gibbs sampling*.

Como vantagens da abordagem proposta, destacamos os seguintes pontos:

(i) O modelo Bayesiano hierárquico proposto é muito flexível e adapta-se a quantidade de valores registrados nos dias;

(ii) O procedimento de inferência é baseado na implementação de um algoritmo *Gibbs sampling*, que pode ser facilmente implementado em *softwares* estatísticos, tal como o *software* `R`.

(iii) A curva de crescimento prevista para o dia $(n + 1)$ é obtida utilizando apenas os dados históricos; e

(iv) Não é necessário ajustar um conjunto de modelos e depois compará-los usando algum critério de seleção de modelos.

O bom desempenho da abordagem proposta juntamente com as quatro vantagens descritas acima, foram ilustradas através de dois estudos de simulação e de uma aplicação a dados reais. Os resultados obtidos mostram que a abordagem proposta é uma alternativa eficiente para modelar a energia solar gerada nos dias considerados no estudo e também para prever a energia solar que será gerada no dia seguinte.

Do ponto de vista prático, os resultados mostraram que a modelagem proposta e o processo de estimação dos parâmetros foram satisfatoriamente precisos na predição da geração de energia para o dia seguinte; apresentando valor médio de erro percentual absoluto entre os valores registrados e o valores preditos inferiores a 10%.

Embora a abordagem proposta tenha sido descrita para prever a curva de crescimento para o dia seguinte, ela também pode ser usada para prever a produção de energia em intervalos mais curtos, tal como, a produção horária de energia e a produção que será gerada na próxima hora. Uma extensão da abordagem proposta é a inclusão de variáveis explicativas na modelagem, uma vez que a energia produzida é influenciada por variáveis ambientais, tais como a temperatura e a irradiância. Todas as implementações computacionais foram feitas utilizando o *software* `R` e os códigos podem ser obtidos através de contato por email a autora desta dissertação.

Apêndice

Neste apêndice apresentamos detalhes sobre a obtenção de expressões matemática citadas no texto e alguns códigos em R implementados.

Apêndice 1: F.d.p. conjunta de $\mathbf{X} = (X_1, X_2)$

Considere, a seguinte transformação

$$X_1 = \sigma_1 Z_1 + \mu_1 \quad \text{e} \quad X_2 = \sigma_2 \left[\rho Z_1 + \left(1-\rho^2\right)^{1/2} Z_2\right] + \mu_2$$

onde μ_1, μ_2, σ_1, σ_2 são constantes, tais que, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in \mathbb{R}^+$ e $-1 < \rho < 1$.

As derivadas parciais em relação a Z_1 e Z_2 são dada por

$$\frac{dX_1}{dZ_1} = \sigma_1; \quad \frac{dX_1}{dZ_2} = 0; \quad \frac{dX_2}{dZ_1} = \sigma_2\rho \quad \text{e} \quad \frac{dX_2}{dZ_2} = \sigma_1\left(1-\rho^2\right)^{1/2}.$$

O Jacobiano da transformação é dado por

$$J = \begin{vmatrix} \frac{dX_1}{dZ_1} & \frac{dX_1}{dZ_2} \\ \frac{dX_2}{dZ_1} & \frac{dX_2}{dZ_2} \end{vmatrix} = \begin{vmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\left(1-\rho^2\right)^{1/2} \end{vmatrix} = \left(1-\rho^2\right)^{1/2}\sigma_1\sigma_2.$$

Além disso, temos que

$$Z_1 = \frac{X_1-\mu_1}{\sigma_1} \quad \text{e} \quad Z_2 = \frac{X_2-\mu_2}{\sigma_2\left(1-\rho^2\right)^{1/2}} - \frac{\rho\left(X_1-\mu_1\right)}{\sigma_1\left(1-\rho^2\right)^{1/2}}.$$

Logo,

$$Z_1^2 = \frac{\left(X_1-\mu_1\right)^2}{\sigma_1^2} \quad \text{e} \quad Z_2^2 = \frac{\left(X_2-\mu_2\right)^2}{\sigma_2^2\left(1-\rho^2\right)} - \frac{2\rho\left(X_1-\mu_1\right)\left(X_2-\mu_2\right)}{\sigma_1\sigma_2\left(1-\rho^2\right)} + \frac{\rho^2\left(X_1-\mu_1\right)^2}{\sigma_1^2\left(1-\rho^2\right)};$$

e

$$Z_1^2 + Z_2^2 = \frac{\left(X_1-\mu_1\right)^2}{\sigma_1^2\left(1-\rho^2\right)} + \frac{\left(X_2-\mu_2\right)^2}{\sigma_2^2\left(1-\rho^2\right)} - \frac{2\rho\left(X_1-\mu_1\right)\left(X_2-\mu_2\right)}{\sigma_1\sigma_2\left(1-\rho^2\right)}.$$

Portanto a função densidade de probabilidade conjunta de $\mathbf{X} = (X_1, X_2)$ é dada por

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= f_{\mathbf{Z}}(\mathbf{x}) \left|J^{-1}\right| \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right.\right. \\ &\quad \left.\left. -2\rho\frac{(X_1-\mu_1)(X_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}. \end{aligned}$$

para $\mathbf{x} = (x_1, x_2)$, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

Apêndice 2: Matriz Σ^{-1}

Considere Σ uma matriz de dimensão 2×2 dada por

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

onde σ_1^2 é a variância da variavel aleatória X_1, σ_2^2 é a variância da variável aleatória X_2 e $\sigma_{12} = Cov(X_1, X_2) = \rho\sigma_1\sigma_2$ é a covariância de (X_1, X_2).

O determinante de Σ é

$$\Sigma = \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2\left(1-\rho^2\right).$$

Portanto, a matriz Σ^{-1} é dada por

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{|\Sigma|}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \\ &= \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \\ &= \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_1^2} \end{pmatrix} \end{aligned}$$

Apêndice 3: Distribuição condicional para $\mathbf{f(t)}$

Multiplicando a função de verossimilhança pela função densidade de probabilidades *a priori* para $\mathbf{f(t)}$, obtemos:

$$\begin{aligned}
\pi(\mathbf{f(t)}|\bullet) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f(t)}\right)^T\Sigma^{-1}\left(\mathbf{y}_i - c_i\mathbf{f(t)}\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left(\mathbf{f(t)}-\mathbf{m}\right)^T\Sigma_f^{-1}\left(\mathbf{f(t)}-\mathbf{m}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{y}_i^T\Sigma^{-1}\mathbf{y}_i - \mathbf{y}_i^T\Sigma^{-1}c_i\mathbf{f(t)} - c_i\mathbf{f(t)}^T\Sigma^{-1}\mathbf{y}_i + c_i^2\mathbf{f(t)}^T\Sigma^{-1}\mathbf{f(t)}\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\mathbf{f(t)}^T\Sigma_f^{-1}\mathbf{f(t)} - \mathbf{f(t)}^T\Sigma_f^{-1}\mathbf{m} - \mathbf{m}^T\Sigma_f^{-1}\mathbf{f(t)} + \mathbf{m}^T\Sigma_f^{-1}\mathbf{m}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathbf{f(t)}^T\Sigma_f^{-1}\mathbf{f(t)} + \sum_{i=1}^{n}c_i^2\mathbf{f(t)}^T\Sigma^{-1}\mathbf{f(t)} - 2\sum_{i=1}^{n}c_i\mathbf{y}_i^T\Sigma^{-1}\mathbf{f(t)}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathbf{f(t)}^T\Sigma_f^{-1}\mathbf{f(t)} + \mathbf{f(t)}^T c^*\Sigma^{-1}\mathbf{f(t)} - 2\mathbf{f(t)}\Sigma^{-1}\bar{yc}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathbf{f(t)}^T\left(c^*\Sigma^{-1}+\Sigma_f^{-1}\right)\mathbf{f(t)} - 2\mathbf{f(t)}n\Sigma^{-1}\bar{yc}\right]\right\}; \qquad (5.1)
\end{aligned}$$

sendo $c^* = \sum_{i=1}^{n} c_i^2$ e $\bar{yc} = \frac{1}{n}\sum_{i=1}^{n} c_i\mathbf{y}_i$.

Para, prosseguirmos com o desenvolvimento, precisamos relembrar o seguinte resultado.

Resultado: Sejam $\mathbf{x}$ e $\mathbf{b}$ vetores de dimensão k e n uma matriz kxk. Então, $\mathbf{x}^T n\mathbf{x} - 2\mathbf{b}^T\mathbf{x} = (\mathbf{x}-n^{-1}\mathbf{b})^T n(\mathbf{x}-n^{-1}\mathbf{b}) - \mathbf{b}^T n^{-1}\mathbf{b}$. Este resultado pode ser verificado multiplicando a forma quadrática

$$\begin{aligned}
(\mathbf{x}-n^{-1}\mathbf{b})^T n(\mathbf{x}-n^{-1}\mathbf{b}) &= (\mathbf{x}^T n - nn^{-1}\mathbf{b}^T)(\mathbf{x}-n^{-1}\mathbf{b}) \\
&= \mathbf{x}^T n\mathbf{x} - \mathbf{x}^T nn^{-1}\mathbf{b} - \mathbf{b}^T\mathbf{x} + \mathbf{b}^T n^{-1}\mathbf{b} \\
&= \mathbf{x}^T n\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + \mathbf{b}^T n^{-1}\mathbf{b}.
\end{aligned}$$

Aplicando o resultado em (5.1) para $n = c^*\Sigma^{-1}$ e $\mathbf{b} = n\Sigma^{-1}\bar{yc}$, temos que:

$$\begin{aligned}
\pi(\mathbf{f(t)}|\bullet) &\propto \exp\left\{-\frac{1}{2}\left[\mathbf{f(t)}^T n\mathbf{f(t)} - 2\mathbf{f(t)}n\Sigma^{-1}\bar{yc}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\mathbf{f(t)}-n^{-1}\mathbf{b}\right)^T n\left(\mathbf{f(t)}-n^{-1}\mathbf{b}\right) - \mathbf{b}^T n^{-1}\mathbf{b}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\mathbf{f(t)}-n^{-1}\mathbf{b}\right)^T n\left(\mathbf{f(t)}-n^{-1}\mathbf{b}\right)\right]\right\}.
\end{aligned}$$

Ou seja, temos o núcleo de uma distribuição normal multivariada de parâmetros:

$$\begin{aligned}\mu_{post} &= n^{-1}\mathbf{b} = n\Sigma^{-1}\left(c^{*}\Sigma^{-1}+\Sigma_{f}^{-1}\right)^{-1}\bar{y}c \\ \Sigma_{post} &= n^{-1} = \left(c^{*}\Sigma^{-1}+\Sigma_{f}^{-1}\right)^{-1}\end{aligned}$$

Portanto,

$$\mathbf{f}(\mathbf{t})|\bullet \sim \mathcal{N}_k\left(\mu_{post};\Sigma_{post}\right)$$

Apêndice 4: Distribuição condicional para Σ

Multiplicando a função de verossimilhança pela função densidade de probabilidades *a priori* para Σ, obtemos:

$$\pi\left(\Sigma|\bullet\right) \propto |\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^{T}\Sigma^{-1}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\right\}|\Sigma|^{-\frac{\delta+k+1}{2}}\exp\left\{-tr\left(\Sigma^{-1}\mathbb{V}\right)\right\}$$

Para, prosseguirmos com o desenvolvimento, precisamos relembrar o seguinte resultado.

Resultado: Seja $\mathbf{x}$ um vetor de dimensão k e A uma matriz kxk. Temos que:

$$\sum_{i=1}^{n}\mathbf{x}_i^{T}A\mathbf{x}_i = tr\left(A\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{T}\right) \qquad \text{e} \qquad tr(A+B) = tr(A)+tr(B)$$

Assim, temos que:

$$\begin{aligned}\pi\left(\Sigma|\bullet\right) &\propto |\Sigma|^{-\frac{\delta+k+n+1}{2}}\exp\left\{-\frac{1}{2}\left[tr\left(\Sigma^{-1}\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^{T}\right)+tr\left(\Sigma^{-1}\mathbb{V}\right)\right]\right\} \\ &\propto |\Sigma|^{-\frac{\delta+k+n+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^{T}+\Sigma^{-1}\mathbb{V}\right]\right\} \\ &\propto |\Sigma|^{-\frac{\delta+k+n+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\mathbb{V}+\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^{T}\right)\right]\right\}.\end{aligned}$$

Ou seja, temos o núcleo de uma distribuição inversa Wishart de parâmetros:

$$\begin{aligned}\delta_{post} &= \delta+k+n \\ \mathbb{V}_{post} &= \mathbb{V}+\sum_{i=1}^{n}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^{T}.\end{aligned}$$

Portanto,

$$\Sigma|\bullet \sim \mathcal{IW}\left(\delta_{post};\mathbb{V}_{post}\right).$$

Apêndice 5: Distribuição condicional para C_i

Multiplicando a função de verossimilhança pela função densidade de probabilidades *a priori* para C_i, obtemos:

$$\begin{aligned}
\pi(c_i|\bullet) &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)^T\Sigma^{-1}\left(\mathbf{y}_i - c_i\mathbf{f}(\mathbf{t})\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}\left(c_i-\mu\right)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i^T\Sigma^{-1}\mathbf{y}_i - \mathbf{y}_i^T\Sigma^{-1}c_i\mathbf{f}(\mathbf{t}) - c_i\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{y}_i + c_i^2\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{f}(\mathbf{t})\right]\right. \\
&\quad \left.-\frac{1}{2}\left[\frac{c_i^2}{\sigma^2} - \frac{2\mu c_i}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[c_i^2\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{f}(\mathbf{t}) + \frac{c_i^2}{\sigma^2} - 2c_i\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{y}_i - 2c_i\frac{\mu}{\sigma^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[c_i^2\left(\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{f}(\mathbf{t}) + \frac{1}{\sigma^2}\right) - 2c_i\left(\mathbf{f}(\mathbf{t})\Sigma^{-1}\mathbf{y}_i + \frac{\mu}{\sigma^2}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[c_i^2\left(\frac{\mathbf{f}(\mathbf{t})^T\sigma^2\Sigma^{-1}\mathbf{f}(\mathbf{t}) + 1}{\sigma^2}\right) - 2c_i\left(\frac{\mathbf{f}(\mathbf{t})\sigma^2\Sigma^{-1}\mathbf{y}_i + \mu}{\sigma^2}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2_{post}}\left[c_i^2 - 2c_i\frac{\mathbf{f}(\mathbf{t})^T\sigma^2\Sigma^{-1}\mathbf{y}_i + \mu}{\mathbf{f}(\mathbf{t})\sigma^2\Sigma^{-1}\mathbf{f}(\mathbf{t}) + 1}\right]\right\}.
\end{aligned}$$

Ou seja, temos o núcleo de uma distribuição normal com média

$$\begin{aligned}
\mu_{post(c_i)} &= \frac{\mathbf{f}(\mathbf{t})^T\sigma^2\Sigma^{-1}\mathbf{y}_i + \mu}{\mathbf{f}(\mathbf{t})\sigma^2\Sigma^{-1}\mathbf{f}(\mathbf{t}) + 1} \\
\sigma^2_{post(c_i)} &= \frac{\sigma^2}{\mathbf{f}(\mathbf{t})^T\Sigma^{-1}\mathbf{f}(\mathbf{t}) + 1}.
\end{aligned}$$

Portanto,

$$C_i \sim \mathcal{N}trunc\left(0; \mu_{post(c_i)}, \sigma^2_{post(c_i)}\right),$$

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

Referências Bibliográficas

Abdelhak, K.; Razika, I.; Ali B.; Abdelmalek, A.; Müslüm, A.; Nacer, L. and Nabila, I. Solar photovoltaic power prediction using artificial neural network and multiple regression considering ambient and operating conditions. *Energy Conversion and Management*, **288**, 117186, 2023.

Aldrich, J. R. A. Fisher and the Making of Maximum Likelihood 1912–1922. *Statistical Science*, **12**(3), 162–176, 1997.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723, 1974.

Alkandri, M. and Ahmad, I. Solar power generation forecasting using ensemble approach based on deep learning and statistical methods. *Applied Computing and Informatics*, Vol. ahead-of-print, No. ahead-of-print, 2020. https://doi.org/10.1016/j.aci.2019.11.002.

Amer, H. N.; Dahlan, N. Y.; Azmi, A. M.;Latip, M. F. A.; Onn, M. S. and Tumian, A. Solar power prediction based on Artificial Neural Network guided by feature selection for Large-scale Solar Photovoltaic Plant. *Energy Reports*, **9**, Supplement 12, 262-266, 2023. https://doi.org/10.1016/j.egyr.2023.09.141. ISSN 2352-4847.

Asri, R., Friansa, K. and Siregar, S. Predicting Solar Irradiance Using Regression Model (Case Study: ITERA Solar Power Plant) *IOP Conf. Ser.: Earth Environ. Sci.*, **830**, 012080, 2021.

Bimenyimana, S.; Osarumwense, G. N. and Lingling, L. Output Power Prediction of Photovoltaic Module Using Nonlinear Autoregressive Neural Network. *Journal of Energy, Environmental & Chemical Engineering*, **2**(4), 32-40, 2017.

Wlodzimierz, B. *The Normal Distribution: Characterizations with Applications*. Springer-Verlag, 1995. ISBN 978-0-387-97990-8.

Cramer, J. S. The early origins of the logit model. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, **35**(4): 613-626, 2004.

Casella, G. & Berger, R. L. *Statistical Inference*. Belmont, CA: Duxbury, 2002.

Casella, G. & Robert, C. & Wells, M. Mixture models, latent variables and partitioned importance sampling. *Technical Report*, 2000-03, CREST, INSEE, Paris.

El-Aal, S. A.; ALQABLI, M. A. and NAIM, A. A. Forecasting solar photovoltaic energy production using linear regression-based techniques. *Journal of Theoretical and Applied Information Technology*, **101**(09), 3326-3337, 2023.

Erten, M. Y. and Aydilek, H. Solar Power Prediction using Regression Models *International Journal of Engineering Research and Development*, **14**(3), s333-s342, 2022.

Gallavotti, G. Ergodicity, ensembles, irreversibility in Boltzmann and beyond. *Journal of Statistical Physics*, **78**(5-6), 1571-1589, 1995.

Gelfand, A. E. & Smith, Adrian F. M. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**(410), 398–409, 1990. doi:10.1080/01621459.1990.10476213

Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–472, 1992.

Gelman, A. & Carlin, J. B. & Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

Geman, S. & Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741, (1984). doi:10.1109/TPAMI.1984.4767596.

Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, **115**, 513-585, 1825.

Grimaldi, S. & Kao, S. C. & Castellarin, A. & Papalexiou, S. M. & Viglione, A. & Laio, F. & Aksoy, H. & Gedikli, A. "Statistical Hydrology", Treatise on Water Science, Editor(s): Peter Wilderer, Elsevier, Pages 479-517, (2011).

S. Ibrahim, I. Daut, Y.M. Irwan, M. Irwanto, N. Gomesh, Z. Farhana, Linear Regression Model in Estimating Solar Radiation in Perlis, *Energy Procedia*, **18**, 1402-1412, 2012.

Mellit, A.; Saglam, S. and Kalogirou, S.A. Artificial neural network-based model for estimating the produced power of a photovoltaic module. *Renewable Energy*, **60**, 71-78, 2013. https://doi.org/10.1016/j.renene.2013.04.011.

Orbanz, P. and Teh, Y. W. (2011). Bayesian Nonparametric Models. *In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_66

Pamain, A.; Rao, P.V. K. and Tilya, F. N. Prediction of photovoltaic power output based on different non-linear autoregressive artificial neural network algorithms. *Global Energy Interconnection*, **5**(2), 226-235, 2022.

Parida, B., Iniyan, S., Goie, R. A review of solar photovoltaic technologies. *Renew Sustain Energy Review*, 15:1625-36, 2011.

Patel, J. K. and Campbell, R. B. *Handbook of the Normal Distribution*, 2nd ed., CRC Press, 1996. ISBN 978-0-8247-9342-5.

R Core Team. R Foundation for Statistical Computing; Vienna, Austria: 2020. R: A language and environment for statistical computing, http://www. R-project., 2023.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*, the MIT Press, 2006.

Rogier, J. K and Nawaz M. Forecasting Photovoltaic Power Generation via an IoT Network Using Nonlinear Autoregressive Neural Network. *Procedia Computer Science*, **151**, 643-650, 2019.

Sampaio, P. G. V., González, M. O. A. Photovoltaic solar energy: Conceptual framework. *Renewable and Sustainable Energy Reviews*, **74**, 590–601, 2017.

Saraiva, E. F.; Sauer, L. and Pereira, C. A. B. A hierarchical Bayesian approach for modeling the evolution of the 7-day moving average of the number of deaths by COVID-19 *Journal of Applied Statistics*, **50**(10), 2194-2208, 2023.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464, 1978.

Schulz, E.; Speekenbrink, M. and Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psycology*, **85**, 1-16, 2018.

Sharadga, H., Hajimirza, S. and Balog, R. S. Time series forecasting of solar power generation for large-scale photovoltaic plants *Renewable Energy*, V. 150, 797–807, 2020.

Silveira, J. L., Tuna, C. E., Lamas, W. Q. The need of subsidy for the implementation of photovoltaic solar energy as supporting of decentralized electrical power generation in Brazil. *Renew Sustain Energy Review*, 20:133-41, 2013.

Singh, B. and Pozo, D. A Guide to Solar Power Forecasting using ARMA Models. *IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, Bucharest, Romania, pp. 1-4, 2019.

Souza, G.; Santos, R. R. and Saraiva, E. F. A Log-logistic predictor for power generation in photovoltaic systems. *Energies*, **15**, 5973, 2022.

Vats, D. & Knudson, C. Revisiting the Gelman-Rubin Diagnostic. *Statistical Science*, **4**(36), 518–529, 2021. doi 0.1214/20-STS812.

Wang, J. An Intuitive Tutorial to Gaussian Processes Regression, (2021). https://arxiv.org/pdf/2009.10862.pdf.