Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Programa de Pós-graduação em Estatística

Gisele de Oliveira Maia

Modelos de Séries Temporais Semiparamétricos com Fator Latente

Belo Horizonte

2020

Gisele de Oliveira Maia

Modelos de Séries Temporais Semiparamétricos com Fator Latente

Dissertação apresentada ao Programa de Pós-graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Estatística.

Orientador: Wagner Barreto de Souza

Coorientador: Fernando de Souza Bastos

Belo Horizonte

2020

Agradecimentos

Meu eterno agradecimento aos meus pais, Paulo Sergio e Maria Aparecida, e irmãs Amanda e Paula por todo amor, força e apoio em todo meu percurso desde a graduação.

Agradeço também a toda minha família e amigos de Inhapim, Juiz de Fora e Belo Horizonte.

Meu muito obrigada ao meu orientador Wagner, por toda paciência, ajuda e principalmente aprendizado que obtive com ele. Agradeço também ao meu coorientador Fernando por toda ajuda. Agradeço aos meus orientadores da graduação, Julio Akashi e Clécio Ferreira.

Agradeço a todos os professores que tive durante toda minha trajetória escolar. Agradeço também a todos os funcionários do departamento de Estatística da UFMG.

Agradeço a todos que de alguma forma me ajudaram a realizar este trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro durante o Mestrado.

Resumo

Introduzimos uma classe de modelos de séries temporais semiparamétricos assumindo uma abordagem de quase-verossimilhança conduzida por um processo latente. Mais especificamente, dado o processo latente, apenas especificamos a média e variância condicionais das séries temporais e utilizamos uma abordagem de quase-verossimilhança para estimar os parâmetros relacionados à média. Essa metodologia proposta possui três características marcantes: (i) nenhuma forma paramétrica é assumida para a distribuição condicional das séries temporais, dado o processo latente; (ii) capaz de modelar séries temporais não-negativas, contagens, limitadas/binárias e com valores reais; (iii) não se assume que o parâmetro de dispersão seja conhecido. Além disso, obtemos expressões explícitas para os momentos marginais e para a função de autocorrelação das séries temporais, para que o método de momentos possa ser empregado para estimar o parâmetro de dispersão e também os parâmetros relacionados ao processo latente. Resultados simulados com o objetivo de verificar o procedimento de estimação proposto são apresentados. A análise de dados reais sobre séries temporais de taxa de desemprego e insolação total ilustram o desempenho de nossa metodologia em situações práticas.

Palavras-chave: Série Temporal Limitada. Processo Gaussiano. Análise de Regressão. Processo Gama Deslocada. Estimação Quase-verossimilhança.

Abstract

We introduce a class of semiparametric time series models by assuming a quasi-likelihood approach driven by a latent factor process. More specifically, given the latent process, we only specify the conditional mean and variance of the time series and enjoy a quasi-likelihood function for estimating parameters related to the mean. This proposed methodology has three remarkable features: (i) no parametric form is assumed for the conditional distribution of the time series given the latent process; (ii) able for modelling non-negative, count, bounded/binary and real-valued time series; (iii) dispersion parameter is not assumed to be known. Further, we obtain explicit expressions for the marginal moments and for the autocorrelation function of the time series process so that a method of moments can be employed for estimating the dispersion parameter and also parameters related to the latent process. Simulated results aiming to check the proposed estimation procedure are presented. Real data analysis on unemployment rate and total insolation time series illustrate the potencial for practice of our methodology.

Keywords: Bounded Time Series. Gaussian Process. Regression analysis. Shifted Gamma Process. Quasi-likelihood estimation.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Existem situações em que uma sequência de observações da variável de interesse é observada ao longo do tempo e com isso, naturalmente, tem-se uma estrutura de correlação temporal entre as observações, e o pressuposto de independência é violado. Nesse contexto, Box & Jenkins (1970) introduziram os modelos de séries temporais para lidar com tais situações. A análise e previsão de séries temporais é uma área muito bem desenvolvida, com grande diversidade de trabalhos, como exemplo, Hannan (1971), Brillinger (1981), Liang & Zeger (1986), Brockwell & Davis (1991), Grunwald et al. (1993), Hamilton (1994), Davis et al. (1999), Terui & Dijk (2002), McCabe & Martin (2005), Shumway & Stoffer (2017), dentre vários outros trabalhos.

Cox (1981) caracterizou duas classes de modelos para dados dependentes no tempo, modelos *observation-driven* e *parameter-driven*. Em modelos *observation-driven*, as observações da série temporal $\{Y_t\}$ no tempo $t \in \mathbb{N}$, são condicionadas às observações passadas $Y_1, \cdots, Y_{t-1}$, e desta forma a independência entre as observações é alcançada. Exemplos de trabalhos que estudam estes modelos são Zeger & Qaqish (1988), Benjamin et al. (2003), Davis et al. (2003), Rocha & Cribari-Neto (2009), dentre outros.

Nos modelos *parameter-driven* a autocorrelação é introduzida através de um fator latente $\{\alpha_t\}$, e ao condicionarmos a série temporal $\{Y_t\}$ no fator latente, as observações tornam-se independentes. Estudos voltados para modelos *parameter-driven* em séries temporais de dados de contagem são muito desenvolvidos e encontramos como referências, Zeger (1988), Davis et al. (2000) e Davis & Wu (2009). Zeger (1988) assume que a série temporal de dados de contagem condicionada a um processo autoregressivo latente estacionário é especificada pelos dois primeiros momentos. Sob estas condições a função de quase-verossimilhança foi desenvolvida para estimação dos coeficientes da regressão. Os parâmetros referentes ao processo autoregressivo são estimados pelo método dos momentos.

Davis et al. (2000) estudaram o modelo Poisson para séries temporais de dados de contagem. Eles trabalharam com a distribuição condicional da série temporal dado o fator latente, seguindo a distribuição Poisson. Desta forma, os autores propõem a estimação dos coeficientes da regressão por meio da maximização de uma pseudo-verossimilhança que é baseada nos modelos lineares generalizados com o fator latente suprimido. Os autores mostraram que os estimadores dos coeficientes da regressão são assintoticamentes normais. Assim, calcularam a variância assintótica para os estimadores dos coeficientes. Os parâmetros referentes ao fator latente são estimados pelo método dos momentos.

Davis & Wu (2009) estenderam a teoria desenvolvido por Davis et al. (2000) para dados que são condicionalmente superdispersos, nesse caso, assumiram que as observa-

ções da série temporal condicionadas ao fator latente possuem distribuição Binomial Negativa. Além disso, na estimação dos coeficientes da regressão maximizaram a pseudo-verossimilhança. Para a estimação dos parâmetros do fator latente utilizaram o método de mínimos quadrados ordinários. Também apresentaram todo referencial teórico para o cálculo da variância assintótica dos estimadores dos coeficientes da regressão.

Outras referências de trabalhos que estudam os modelos *parameter-driven* são Brännäs & Johansson (1994), que abordaram dados de contagem, estudando o modelo proposto por Zeger (1988), mas com foco no estimador de máxima verossimilhança do modelo Poisson. Jørgensen et al. (1995) propuseram o modelo de espaço de estados não estacionário para dados de contagem multivariados longitudinais, em que o processo latente segue um Gamma Markov. Jørgersen et al. (1996) generalizaram o modelo apresentado por Jørgensen et al. (1995) para o modelo baseado na distribuição da classe exponencial Tweedie para modelos de dispersão (Jørgensen, 1987). Jørgersen & Song (2007) estenderam o processo estudado por Jørgersen et al. (1996) para modelo espaço de estados estacionário. Temos também como referência no assunto, Fahrmeir & Tutz (1994), Chan & Ledolter (1995) e Davis & Rodriguez (2005).

Há vantagens e desvantagens em relação a cada uma das duas classes definidas por Cox (1981). Nos modelos *observation-driven* obtemos a função de verossimilhança e consequentemente, estimamos os parâmetros com facilidade, mas a desvantagem está na dificuldade de provas de propriedades assintóticas, estacionalidade e ergodicidade. Para os modelos *parameter-driven*, a prova de propriedades assintóticas, estacionalidade e ergodicidade não são obtidas de forma trabalhosa, enquanto que não conseguimos expressar de forma fechada a função de verossimilhança, o que influencia na dificuldade de estimação dos parâmetros.

Nesta dissertação utilizamos a abordagem de modelos *parameter-driven*. Especificamente, seja $\{Y_t\}$ e $\{\alpha_t\}$ denotando a variável observada e o fator latente, respectivamente, no tempo t. A variável observada $\{Y_t\}$ será condicionada ao fator latente e às covariáveis. Uma das expressões que definem o modelo proposto é definida na forma $g(\mu_t) = x_t^\top \boldsymbol{\beta} + \alpha_t$, onde x_t é o vetor de covariáveis com dimensão $q \times 1$, $\boldsymbol{\beta}$ o vetor com os coeficientes da regressão com dimensão $q \times 1$, $g(\cdot)$ a função de ligação e $\mu_t = E(Y_t|x_t, \alpha_t)$. Especificamos somente o primeiro e segundo momento da distribuição condicional $Y_t|\alpha_t$. Nossa proposta aqui é trabalhar com a estimação dos coeficientes da regressão considerando a função de quase-verossimilhança. Trabalhamos com duas configurações para o fator latente, ele será definido como um processo AR(1) Gaussiano e também como um processo Gama deslocado. Na estimação do parâmetro de dispersão referente ao modelo de quase-verossimilhança e aos parâmetros do fator latente utilizamos o método dos momentos. Ilustramos nossa metodologia aplicada a dados temporais de contagem, limitados/binários, não-negativos e com valores reais.

Uma visão geral de trabalhos que discutem a abordagem de quase-verossimilhança na análise de regressão para séries temporais pode ser visto em Zeger & Qaqish (1988) que consideraram modelos de Markov na classe *observation-driven*. Eles focaram em dados com distribuição Gaussiana e Gamma. Christou & Fokianos (2014) estudaram a inferência e diagnóstico para séries temporais de dados de contagem, com mais interesse no processo Binomial Negativa, com a inclusão de um mecanismo de retorno (Fokianos et al., 2009). Eles abordaram a função de quase-verossimilhança, baseando-se na função de log-verossimilhança de Poisson. Christou & Fokianos (2015) estenderam o estudo realizado por Christou & Fokianos (2014) e trabalharam com modelos autoregressivos mistos não-linear Poisson. Outras referências sobre o assunto são Zeger & Liang (1986), Heyde (1997), Berkes et al. (2003), Francq & Zakoian (2004) e Mikosch & Straumann (2006).

A proposta desta dissertação é apresentar um modelo flexível e semiparamétrico, dado que não estamos especificando a distribuição condicional. Diferente de como ocorre nos trabalhos de Davis et al. (2000), Davis & Wu (2009) e Jørgensen & Song (2007). Além disso, o modelo introduzido nesta dissertação pode ser utilizado para dados temporais contínuos, discretos e limitados. Zeger (1988) e Brännäs & Johansson (1994) trabalharam com a abordagem semiparamétrica, mas consideraram somente dados de contagem. Zeger & Qaqish (1988) e Christou & Fokianos (2015) lidaram com a abordagem semiparamétrica, mas somente no contexto dos modelos *observation-driven*. Portanto, o modelo proposto une a classe de modelos *parameter-driven* com a abordagem semiparamétrica, com o acréscimo de lidar com uma grande variedade de dados temporais.

A dissertação está estruturada da seguinte forma. No Capítulo 2 introduzimos os conceitos sobre a função de quase-verossimilhança e estimação dos parâmetros através da maximização da função. Na sequência, propomos o modelo de séries temporais semiparamétrico com fator latente, discutindo as definições para cada suporte da série temporal. Apresentamos quantidades que posteriormente serão utilizadas na parte inferencial do modelo. No Capítulo 3 apresentamos a abordagem utilizada para estimação dos parâmetros, como também um estudo de simulação para verificação do método proposto. No Capítulo 4 apresentamos duas aplicações a dados reais. Uma aplicação é realizada a dados não-negativos e outra a dados de proporção. No Capítulo 5 apresentamos as conclusões do trabalho e pesquisas futuras a serem abordadas.

2 Modelos de Séries Temporais Semiparamétricos com Fator Latente

2.1 Modelo de Quase-Verossimilhança

Modelos normais lineares são empregados em diversas áreas da ciência e, durante anos, algum tipo de transformação da variável de interesse era sugerida, sempre que tal variável não apresentasse normalidade. A transformação mais conhecida foi proposta por Box e Cox (1964). Outra proposta foi dada por Nelder e Wedderburn (1972), que introduziram os modelos lineares generalizados (MLGs). Tais modelos podem ser ajustados a dados de diferentes distribuições da família exponencial.

Estendendo a ideia dos MLGs para situações mais gerais, Wedderburn (1974) propôs uma função denominada função de quase-verossimilhança, na qual estão contidas algumas funções de verossimilhança da família exponencial. Nesta seção utilizamos como referêncial teórico o livro de Paula (2004).

Seja Y uma variável aleatória de interesse. A função log-quase-verossimilhança é definida por

$$Q(\mu; y) = \frac{1}{\phi}\int_y^\mu \frac{y-t}{V(t)}\ dt,$$

em que $\phi > 0$ é um parâmetro de dispersão, $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ e $V(t)$ é uma função positiva e conhecida. Segue que

$$\begin{aligned}\frac{\partial Q(\mu; y)}{\partial \mu} &= \frac{y-t}{\phi V(t)}\bigg|_y^\mu \\ &= \frac{y-\mu}{\phi V(\mu)}.\end{aligned}$$

Para a construção da função de quase-verossimilhança não é necessário definir uma distribuição para a variável de interesse Y, apenas ter de informação o primeiro e segundo momentos de Y.

Assumimos as seguintes condições de regularidade para o modelo

$$E\left\{\frac{\partial Q(\mu; Y)}{\partial \mu}\right\} = 0,$$

e

$$E\left[\left\{\frac{\partial Q(\mu;Y)}{\partial \mu}\right\}^2\right] = -E\left\{\frac{\partial^2 Q(\mu;Y)}{\partial^2 \mu}\right\}.$$

Com isto, segue que

$$E(Y) = \mu \quad e \quad Var(Y) = \phi V(\mu).$$

A média da variável resposta é $\mu \in \mathbb{R}$ e a variância de Y é proporcional a $V(\mu)$.

Uma terceira propriedade dada diz que a informação a respeito de μ quando se conhece apenas a média e variância é menor do que a informação a respeito de μ quando se conhece a distribuição de Y, ou seja

$$-E\left\{\frac{\partial^2 Q(\mu;Y)}{\partial \mu^2}\right\} \leq -E\left\{\frac{\partial^2 L(\mu;Y)}{\partial \mu^2}\right\}$$

em que $L(\mu;Y)$ é a função de verossimilhança de Y.

Supondo que $Y_1, \ldots, Y_n$ são variáveis aleatórias independentes e denotando por $Q(\mu_i, y_i)$, $i = 1, \ldots, n$ a função log-quase-verossimilhança, temos que

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} Q(\mu_i; y_i).$$

Consideramos a seguinte estrutura de regressão para a média

$$g(\mu_i) = \eta_i = x_i^\top \beta,$$

em que $x_i = (x_{i1}, \ldots, x_{iq})^\top$ contém os valores das variáveis explicativas, $\beta = (\beta_1, \ldots, \beta_q)^\top$ é o vetor de coeficientes da regressão e $g(\cdot)$ é uma função de ligação.

A função quase-escore para $\boldsymbol{\beta}$ é expressa na forma

$$U_{\boldsymbol{\beta}} = \frac{\partial Q(\mu; y)}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} D^\top V^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

em que $D = \dfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = W^{1/2} V^{1/2} X$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $y = (y_1, \ldots, y_n)^\top$, $V = \text{diag}(V_1, \ldots, V_n)$, $W = \text{diag}(w_1, \ldots, w_n)$, com $w_i = (d\mu/d\eta)_i^2 / V_i$ e X é uma matriz $n \times q$ de linhas $x_i^\top$, $i = 1, \ldots, n$. A matriz de quase-informação para $\boldsymbol{\beta}$ é dada por

$$K_{\boldsymbol{\beta}\boldsymbol{\beta}} = -E\left\{\frac{\partial^2 Q(\boldsymbol{\mu}; \mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right\} = \frac{1}{\sigma^2} D^\top V^{-1} D.$$

Através da solução da equação $U_{\boldsymbol{\beta}} = 0$ obtemos a estimativa de quase-verossimilhança para $\boldsymbol{\beta}$. Tendo como referência o livro de Paula (2004), o método de Fisher é utilizado na solução. Temos o seguinte processo iterativo

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{(-U'_{\boldsymbol{\beta}})^{-1}\right\}^{(m)} U^{(m)}_{\boldsymbol{\beta}},$$

com $m = 0, 1, 2, \ldots$. Na aplicação do método escore de Fisher substituímos a matriz $-U'_{\boldsymbol{\beta}}$ pelo correspondente valor esperado $K_{\boldsymbol{\beta}\boldsymbol{\beta}}$. Assim, segue que

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{D^{(m)\top} V^{-(m)} D^{(m)}\right\}^{-1} D^{(m)\top} V^{-(m)} \left\{y - \mu^{(m)}\right\}, m = 0, 1, 2, \ldots.$$

O processo iterativo precisa ser iniciado com uma quantidade $\boldsymbol{\beta}^{(0)}$. O parâmetro de dispersão ϕ é estimado, separadamente, pelo método dos momentos. Podemos verificar que

$$Var\left\{\frac{(Y_i - \mu_i)}{\sqrt{\phi}\sqrt{V(\mu_i)}}\right\} = 1,$$
$$Var\left\{\frac{(Y_i - \mu_i)}{\sqrt{V(\mu_i)}}\right\} = \phi,$$

portanto

$$\widehat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}. \tag{2.1}$$

2.2 Modelos de Séries Temporais Semiparamétricos: Definição

Seja $\{Y_t\}$ uma série temporal e suponha que para $t \in \mathbb{Z}$, x_t é uma covariável com dimensão $q \times 1$ observada e cujo primeiro componente é um. Em alguns casos, x_t pode depender do tamanho n da amostra e formar uma matriz triangular x_{nt}. Assumindo que, condicionado ao fator latente $\{\alpha_t\}$, as variáveis aleatórias $Y_1, \ldots, Y_n$ são condicionalmente independentes. A proposta de trabalhar com o modelo semiparamétrico surge a partir da não-especificação da distribuição de $Y_t|\alpha_t$. Apenas definimos o primeiro e segundo momento da distribuição condicional.

A nossa classe proposta de modelos é definida completamente através das três expressões dadas na forma

$$g(\widetilde{\mu}_t) = x_{nt}^{\top}\beta + \alpha_t,$$
$$E(Y_t|\alpha_t) = \widetilde{\mu}_t,$$
$$Var(Y_t|\alpha_t) = \phi V(\widetilde{\mu}_t),$$

em que $\beta = (\beta_1, ..., \beta_q)^\top$ é o vetor de coeficientes da regressão, $g(\cdot)$ a função de ligação, ϕ o parâmetro de dispersão e $V(\widetilde{\mu}_t)$ a função de variância.

A primeira expressão parte da definição de MLGs, onde temos a relação entre a média de $Y_t|\alpha_t$ e o preditor linear, através de uma função de ligação. No preditor linear temos o acréscimo do fator latente. As duas expressões, onde definimos a esperança e a variância da distribuição condicional, são oriundas da definição dos modelos de quase-verossimilhança. As três expressões dadas são suficientes para definir o modelo e através das mesmas calcular quantidades marginais para $\{Y_t\}$ que, posteriormente, serão utilizadas na estimação dos parâmetros do modelo. Obtemos expressões para a esperança, a variância, a função de autocovariância e a função de autocorrelação da série temporal, pois estas são as quantidades necessárias para as equações de estimação dos parâmetros.

Uma das contribuições do modelo proposto é poder aplicá-lo em uma grande variedade de dados de séries temporais. Com isso, trabalhamos com diferentes suportes para as observações da série temporal, a saber dados não-negativos, dados limitados/binários e dados com valor real. Para cada suporte da série temporal, apenas definimos uma função de ligação e a função de variância. Não definimos uma distribuição para $Y_t|\alpha_t$, apenas o primeiro e segundo momento. Apresentamos toda construção e definições do modelo, como também a abordagem para estimação dos parâmetros. Temos como parâmetros do modelo, os coeficientes da regressão, o parâmetro de dispersão do modelo de quase-verossimilhança e os parâmetros referentes ao fator latente.

Trabalhos como de Davis et al. (2000) e Davis & Wu (2009) comprovaram que os coeficientes da regressão estimados pela maximização da função pseudo-verossimilhança, ignorando a presença do fator latente, apresentam bons resultados, mas o mesmo não ocorre para os erros padrão das estimativas dos coeficientes calculados pelo modelo de verossimilhança. Pela forma que os erros padrão são calculados, ignorando a presença do fator latente e a estrutura temporal do modelo, obtemos estimadores inconsistentes em relação ao modelo. Por isso, os autores realizaram simulações para o cálculo dos desvios padrão empíricos e também apresentaram referencial teórico para o cálculo dos erros padrão assintóticos dos estimadores dos coeficientes. Observando que os desvios padrão empíricos estavam de acordo com os erros padrão assintóticos. Desta forma, nesta dissertação a estimação do vetor de coeficientes $\boldsymbol{\beta}$ é realizada por meio da maximização da função de quase-verossimilhança, ignorando a presença do fator latente. Observamos, por intermédio de estudos de simulações, que esta abordagem também produz resultados consistentes, mas apresentando o mesmo problema para os erros padrão calculados pelo modelo de quase-verossimilhança. Como solução, utilizamos o método de Monte Carlo para o calculo dos desvios padrão empíricos que serão utilizados como uma opção para os erros padrão das estimativas dos coeficientes. Para a estimação do parâmetro de dispersão e parâmetros do fator latente trabalhamos com o método dos momentos.

Segundo Davis et al. (2000) e Davis & Wu (2009) devemos obedecer a seguinte suposição para a validade do modelo proposto.

Suposição: Seja $\{Y_t\}$ uma série temporal. Assumimos que a média marginal de Y_t satisfaz a seguinte condição

$$E(Y_t) = E(h(x_{nt}^\top \beta + \alpha_t)) = h(x_{nt}^\top \beta), \tag{2.2}$$

em que h é a inversa da função de ligação g. Ou seja, a média marginal da série temporal não deve depender de momentos do fator latente.

Apresentamos duas configurações que utilizaremos nesta dissertação para o fator latente α_t dependente do tempo incluído no modelo. Seguindo os estudos de Zeger (1988), Davis et al. (2000) e Davis & Wu (2009) assumiremos o fator latente como um processo autoregressivo Gaussiano de ordem um (denotado por AR(1) Gaussiano) para os modelos de contagem, não-negativo e com valor real. Mais explicitamente, temos que o processo AR(1) Gaussiano possui a seguinte dinâmica estocástica

$$\alpha_t = c + \rho\alpha_{t-1} + \eta_t,$$

em que $c \in \mathbb{R}$ é o intercepto, ρ o parâmetro de autocorrelação do processo e η_t o ruído branco normal independente e identicamente distribuído com média zero e variância σ_η^2. A condição $|\rho| < 1$ é suficiente para que α_t seja estacionário.

A média e variância do processo são dadas por

$$E(\alpha_t) = c + \rho E(\alpha_{t-1}) + E(\eta_t) = \frac{c}{1-\rho},$$

e

$$Var(\alpha_t) = \rho^2 Var(\alpha_{t-1}) + Var(\eta_t) = \frac{\sigma_\eta^2}{1-\rho^2} = \sigma^2.$$

Desta forma, temos que $\alpha_t \sim N\left(\frac{c}{1-\rho}, \sigma^2\right)$. Para o caso em que não há o intercepto, $\alpha_t = \rho\alpha_{t-1} + \eta_t$, o fator latente apresentará a seguinte configuração $\alpha_t \sim N(0, \sigma^2)$.

O parâmetro de autocorrelação do processo ρ, é dado como a razão entre a função de autocovariância (γ) e a variância (σ^2) do processo. Na Figura 1, temos a representação gráfica de uma trajetória simulada de um processo estacionário AR(1) Gaussiano, com os respectivos gráficos da função de autocorrelação (FAC) e função de autocorrelação parcial (FACP) na Figura 2. Simulamos um processo AR(1) Gaussiano com a seguinte configuração $c = -0,024, n = 160, \rho = 0,82, \sigma^2 = 0,262$ e $\sigma_\eta^2 = 0,086$.

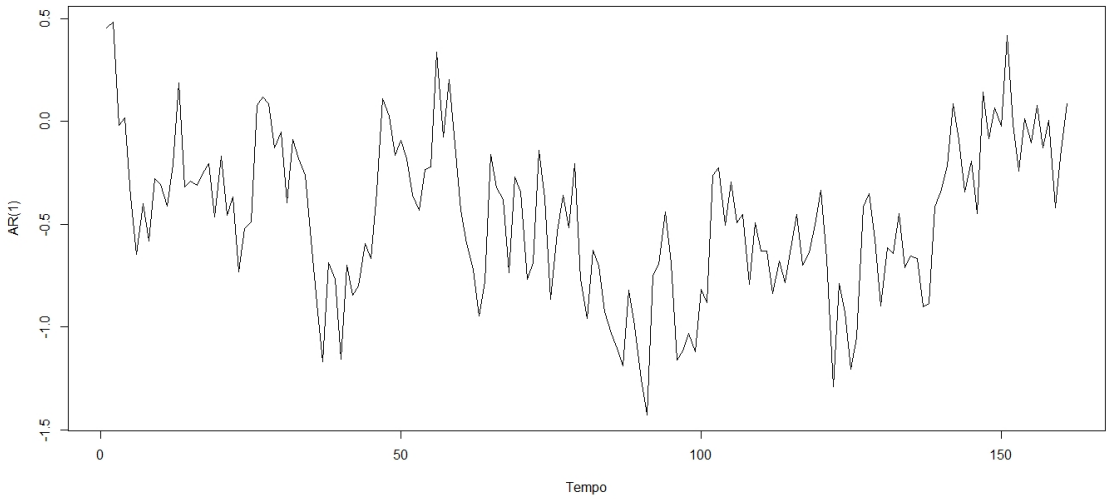

Figura 1 – Representação gráfica de uma trajetória simulada de um processo estacionário AR(1).

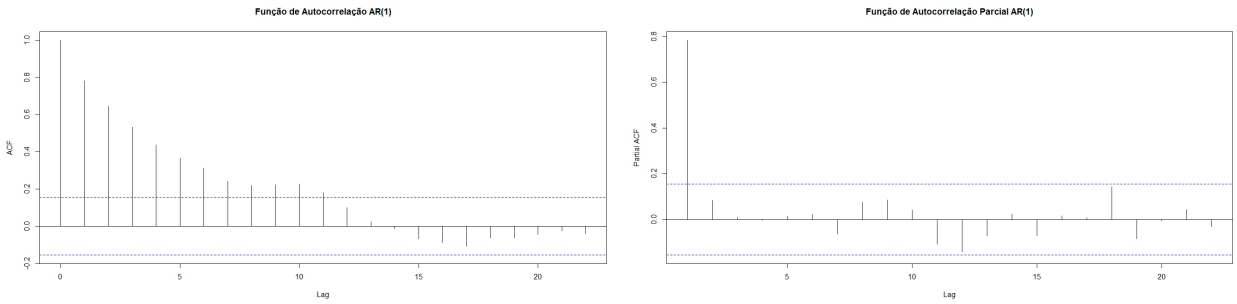

Figura 2 – Gráfico FAC (à esquerda) e FACP (à direita) para uma trajetória simulada de um processo estacionário AR(1).

Na Figura 1, observamos o comportamento estacionário do processo, ou seja, a média e variância são constantes ao longo do tempo. No gráfico FAC para modelos AR(1), Figura 2 à esquerda, temos o comportamento semelhante a uma queda exponencial, para o caso onde $\rho > 0$. Quando $\rho < 0$ o comportamento é semelhante a uma queda exponencial alternando entre valores positivos e negativos. O gráfico FACP, Figura 2 à direita, apresenta apenas um pico significativo na primeira defasagem, indicando a ordem do processo. O pico é positivo se $\rho > 0$ e negativo de $\rho < 0$.

Para o fator latente incluído no modelo apresentamos apenas sua distribuição, média e variância, mas em modelos de espaço de estados, como estudado por Jørgensen & Song (2007), mais quantidades para o fator latente são calculadas e utilizadas na metodologia proposta pelos autores como uma extensão do filtro de Kalman (Kalman, 1960).

Agora, suponha que, para $t, j \in \mathbb{N}$, $\alpha_{t+j} \sim N(\tau_1, \kappa_1^2)$, $\alpha_t \sim N(\tau_2, \kappa_2^2)$ e $\omega(j) = Corr(\alpha_{t+j}, \alpha_t)$. A distribuição condicional de $\alpha_{t+j}|\alpha_t$ é dada por

$$f(\alpha_{t+j}|\alpha_t = z) = c(\kappa_1^2, \kappa_2^2, \omega(j), \tau_2, z) exp\left\{-\frac{1}{2\kappa_1^2(1-\omega^2(j))}\left[\alpha_{t+j} - \tau_1 - \frac{\omega(j)\kappa_1}{\kappa_2}(z-\tau_2)\right]^2\right\},$$

em que a constante $c(\kappa_1^2, \kappa_2^2, \omega(j), \tau_2, z)$ é dada por $(2\pi\kappa_1^2(1-\omega^2(j)))^{-1/2}$.

Esta é a densidade da distribuição normal com média $\tau_1 + \frac{\omega(j)\kappa_1}{\kappa_2}(z-\tau_2)$ e variância $\kappa_1^2(1-\omega(j)^2)$. Assim, seque que

$$\alpha_{t+j}|\alpha_t = z \sim N\left(\tau_1 + \frac{\omega(j)\kappa_1}{\kappa_2}(z-\tau_2), \kappa_1^2(1-\omega^2(j))\right).$$

Agora discutiremos a configuração que será considerada para o fator latente nos modelos limitados e binários. Ela será baseada no processo autoregressivo de ordem um Gama (com média 1) proposta por Sim (1990). Dizemos que a sequência $\{Z_t\}_{t\in\mathbb{N}}$ segue um processo autoregressivo de ordem um Gama (denotado por GAR(1)) se satisfaz

$$Z_t = \kappa \odot Z_{t-1} + \eta_t, \quad t \in \mathbb{N}, \quad Z_0 \sim \mathrm{G}(1/\sigma^2, 1/\sigma^2),$$

onde o operador $\odot$ é definido por $\kappa \odot Z_{t-1} \stackrel{d}{=} \sum_{i=1}^{N_{t-1}} W_i$, com $N_{t-1}|Z_{t-1} = z \sim \text{Poisson}(\alpha\rho z)$, $\{W_i\}_{i=1}^{\infty} \overset{iid}{\sim} \text{Exponencial}(\kappa)$ e $\{\eta_t\}_{t=1}^{\infty} \overset{iid}{\sim} \mathrm{G}(\sigma^2, \kappa)$ é assumido ser independente e $\kappa = \frac{1}{\sigma^2(1-\rho)}$, para $\sigma^2 > 0$ e $\rho \in (0,1)$. Aqui, $G(\sigma^2, \kappa)$ denota a distribuição Gama com parâmetros de forma e escala σ^2 e κ, respectivamente.

O processo GAR depende dos parâmetros σ^2 e ρ. O parâmetro ρ controla a dependência desse processo, já que $\text{corr}(Z_{t+k}, Z_t) = \rho^k$ para $t, k \in \mathbb{N}$. As marginais deste modelo possuem distribuição Gama com média 1 e variância σ^2, portanto o modelo é estacionário; veja Sim, (1990). Uma forte propriedade de mistura deste processo foi estabelecida recentemente por Barreto-Souza e Ombao (2019). Portanto, definimos o processo latente $\{\alpha_t\}_{t\in\mathbb{N}}$ para os casos limitados e binários por

$$\alpha_t = Z_t + \log E\left(\exp(-Z_t)\right) = Z_t - \frac{1}{\sigma^2}\log(1+\sigma^2), \quad t \in \mathbb{N}. \tag{2.3}$$

O processo Gama deslocado $\{\alpha_t\}_{t\in\mathbb{N}}$ acima deve satisfazer a suposição dada em (2.2). Uma abordagem semelhante foi considerada por Davis & Davis (2009) para lidar com dados binários. Nesse artigo, os autores assumiram uma espécie de processo exponencial deslocado. Com base em nossa abordagem, o termo deslocado é muito simples, em contraste com o termo do processo exponencial considerado em Davis & Wu (2009) (consulte a Experiência 2, página 743).

Nas próximas subseções apresentamos as definições do modelo proposto de acordo com o suporte da série temporal. Focamos no cálculo das quantidades marginais para a série temporal, dado que serão utilizadas na estimação dos parâmetros do fator latente.

2.2.1 Modelo para Séries Temporais Limitadas e Binárias

Sejam $Y_1, \ldots, Y_t$ as observações da série temporal com suporte em (0,1), $\{0,1\}$ ou $\{0,1,\ldots,m\}$, com $m \in \mathbb{Z}^+$. Portanto, aqui são permitidas dados de séries temporais de proporções/taxas (contínuas limitadas), binárias e binomiais . Definimos o modelo usando como função de ligação $-\log(\widetilde{\mu}_t)$ e função de variância $V(\widetilde{\mu}_t) = \widetilde{\mu}_t(1-\widetilde{\mu}_t)$, para $\widetilde{\mu}_t \in (0,1)$.

Por conveniência definimos o processo $\epsilon_t = \exp(-\alpha_t)$, que é uma série temporal não-negativa estritamente estacionária quando α_t é estritamente estacionária. Assumimos o fator latente α_t como um processo Gama deslocado, definido em (2.3). A condição $E(\epsilon_t) = 1$, é imposta por razões de identificabilidade, pois se $E(\epsilon_t) \neq 1$, a média pode ser absorvida pelo intercepto da regressão linear e além disso, força a média marginal da série temporal a não depender de momentos do fator latente. Desta forma, podemos definir o modelo para séries temporais limitadas e binárias como na forma abaixo

$$\begin{aligned} -\log \widetilde{\mu}_t &= x_{nt}^\top \beta + \alpha_t, \\ E(Y_t|\alpha_t) &= \widetilde{\mu}_t = \exp(-x_{nt}^\top \beta)\epsilon_t, \\ \mathrm{Var}(Y_t|\alpha_t) &= \phi V(\widetilde{\mu}_t) = \phi\widetilde{\mu}_t(1-\widetilde{\mu}_t), \end{aligned}$$

em que $\phi = 1$ e $\phi = m$ para o caso binário e binomial, respectivamente. Para o caso contínuo limitado, temos que $0 < \phi < 1$. O vetor β é tal que $x_{nt}^\top \beta > 0$, desde que $\widetilde{\mu}_t \in (0,1)$ para todo $t \in \mathbb{N}$. Calculando quantidades marginais para a série temporal, temos que a esperança Y_t é dada por

$$E(Y_t) = E(E(Y_t|\alpha_t)) = E(\exp(-x_{nt}^\top\beta)\epsilon_t) = \exp(-x_{nt}^\top\beta)E(\epsilon_t) = \exp(-x_{nt}^\top\beta) = \mu_t.$$

desde que $E(\epsilon_t) = E(\exp(-\alpha_t)) = \frac{E(\exp(-Z_t))}{E(\exp(-Z_t))} = 1$. Depois de manipulações algébricas, obtemos que a variância de Y_t é

$$\begin{aligned} Var(Y_t) &= E\left\{Var(Y_t|\alpha_t)\right\} + Var\left\{E(Y_t|\alpha_t)\right\} \\ &= \phi\mu_t + \mu_t\left\{(1-\phi)\left(\frac{(1+\sigma^2)^2}{1+2\sigma^2}\right)^{1/\sigma^2} - 1\right\}. \end{aligned} \tag{2.4}$$

Para a função de autocovariância temos, para $k \neq 0$,

$$\begin{aligned} Cov(Y_{t+k}, Y_t) &= E\left\{Cov\left(Y_{t+k}, Y_t\right)|\alpha_t\right\} + Cov\left\{E(Y_{t+k}|\alpha_{t+k}), E(Y_t|\alpha_t)\right\} \\ &= 0 + Cov\left\{E(Y_{t+k}|\alpha_{t+k}), E(Y_t|\alpha_t)\right\} \\ &= \mu_{t+k}\mu_t(1+\sigma^2)^{2/\sigma^2}Cov(\exp(-Z_{t+k}), \exp(-Z_t)). \end{aligned}$$

Da equação (2.6) em Sim (1990), obtemos uma expressão explicita para a função conjunta de Laplace de (Z_{t+k}, Z_t). Usando esta expressão, obtemos que

$$Cov(Y_{t+k}, Y_t) = \mu_{t+k}\mu_t\left\{\left(\frac{(1+\sigma^2)^2}{1+2\sigma^2+(\sigma^2)^2(1-\rho^k)}\right)^{1/\sigma^2} - 1\right\}. \tag{2.5}$$

A expressão para a função de autocorrelação é obtida imediatamente usando (2.4) e (2.5). O modelo de série temporal para dados binários proposto aqui é uma alternativa ao modelo discutido por Davis e Wu (2009), pois estamos usando um processo latente diferente. Mais uma vez chamamos a atenção de que o termo deslocado considerado aqui é mais simples que o termo nesse artigo, que envolve uma multiplicação de número infinito de termos. Além disso, nossa metodologia proposta nos permite lidar com dados de séries temporais contínuas limitadas.

2.2.2 Modelo para Séries Temporais Não-negativas

Seja $Y_1, \ldots, Y_t$ uma série temporal com observações pertencentes ao conjunto dos números reais não-negativos, ou seja, inclui dados discretos e contínuos não-negativos. Definimos a função de ligação como $\log(\tilde{\mu}_t)$ e função de variância $V(\mu_t) = \tilde{\mu}_t^p$, onde p é uma constante positiva. Dados de contagem e quaisquer dados cujo domínio esteja definido nos reais não-negativos estão inclusos nesta definição.

Definimos $\epsilon_t = \exp(\alpha_t)$, como uma série temporal não-negativa estritamente estacionária, seguindo a distribuição Log-Normal com média igual a 1, variância igual a $\sigma_\epsilon^2 = \exp(\sigma^2) - 1$, e função de autocovariância e autocorrelações dadas, respectivamente, por

$$\gamma_\epsilon(k) \equiv \text{cov}(\epsilon_{t+k}, \epsilon_t) = \exp(\gamma(k)) - 1$$

e

$$\rho_\epsilon(k) \equiv \text{corr}(\epsilon_{t+k}, \epsilon_t) = \frac{\exp(\gamma(k)) - 1}{\exp(\sigma^2) - 1},$$

onde $\gamma(k)$ é a função de autocovariância do fator latente AR(1) Gaussiano. A função de autocorrelação do fator latente AR(1) Gaussiano é dada por $\rho(k) = \gamma(k)/\sigma^2$. Desta forma, temos que o modelo para séries temporais não-negativas é definido completamente na forma

$$\begin{aligned}
\log \tilde{\mu}_t &= x_{nt}^\top \beta + \alpha_t, \\
E(Y_t|\alpha_t) &= \tilde{\mu}_t = \exp(x_{nt}^\top \beta + \alpha_t) = \exp(x_{nt}^\top \beta)\epsilon_t, \\
\text{Var}(Y_t|\alpha_t) &= \phi V(\tilde{\mu}_t) = \phi \tilde{\mu}_t^p,
\end{aligned}$$

Das expressões acima, obtemos a esperança de Y_t

$$E(Y_t) = E(E(Y_t|\alpha_t)) = E(\exp(x_{nt}^\top \beta)\epsilon_t) = \exp(x_{nt}^\top \beta)E(\epsilon_t),$$

como pela condição a $E(\epsilon_t)$ é igual a um, decorre que o valor do intercepto c do fator latente deve ser igual a $-\frac{(1-\rho)\sigma^2}{2}$. Assim, $\alpha_t \sim N\left(-\frac{\sigma^2}{2}, \sigma^2\right)$.

Então, satisfeita a condição temos que

$$E(Y_t) = \exp(x_{nt}^\top \beta) = \mu_t,$$

e a variância de Y_t é dada por

$$\begin{aligned} Var(Y_t) &= E\left\{Var(Y_t|\alpha_t)\right\} + Var\left\{E(Y_t|\alpha_t)\right\} \\ &= E\left\{\phi(\tilde{\mu}_t)^p\right\} + Var\left\{e^{x_{nt}^T}e^{\alpha_t}\right\} \\ &= \phi\mu_t^p(\sigma_\epsilon^2+1)^{\frac{p^2-p}{2}} + \mu_t^2\sigma_\epsilon^2. \end{aligned}$$

Para os casos particulares, onde $p = 1$, $p = 2$ e $p = 3$ temos as seguintes variâncias, respectivamente,

$$\begin{aligned} Var(Y_t) &= \phi\mu_t + \sigma_\epsilon^2\mu_t^2, \\ Var(Y_t) &= \phi\sigma_\epsilon^2\mu_t^2 + \phi\mu_t^2 + \sigma_\epsilon^2\mu_t^2, \\ Var(Y_t) &= \phi(\sigma_\epsilon^2+1)^3\mu_t^3 + \sigma_\epsilon^2\mu_t^2. \end{aligned}$$

Para $k \neq 0$, temos a função de autocovâriancia dada por

$$\begin{aligned} Cov(Y_{t+k}, Y_t) &= Cov\left\{E(Y_{t+k}|\alpha_{t+k}), E(Y_t|\alpha_t)\right\} \\ &= Cov\left\{e^{x_{n,t+k}^\top\beta}e^{\alpha_{t+k}}, e^{x_{nt}^\top\beta}e^{\alpha_t}\right\} \\ &= \mu_{t+k}\mu_t\gamma_\epsilon(k). \end{aligned}$$

Por fim, a função de autocorrelação é dada na seguinte forma

$$\begin{aligned} Corr(Y_{t+k}, Y_t) &= \frac{Cov(Y_{t+k}, Y_t)}{\sqrt{Var(Y_{t+k})Var(Y_t)}} \\ &= \frac{\mu_{t+k}\mu_t\gamma_\epsilon(k)}{\sqrt{\left[\phi\mu_{t+k}^p(\sigma_\epsilon^2+1)^{\frac{p^2-p}{2}} + \mu_{t+k}^2\sigma_\epsilon^2\right]\left[\phi\mu_t^p(\sigma_\epsilon^2+1)^{\frac{p^2-p}{2}} + \mu_t^2\sigma_\epsilon^2\right]}} \\ &= \frac{\rho_\epsilon(k)}{\sqrt{\left[\phi\sigma_\epsilon^{-2}\mu_{t+k}^{p-2}(\sigma_\epsilon^2+1)^{\frac{p^2-p}{2}} + 1\right]\left[\phi\sigma_\epsilon^{-2}\mu_t^{p-2}(\sigma_\epsilon^2+1)^{\frac{p^2-p}{2}} + 1\right]}}. \end{aligned}$$

Para os valores de $p = 1$, $p = 2$ e $p = 3$ temos as funções de autocorrelação, respectivamente,

$$\begin{aligned} Corr(Y_{t+k}, Y_t) &= \frac{\rho_\epsilon(k)}{\sqrt{\left[\phi\sigma_\epsilon^{-2}\mu_{t+k}^{-1} + 1\right]\left[\phi\sigma_\epsilon^{-2}\mu_t^{-1} + 1\right]}}, \\ Corr(Y_{t+k}, Y_t) &= \frac{\rho_\epsilon(k)}{\sqrt{\left[\phi\sigma_\epsilon^{-2}(\sigma_\epsilon^2+1) + 1\right]\left[\phi\sigma_\epsilon^{-2}(\sigma_\epsilon^2+1) + 1\right]}} = \frac{\rho_\epsilon(k)}{\phi\sigma_\epsilon^{-2}(\sigma_\epsilon^2+1) + 1}, \\ Corr(Y_{t+k}, Y_t) &= \frac{\rho_\epsilon(k)}{\sqrt{\left[\phi\sigma_\epsilon^{-2}\mu_{t+k}(\sigma_\epsilon^2+1)^3 + 1\right]\left[\phi\sigma_\epsilon^{-2}\mu_t(\sigma_\epsilon^2+1)^3 + 1\right]}}. \end{aligned}$$

2.2.3 Modelo para Séries Temporais com Valor Real

Seja $Y_1, \ldots, Y_t$ uma série temporal com observações pertencentes ao conjunto dos números reais. Definimos a função de ligação como a identidade, a função de variância

$V(\widetilde{\mu}_t) = 1$ e utilizamos a configuração para o fator latente AR(1) Gaussiano, sem a presença do intercepto. Temos a definição completa do modelo para séries temporais reais abaixo

$$\begin{aligned}\widetilde{\mu}_t &= x_{nt}^\top \beta + \alpha_t,\\ E(Y_t|\alpha_t) &= \widetilde{\mu}_t = x_{nt}^\top \beta + \alpha_t,\\ \text{Var}(Y_t|\alpha_t) &= \phi V(\widetilde{\mu}_t) = \phi,\end{aligned}$$

Como resultado imediato da configuração do fator latente sem intercepto, temos que $\alpha_t \sim N(0, \sigma^2)$. Portanto, podemos escrever a esperança e variância de Y_t na forma

$$E(Y_t) = E(E(Y_t|\alpha_t)) = E(x_{nt}^\top \beta + \alpha_t) = x_{nt}^\top \beta + E(\alpha_t) = x_{nt}^\top \beta = \mu_t,$$

e

$$\begin{aligned}Var(Y_t) &= E\left\{Var(Y_t|\alpha_t)\right\} + Var\left\{E(Y_t|\alpha_t)\right\}\\ &= E\left\{\phi\right\} + Var\left\{x_{nt}^\top \beta + \alpha_t\right\}\\ &= \phi + \sigma^2.\end{aligned}$$

Para a função de autocovâriancia, onde $k \neq 0$, temos que

$$\begin{aligned}Cov(Y_{t+k}, Y_t) &= Cov\left\{E(Y_{t+k}|\alpha_{t+k}), E(Y_t|\alpha_t)\right\}\\ &= Cov\left\{x_{n,t+k}^\top \beta + \alpha_{t+k}, x_{nt}^\top \beta + \alpha_t\right\}\\ &= \rho(k)\sigma^2.\end{aligned}$$

Por fim, a função de autocorrelação é dada na seguinte forma

$$Corr(Y_{t+k}, Y_t) = \frac{Cov(Y_{t+k}, Y_t)}{\sqrt{Var(Y_{t+k})Var(Y_t)}} = \frac{\rho(k)\sigma^2}{\sqrt{(\phi + \sigma^2)^2}} = \frac{\rho(k)}{\frac{\phi}{\sigma^2} + 1}.$$

3 Estimação dos Parâmetros

Neste capítulo apresentamos a metodologia utilizada para a estimação dos parâmetros do modelo. O vetor de parâmetros do modelo $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \sigma^2, \rho)^\top$ é constituído pelos coeficientes da regressão, o parâmetro de dispersão e a variância e autocorrelação do fator latente. Para a estimação dos coeficientes da regressão maximizamos o logaritmo da função de quase-verossimilhança, ignorando a presença do fator latente. Para os parâmetros do fator latente e o parâmetro de dispersão utilizamos o método dos momentos. Em alguns casos, o método dos momentos geram estimativas fora do espaço paramétrico. No entanto, este é um bom ponto para começar.

Wedderburn (1974) propôs que a estimação do parâmetro de dispersão ϕ fosse realizada separadamente da função de quase-verossimilhança. Assim, propôs a estimação através do método dos momentos, como apresentado na Eq. (2.1). Para o modelo proposto nesta dissertação o parâmetro ϕ não é bem estimado usando esta equação. Desta forma, procedemos utilizando a abordagem apresentada a seguir para estimação do parâmetro.

Como discutido em Davis et al. (2000) e Davis e Wu (2009), os coeficientes da regressão são bem estimados quando ignoramos a presença do fator latente na construção da função de verossimilhança, mas o mesmo não ocorre para os erros padrão das estimativas dos coeficientes. Desta forma, também observamos o mesmo comportamento quando trabalhamos com os modelos de quase-verossimilhança. Este comportamento será evidenciado no capítulo 4, onde observaremos que os erros padrão calculados pelo modelo de quase-verossimilhança (ignorando a presença do fator latente no modelo) é significativamente diferente dos desvios padrão empíricos calculados levando em consideração a presença do fator latente no modelo. Assim, apresentamos uma proposta para o cálculo dos erros padrão das estimativas. Utilizamos o método de Monte Carlo para o cálculo dos desvios padrão empíricos.

3.1 Séries Temporais Limitadas e Binárias

Para séries temporais com suporte limitado ou binário, supomos que $V(\mu_t) = \mu_t(1-\mu_t)$, onde $\mu_t = e^{-x_{nt}^\top \beta}$ ignorando a presença do fator latente. Para o caso onde as observações de Y_t estão definidas no intervalo $(0, 1)$, o logaritmo da função de quase-

verossimilhança é dado na forma

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu \frac{y-t}{V(t)}dt \\ &= \int_y^\mu \frac{y-t}{t(1-t)}dt \\ &= y\ln\left(\frac{1-y}{y}\right) + y\ln\left(\frac{\mu}{1-\mu}\right) + \ln\left(\frac{1-\mu}{1-y}\right). \end{aligned} \tag{3.1}$$

Para $Y_t = 0$ e $Y_t = 1$, obtemos respectivamente $Q(0,\mu) = \log(1-\mu)$ e $Q(1,\mu) = \log\mu$. Através da maximização do logaritmo da função de quase-verossimilhança obtemos o estimador $\widehat{\boldsymbol{\beta}}$ para o vetor de coeficientes da regressão.

Na estimação dos parâmetros ϕ, σ^2 e ρ utilizamos o método dos momentos. Definimos $\widehat{\mu}_t = e^{-x_{nt}^\top\widehat{\beta}}$. Igualando a variância amostral à variância teórica, dada em (2.4), temos

$$\sum_{t=1}^n (Y_t - \widehat{\mu}_t)^2 = \sum_{t=1}^n \widehat{\phi}\widehat{\mu}_t + \sum_{t=1}^n \widehat{\mu}_t\left\{(1-\phi)w(\sigma^2) - 1\right\}$$

em seguida, obtemos a expressão

$$\widehat{\phi} = \frac{\sum_{t=1}^n (Y_t - \widehat{\mu}_t)^2 - (w(\widehat{\sigma}^2) - 1)\sum_{t=1}^n \widehat{\mu}_t^2}{\sum_{t=1}^n \widehat{\mu}_t - w(\widehat{\sigma}^2)\sum_{t=1}^n \widehat{\mu}_t^2}.$$

Agora utilizando a expressão para a função da autocovariância dada em (2.5), igualamos a função de autocovariância amostral a função de autocovariância teórica. Assim, obtemos

$$\sum_{t=1}^n (Y_t - \widehat{\mu}_t)(Y_{t+k} - \widehat{\mu}_{t+k}) = \sum_{t=1}^n \widehat{\mu}_{t+k}\widehat{\mu}_t\left\{v(x,y) - 1\right\},$$

e consequentemente

$$v(\widehat{\sigma}^2, \widehat{\rho}^k) = \frac{\sum_{t=1}^{n=k}(Y_t - \widehat{\mu}_t)(Y_{t+k} - \widehat{\mu}_{t+k})}{\sum_{t=1}^{n-k}\widehat{\mu}_t\widehat{\mu}_{t+k}} + 1. \tag{3.2}$$

Temos que $w(x) = \left(\frac{(1+x)^2}{1+2x}\right)^{1/x}$ e $v(x,y) = \left(\frac{(1+x)^2}{1+2x+(x)^2(1-y)}\right)^{1/x}$. Na expressão em (3.2) definimos $k = 1, 2$, e assim, obtemos duas expressões. Desta forma, temos as três expressões para estimação dos três parâmetros.

3.2 Séries Temporais Não-negativas

Para séries temporais com suporte não-negativo, supomos que $V(\mu_t) = \mu_t^p$, onde $\mu_t = e^{x_{nt}^\top\beta}$ ignorando o fator latente. Para $p \neq 1, 2$, o vetor de parâmetros $\boldsymbol{\beta}$ é estimado através da maximização do logaritmo da função de quase-verossimilhança dado a seguir

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu \frac{y-t}{t^p}dt \\ &= \frac{y}{1-p}\left\{\mu^{-p+1} - y^{-p+1}\right\} - \frac{1}{2-p}\left\{\mu^{-p+2} - y^{-p+2}\right\}. \end{aligned} \tag{3.3}$$

Para $p = 1$ e $p = 2$, temos respectivamente $Q(y;\mu) = y(\log\mu - \log y) + y - \mu$ e $Q(y;\mu) = \log(y/\mu) - y/\mu + 1$. Os modelos de quase-verossimilhança com $p = 1, p = 2$ e $p = 3$ têm os casos correspondentes de modelos lineares generalizados Poisson, Gamma e Inversa Gaussiana.

Para estimação de ϕ, σ^2 e ρ pelo método dos momentos, utilizamos as expressões da $Var(Y_t)$ e $Cov(Y_{t+k}, Y_t)$. Igualamos as quantidades amostrais às respectivas quantidades teóricas. Primeiro, utilizando a $Var(Y_t)$ e definindo $\widehat{\mu}_t = e^{x_{nt}^\top \widehat{\beta}}$. Temos as expressões abaixo

$$\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2 = \sum_{t=1}^{n}\left[\widehat{\phi}\widehat{\mu}_t^p(\widehat{\sigma}_\epsilon^2 + 1)^{\frac{p^2-p}{2}} + \widehat{\mu}_t^2\widehat{\sigma}_\epsilon^2\right]$$

$$\widehat{\phi} = \frac{\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2 - (e^{\sigma^2} - 1)\sum_{t=1}^{n}\widehat{\mu}_t^2}{(e^{\sigma^2})^{\frac{p^2-p}{2}}\sum_{t=1}^{n}\widehat{\mu}_t^p}.$$

Agora, utilizando a expressão encontrada para a $Cov(Y_{t+k}, Y_t)$ e igualando-a à autocorrelação amostral, temos que

$$\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)(Y_{t+k} - \widehat{\mu}_{t+k}) = \widehat{\gamma}_\epsilon(k)\sum_{t=1}^{n}\mu_{t+k}\mu_t$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}^k) = \frac{\sum_{t=1}^{n-k}(Y_t - \widehat{\mu}_t)(Y_{t+k} - \widehat{\mu}_{t+k})}{\sum_{t=1}^{n-k}\widehat{\mu}_t\widehat{\mu}_{t+k}} + 1, \quad \text{para } k = 1, 2.$$

Para o caso em que $p = 1$, temos as expressões de estimação

$$\widehat{\phi} = \frac{\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2 - (e^{\sigma^2} - 1)\sum_{t=1}^{n}\widehat{\mu}_t^2}{\sum_{t=1}^{n}\widehat{\mu}_t},$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}) = \frac{\sum_{t=1}^{n-1}(Y_t - \widehat{\mu}_t)(Y_{t+1} - \widehat{\mu}_{t+1})}{\sum_{t=1}^{n-1}\widehat{\mu}_t\widehat{\mu}_{t+1}} + 1,$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}^2) = \frac{\sum_{t=1}^{n-2}(Y_t - \widehat{\mu}_t)(Y_{t+2} - \widehat{\mu}_{t+2})}{\sum_{t=1}^{n-2}\widehat{\mu}_t\widehat{\mu}_{t+2}} + 1.$$

Para $p = 2$, temos

$$\widehat{\phi} = \frac{\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2 - (e^{\sigma^2} - 1)\sum_{t=1}^{n}\widehat{\mu}_t^2}{e^{\sigma^2}\sum_{t=1}^{n}\widehat{\mu}_t^2},$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}) = \frac{\sum_{t=1}^{n-1}(Y_t - \widehat{\mu}_t)(Y_{t+1} - \widehat{\mu}_{t+1})}{\sum_{t=1}^{n-1}\widehat{\mu}_t\widehat{\mu}_{t+1}} + 1,$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}^2) = \frac{\sum_{t=1}^{n-2}(Y_t - \widehat{\mu}_t)(Y_{t+2} - \widehat{\mu}_{t+2})}{\sum_{t=1}^{n-2}\widehat{\mu}_t\widehat{\mu}_{t+2}} + 1.$$

Por fim, para $p = 3$ temos

$$\widehat{\phi} = \frac{\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)^2 - (e^{\sigma^2} - 1)\sum_{t=1}^{n}\widehat{\mu}_t^2}{(e^{\sigma^2})^3\sum_{t=1}^{n}\widehat{\mu}_t^3},$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}) = \frac{\sum_{t=1}^{n-1}(Y_t - \widehat{\mu}_t)(Y_{t+1} - \widehat{\mu}_{t+1})}{\sum_{t=1}^{n-1}\widehat{\mu}_t\widehat{\mu}_{t+1}} + 1,$$

$$\exp(\widehat{\sigma}^2\widehat{\rho}^2) = \frac{\sum_{t=1}^{n-2}(Y_t - \widehat{\mu}_t)(Y_{t+2} - \widehat{\mu}_{t+2})}{\sum_{t=1}^{n-2}\widehat{\mu}_t\widehat{\mu}_{t+2}} + 1.$$

3.3 Séries Temporais Reais

Para séries temporais com suporte real, supomos que $V(\mu_t) = 1$ e $\mu_t = x_{nt}^\top \beta$ ignorando a presença do fator latente. O logaritmo da função de quase-verossimilhança para estimação do vetor $\boldsymbol{\beta}$ é dado por

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu (y - t)dt \\ &= y\mu - \frac{\mu^2}{2} - \frac{y^2}{2}. \end{aligned}$$

Por meio da maximização do logaritmo da função de quase-verossimilhança obtemos os estimadores para os coeficientes da regressão. Pelo método dos momentos, igualando as quantidades amostrais as quantidades teóricas e utilizando as expressões da $Var(Y_t)$ e $Cov(Y_{t+k}, Y_t)$ temos, respectivamente,

$$\frac{1}{n}\sum_{t=1}^{n}(Y_t - x_{nt}^\top \widehat{\beta})^2 = \widehat{\phi} + \widehat{\sigma}^2$$

$$\widehat{\phi} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - x_{nt}^\top \widehat{\beta})^2 - \widehat{\sigma}^2,$$

e

$$\frac{1}{n}\sum_{t=1}^{n}(Y_t - \widehat{\mu}_t)(Y_{t+k} - \widehat{\mu}_{t+k}) = \widehat{\sigma}^2\widehat{\rho}^k. \tag{3.4}$$

Da expressão em (3.4) obtemos

$$\widehat{\rho} = \frac{\sum_{t=1}^{n-2}(Y_t - \widehat{\mu}_t)(Y_{t+2} - \widehat{\mu}_{t+2})}{\sum_{t=1}^{n-1}(Y_t - \widehat{\mu}_t)(Y_{t+1} - \widehat{\mu}_{t+1})},$$

e

$$\widehat{\sigma}^2 = \frac{\left(\sum_{t=1}^{n-1}(Y_t - \widehat{\mu}_t)(Y_{t+1} - \widehat{\mu}_{t+1})\right)^2}{n\sum_{t=1}^{n-2}(Y_t - \widehat{\mu}_t)(Y_{t+2} - \widehat{\mu}_{t+2})}.$$

3.4 Estudo de Simulação

Realizamos três estudos de simulação para avaliar a metodologia apresentada para estimação dos parâmetros do modelo com base na abordagem de quase-verossimilhança combinada com o método dos momentos. Todas as implementações neste documento foram realizadas através do software R Core Team (2019). Ilustramos aqui os casos positivos contínuos, com valor real e limitados. Para todos os casos considerados nesses estudos simulados, coletamos réplicas de Monte Carlo de tamanho 1000 e tamanhos de amostra $n = 500, 1000, 2000$.

3.4.1 Modelo de Séries Temporais Não-Negativas

Para o primeiro caso, supomos a distribuição condicional $Y_t|\alpha_t$ como sendo uma Gamma$(\tilde{\mu}_t, \nu)$, com média $\tilde{\mu}_t = e^{x_{nt}^\top \beta}\epsilon_t$ e variância $Var(Y_t|\alpha_t) = \phi\tilde{\mu}_t^2$, onde $\nu = 1/\phi$. O fator latente α_t será definido como um AR(1) Gaussiano, com $\sigma^2 = 0,5$e $\rho = 0,6$. Para a simulação, utilizamos as seguintes covariáveis $x_{nt} = \{1, cos(2\pi t/12), sin(2\pi t/12)\}$, coeficientes da regressão $\boldsymbol{\beta} = (5; -0,2; 0,4)^\top$ e $\phi = 0.1$. Na Figura 3 apresentamos a trajetória da série temporal e fator latente simulados para este cenário. Desta forma, estimamos os coeficientes da regressão maximizando o logaritmo da função de quase-verossimilhança, ignorando o fator latente como proposto na seção 3.1. Em seguida, estimamos o parâmetro de dispersão e parâmetros do fator latente por meio do método dos momentos.

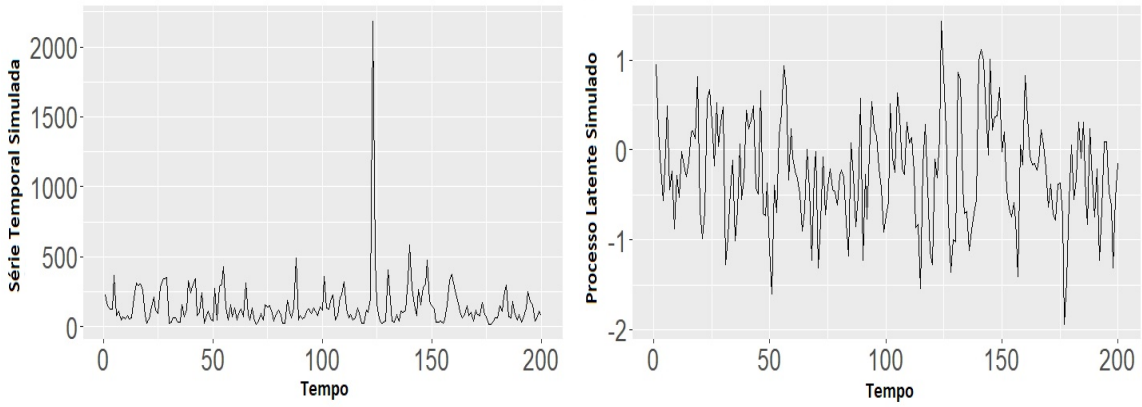

Figura 3 – Trajetórias simuladas de uma série temporal não-negativa (à esquerda) e associado processo latente (à direita) para o primeiro cenário.

Na Tabela 1, apresentamos as médias empíricas e os erros padrão das estimativas de quase-verossimilhança dos β's e as estimativas do método dos momentos (MMs) de ϕ, σ^2 e ρ com seus respectivos erros padrão. Chamamos atenção que os estimadores de MM podem produzir estimativas fora do espaço de parâmetros. Nesses casos, as amostras foram descartadas e uma nova réplica de Monte Carlo foi considerada. Esse é um problema bem conhecido desse tipo de estimador e é atenuado ao trabalhar com tamanhos de amostra grandes ou moderados.

Na Tabela 1, observamos que os estimadores de quase-verossimilhança produziram estimativas aproximadamente não-viesadas do vetor β para todos os tamanhos de amostra considerados. Os estimadores de método dos momentos também forneceram resultados satisfatórios para estimar ϕ, σ^2 e ρ. Esses comentários também são suportados na Figura 4, onde são exibidos os boxplots das estimativas dos parâmetros. A partir desses gráficos, observamos um bom desempenho geral e consistência dos estimadores propostos à medida que o tamanho da amostra aumenta.

Tabela 1 – Médias empíricas e erros padrão das estimativas de quase-verossimilhança de β e método de momentos estimado de ϕ, σ^2 e ρ com base no modelo de séries temporais semiparamétricos para dados contínuos positivos.

	valor	$n = 500$		$n = 1000$		$n = 2000$	
parâmetro	verdadeiro	média	ep	média	ep	média	ep
β_0	5	4,997	0,070	4,998	0,049	4,997	0,035
β_1	$-0{,}2$	$-0{,}199$	0,076	$-0{,}202$	0,054	$-0{,}200$	0,037
β_2	0,4	0,394	0,074	0,398	0,053	0,401	0,039
ϕ	0,1	0,131	0,089	0,115	0,071	0,107	0,059
σ^2	0,5	0,448	0,107	0,475	0,086	0,487	0,058
ρ	0,6	0,626	0,101	0,615	0,075	0,603	0,102

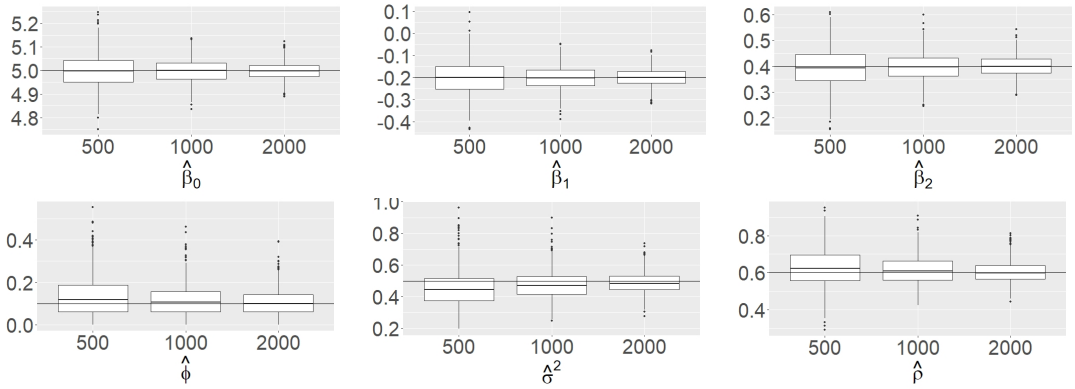

Figura 4 – Boxplots das estimativas de parâmetros com base no modelo de séries temporais semiparamétricos para dados contínuos positivos.

3.4.2 Modelo de Séries Temporais Reais

Agora no segundo cenário considerado, supomos a distribuição condicional $Y_t|\alpha_t$ como sendo uma Normal$(\tilde{\mu}_t, \phi)$, com média $\tilde{\mu}_t = x_{nt}^\top \beta + \alpha_t$ e variância $Var(Y_t|\alpha_t) = \phi$. Utilizamos $\phi = 3$. O fator latente ϵ_t será definido como um AR(1) Gaussiano, com $\sigma^2 = 1$ e $\rho = 0,5$. As covariáveis utilizadas são $x_{nt} = \{1, t/n, cos(2\pi t/6)\}$ e os coeficientes da regressão $\boldsymbol{\beta} = (0,1; 0,5; 0,7)^\top$. Na Figura 5 temos a trajetória simulada da série temporal e do processo latente. Desta forma, estimamos os coeficientes da regressão maximizando o logaritmo da função de quase-verossimilhança, ignorando o fator latente. Em seguida, estimamos o parâmetro de dispersão e parâmetros do fator latente por meio do método dos momentos. As duas metodologias são apresentadas na seção 3.3.

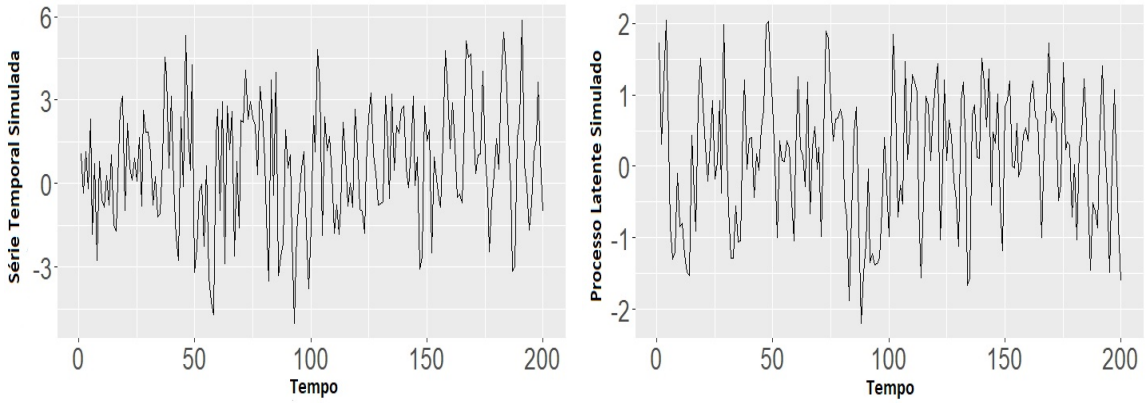

Figura 5 – Trajetórias simuladas de uma série temporal com valor real (à esquerda) e associado processo latente (à direita) para o segundo cenário.

As médias empíricas e erros padrão dos parâmetros do modelo são apresentados na Tabela 2. Os boxplots dessas estimativas, obtidos através da simulação de Monte Carlo, são apresentados na Figura 6. A partir desses resultados, podemos observar um bom desempenho dos estimadores propostos com base na abordagem de quase-verossimilhança combinada com o método de momentos para as séries temporais com valor real considerado.

Tabela 2 – Médias empíricas e erros padrão das estimativas de quase-verossimilhança de β e método de momentos estimado de ϕ, σ^2 e ρ com base no modelo de séries temporais semiparamétricos para dados reais.

	valor	$n = 500$		$n = 1000$		$n = 2000$	
parâmetro	verdadeiro	média	ep	média	ep	média	ep
β_0	0,1	0,106	0,218	0,100	0,152	0,096	0,109
β_1	0,5	0,496	0,382	0,501	0,267	0,502	0,192
β_2	0,7	0,696	0,126	0,697	0,086	0,699	0,060
ϕ	3	2,700	0,810	2,813	0,686	2,832	0,560
σ^2	1	1,280	0,800	1,184	0,685	1,157	0,555
ρ	0,5	0,519	0,230	0,516	0,203	0,499	0,174

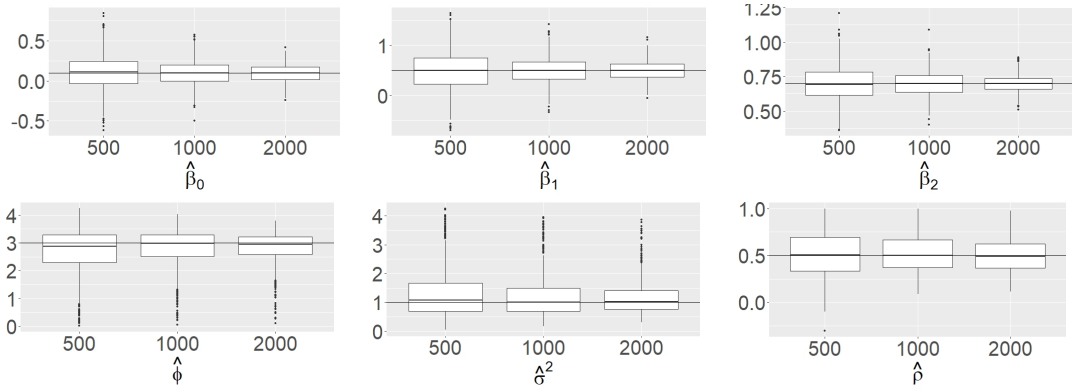

Figura 6 – Boxplots das estimativas de parâmetros com base no modelo de séries temporais semiparamétricos para dados reais.

3.4.3 Modelo de Séries Temporais Limitadas

Nosso último cenário é sobre séries temporais limitadas no intervalo $(0, 1)$. Supomos a distribuição condicional $Y_t|\alpha_t$ como sendo uma Beta$(\tilde{\mu}_t, \lambda)$, com média $\tilde{\mu}_t = e^{-x_{nt}^\top \beta}\epsilon_t$ e variância $\phi\tilde{\mu}_t(1-\tilde{\mu}_t)$, onde $\phi = 1/(1+\lambda)$. O fator latente considerado para este cenário será o processo Gamma deslocado. Desta forma, utilizamos os seguintes valores para os parâmetros do fator latente, $\sigma^2 = 0,3$ e $\rho = 0,8$, o parâmetro de dispersão $\phi = 0,1$, e as seguintes covariáveis no modelo $x_{nt} = \{1, t/n, (t/n)^2\}^\top, t = 1, \ldots, n$, com coeficientes de regressão associados $\beta = (1; 0,3; 0,5)^\top$. Na Figura 7 apresentamos a trajetória simulada da série temporal e processo latente para este caso.

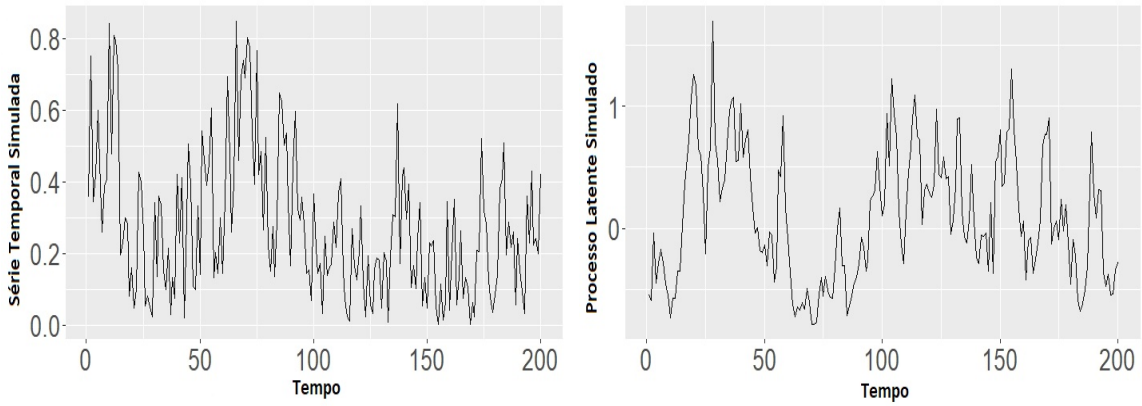

Figura 7 – Trajetórias simuladas de uma série temporal limitada (à esquerda) e associado processo latente (à direita) para este cenário.

Na Tabela 3, apresentamos as médias empíricas e os erros padrão das estimativas de quase-verossimilhança, bem como as estimativas pelo método dos momentos. Nesse caso, podemos observar um viés considerável nas estimativas de quase-verossimilhança para os β's, especialmente para $n = 500$. Por outro lado, vemos um bom desempenho do método dos estimadores de momentos para os parâmetros ϕ, σ^2 e ρ. Essa dificuldade em

estimar os coeficientes de regressão foi relatada por Davis & Wu (2009) em um cenário semelhante. Os autores consideraram um modelo de série temporal binária conduzido por um processo latente exponencial. Nos resultados simulados desse trabalho, assume-se apenas uma interceptação para a média e uma abordagem GLM é considerada para estimar, o que produziu estimativas com considerável viés.

A figura 8 mostra os gráficos boxplots das estimativas dos parâmetros para o caso de série temporal limitada. A partir desses gráficos, temos evidências empíricas de que os estimadores propostos são consistentes para o cenário considerado aqui, mesmo para os estimadores de quase-verossimilhança dos β's.

Tabela 3 – Médias empíricas e erros padrão das estimativas de quase-verossimilhança de β e método de momentos estimado de ϕ, σ^2 e ρ com base no modelo de séries temporais semiparamétricos para dados limitados.

	valor	$n = 500$		$n = 1000$		$n = 2000$	
parâmetro	verdadeiro	média	ep	média	ep	média	ep
β_0	1	0,932	0,174	0,965	0,128	0,989	0,090
β_1	0,3	0,616	0,867	0,437	0,608	0,349	0,429
β_2	0,5	0,228	0,869	0,384	0,595	0,459	0,423
ϕ	0,1	0,096	0,018	0,099	0,012	0,099	0,009
σ^2	0,3	0,333	0,201	0,301	0,101	0,306	0,069
ρ	0,8	0,773	0,107	0,788	0,079	0,792	0,054

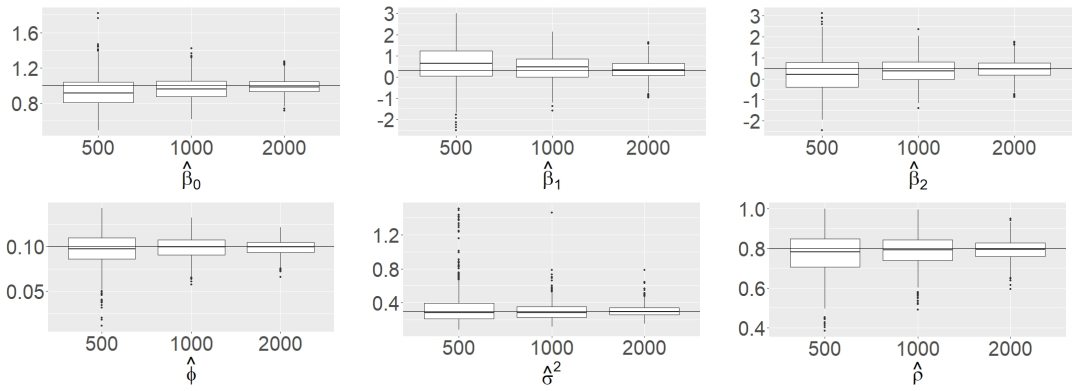

Figura 8 – Boxplots das estimativas de parâmetros com base no modelo de séries temporais semiparamétricos para dados limitados.

4 Aplicações

Neste capítulo, aplicamos os modelos de séries temporais semiparamétricos propostos para analisar séries temporais reais sobre taxa de desemprego e insolação. Ilustramos o desempenho de nossos modelos de séries temporais semiparamétricos em dados limitados e não-negativos.

4.1 Análise de dados de insolação

Nesta seção consideramos os dados temporais reais de insolação total (em horas) mensal da cidade de Belo Horizonte, no estado de Minas Gerais. O período analisado é de Janeiro de 1961 a Janeiro de 2019, totalizando 616 observações. Os dados foram obtidos no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP). Na Figura 9, temos o gráfico da série temporal (à esquerda) e sua respectiva função de autocorrelação (à direita).

Segundo o Núcleo Geoambiental da Universidade Estadual do Maranhão, a insolação é, pois, o intervalo total de tempo (entre o nascimento e o por do sol) em que o disco solar não esteve oculto por nuvens ou fenômenos atmosféricos de qualquer natureza. Desta forma, temos que nossos dados representam o intervalo de tempo total mensal de insolação na cidade de Belo Horizonte. O estudo do período de insolação é importante para atividades turísticas e agrícolas, como também na analise de mecanismos para melhor aproveitá-lo.

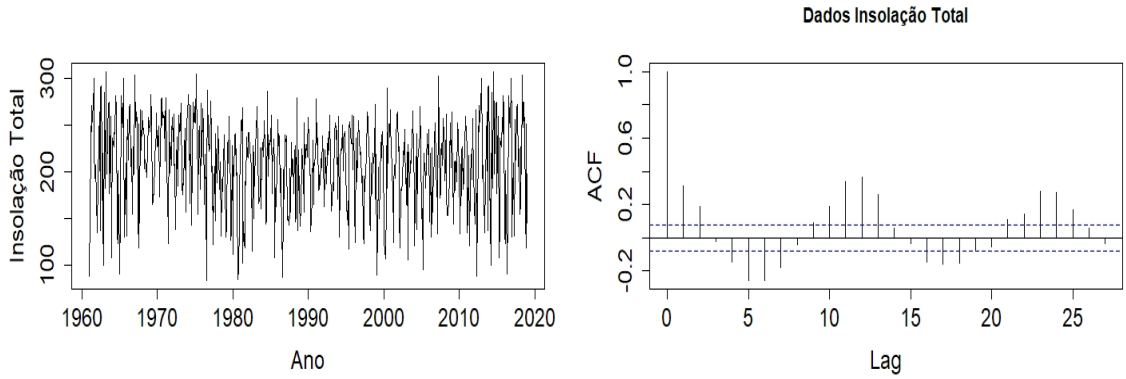

Figura 9 – Gráficos da insolação total mensal da cidade de Belo Horizonte de Janeiro de 1961 a Janeiro de 2019 (à esquerda) e ACF associado (à direita).

Os dados classificam-se como contínuos não-negativos. Assim, utilizamos na aplica-

ção o seguinte Modelo de Séries Temporais Não-negativas definido na subseção 2.2.2

$$\begin{aligned}\log \tilde{\mu}_t &= x_{nt}^\top \beta + \alpha_t,\\ E(Y_t|\alpha_t) = \tilde{\mu}_t &= \exp(x_{nt}^\top \beta + \alpha_t) = \exp(x_{nt}^\top \beta)\epsilon_t,\\ \mathrm{Var}(Y_t|\alpha_t) &= \phi V(\tilde{\mu}_t) = \phi \tilde{\mu}_t^p,\end{aligned}$$

em que $\{\alpha_t\}$ é assumido ser o fator latente, definido na forma $\alpha_t = c + \rho\alpha_{t-1} + \eta_t$, com distribuição $N(-\sigma^2/2, \sigma^2)$ e $\eta_t \sim N(0, \sigma_\eta^2)$.

Especificamos as seguintes covariáveis

$$x_{nt} = \{1, \cos(2\pi t/12), \cos(2\pi t/6), \cos(2\pi t/3)\}^\top,$$

para $t = 1, \ldots, 616$. A componente $cos(2\pi t/12)$ lida com uma possível sazonalidade anual da série temporal, enquanto a componente $cos(2\pi t/6)$ e $cos(2\pi t/3)$ lidam com a sazonalidade semestral e trimestral, respectivamente. Nesta aplicação não definimos uma componente de tendência, de acordo com a análise do gráfico à esquerda da Figura 9, portanto não tendo necessidade da inclusão da componente.

Nesta aplicação trabalhamos com a função de variância $V(\tilde{\mu}_t) = \tilde{\mu}_t^2 (p = 2)$. Na estimação dos coeficientes da regressão maximizamos o logaritmo da função de quase-verossimilhança dada por (3.3), em que $p = 2$. Obtemos as estimativas dos coeficientes e respectivos erros padrão, por meio da função de quase-verossimihança. Por intermédio do método dos momentos obtemos as estimativas para ϕ e parâmetros do fator latente. Os resultados são apresentados na Tabela 4.

Tabela 4 – Estimativas dos parâmetros e respectivos erros padrão do modelo de série temporal semiparamétrico não-negativo (com $p = 2$) para os dados de insolação total.

	Quase+MM		Simulação	
covariáveis/par.	estimativas	erro padrão	estimativas	erro padrão
Intercepto	5,335	0,008	5,335	0,001
$\cos(2\pi t/12)$	0,046	0,012	0,046	0,015
$\cos(2\pi t/6)$	−0,008	0,012	−0,007	0,011
$\cos(2\pi t/3)$	−0,021	0,012	−0,021	0,010
ϕ	0,021	–	0,020	0,006
σ^2	0,022	–	0,022	0,006
ρ	0,600	–	0,606	0,121

Para verificação dos resultados obtidos, realizamos uma simulação pelo método de Monte Carlo com 1000 replicações. Simulamos uma série temporal de tamanho 616 de uma distribuição Gamma, com média igual a $\tilde{\mu}_t = e^{x_{nt}^\top \beta}\epsilon_t$ e variância $\phi\tilde{\mu}_t^2$. Seja $\{\alpha_t\}$ o fator

latente assumindo ser um AR(1) Gaussiano com média $-\sigma^2/2$ e variância σ^2. Utilizamos $\widehat{\boldsymbol{\beta}}, \widehat{\phi}, \widehat{\sigma}^2$ e $\widehat{\rho}$ como os valores dos parâmetros para a simulação.

As médias e desvios padrão empíricos são apresentados na Tabela 4. Os procedimentos de quase-verossimilhança e método de momentos fornecem estimativas semelhantes ao método de Monte Carlo, especialmente para estimar os β's. Usando um nível de significância de 5% e levando em consideração o processo latente, as covariáveis $\cos(2\pi t/12)$ e $\cos(2\pi t/3)$ foram significativas. Essas covariáveis correspondem à sazonalidade anual e trimestral. Por outro lado, ignorando a presença do processo latente, somente a sazonalidade anual é significativa. Isso mostra a importância de considerar uma especificação de modelo adequada; caso contrário, a inferência pode ser comprometida.

Nas figuras 10 e 11, apresentamos os histogramas e qq plots das estimativas padronizadas de quase-verossimilhança dos β's, respectivamente. Esses gráficos indicam novamente uma aproximação normal satisfatória para a distribuição dos estimadores de quase-verossimilhança. Isso está de acordo com nossos resultados simulados fornecidos na Seção 3.4.

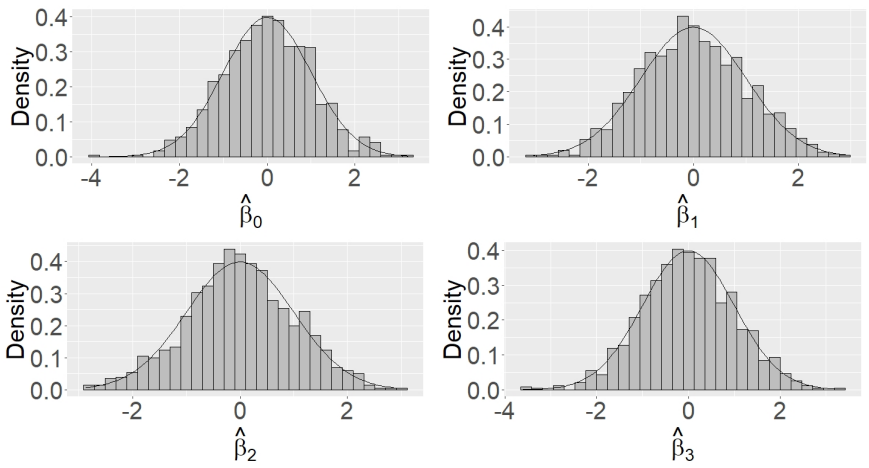

Figura 10 – Histogramas das estimativas padronizadas de quase-verossimilhança dos β's para os dados de insolação total.

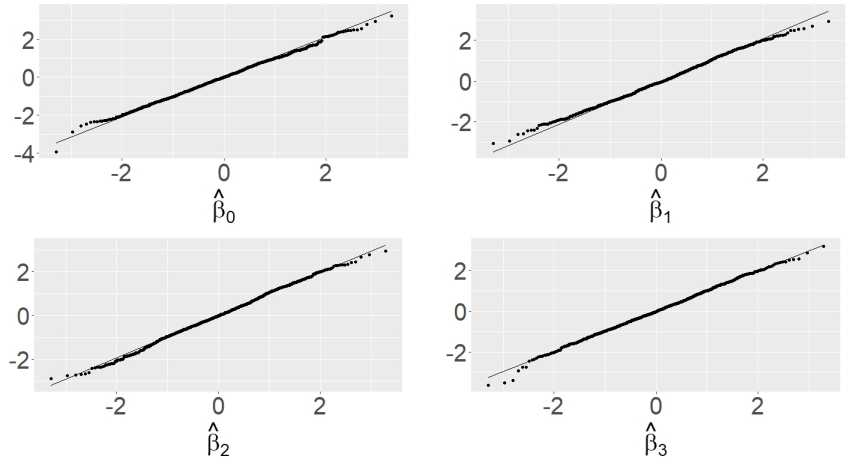

Figura 11 – QQ-plots das estimativas padronizadas de quase-verossimilhança dos β's para os dados de insolação total.

A figura 12 apresenta predições de intervalos um passo à frente para a insolação total de Belo Horizonte de fevereiro de 2019 a novembro de 2019. Os intervalos foram construídos simulando trajetórias (como foi feito para calcular os erros padrão) com base nas estimativas dos parâmetros sobre as séries temporais observadas e os quantis 97,5 % e 2,5 %. Observamos que nossa metodologia proposta fornece uma previsão satisfatória.

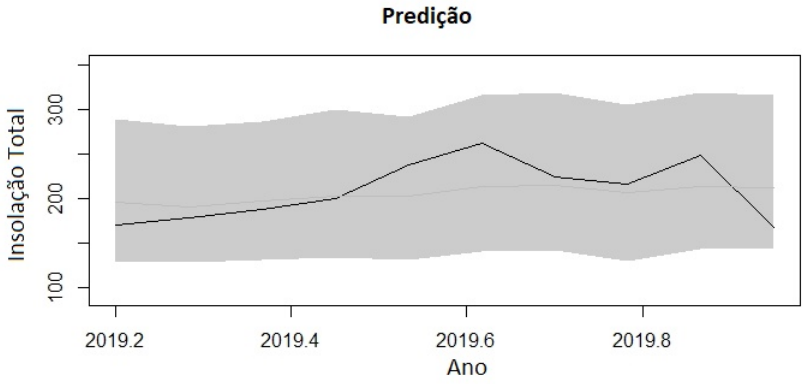

Figura 12 – Predição de um passo à frente para o total de dados de insolação.

4.2 Análise de dados de taxa de desemprego

Os dados consistem na taxa mensal de desemprego do período de março do ano de 2002 a abril de 2015 da cidade de Recife, capital do estado de Pernambuco. Os dados foram obtidos por meio da plataforma de consulta do Instituto de Econômia Aplicada (Ipea). À direita da Figura 13, apresentamos o gráfico da série temporal e à esquerda a respectiva função de autocorrelação da série temporal.

Para obter a taxa de desemprego dois outros indicadores são necessários. O número de pessoas desocupadas e a força de trabalho. Segundo definição da Pesquisa Mensal de Emprego (PME), são classificadas como desocupadas na semana de referência as pessoas com 10 anos ou mais de idade sem trabalho nessa semana, que tomaram alguma providência efetiva para consegui-lo no período de referência de 30 dias e que estavam disponíveis para assumi-lo na semana de referência. Pessoas na força de trabalho na semana de referência compreendem as pessoas ocupadas e as pessoas desocupadas nesse período. Portanto, a taxa de desemprego é o percentual de pessoas desocupadas em relação as pessoas na força de trabalho: [desocupados/força de trabalho] x 100.

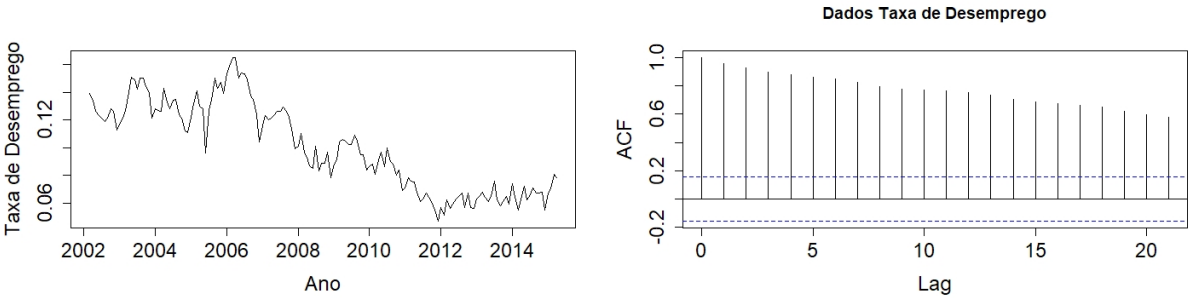

Figura 13 – Gráficos da taxa de desemprego mensal da cidade de Recife de Março de 2002 a Abril de 2015 (à esquerda) e ACF associado (à direita).

Os dados temporais são definidos no intervalo $(0, 1)$. Desta forma, aplicamos os dados no modelo definido na subseção 2.2.1, Modelo para Séries Temporais Limitadas. Abaixo as definições do modelo

$$\begin{aligned} -\log \tilde{\mu}_t &= x_{nt}^\top \beta + \alpha_t, \\ E(Y_t|\alpha_t) &= \tilde{\mu}_t = \exp(-x_{nt}^\top \beta)\epsilon_t, \\ \text{Var}(Y_t|\alpha_t) &= \phi V(\tilde{\mu}_t) = \phi\tilde{\mu}_t(1-\tilde{\mu}_t), \end{aligned}$$

em que processo latente $\{\alpha_t\}$ é assumido ser um processo Gamma deslocado. As covariáveis utilizadas serão as seguintes

$$x_{nt} = \{1, |t-118|/158, \cos(2\pi t/6), \sin(2\pi t/6)\}^\top, \quad t = 1, \ldots, 158.$$

para $t = 1, \ldots, 158$. As componentes $cos(2\pi t/6)$ e $sin(2\pi t/6)$ capturam a sazonalidade semestral da série temporal. Como observado no gráfico à esquerda da Figura 13, a série apresenta duas inclinações significativas, onde o valor 118 é definido como o ponto de mudança no comportamento da série de desemprego de Recife. A partir da observação 118 a série inicia uma tendência de crescimento na taxa, diferente de seu comportamento anterior ao ponto 118, que era de queda. Devido a este comportamento acrescentamos a

componente $|t-118|/158$ no modelo. O ponto 118 corresponde ao mês de dezembro do ano de 2011.

Os coeficientes da regressão são estimados pela maximização do logaritmo da função de quase-verossimilhança, ignorando o fator latente, dada em (3.1). As estimativas para ϕ e os parâmetros do fator latente são obtidas pelo método dos momentos. Os resultados estão presentes na Tabela 5.

Tabela 5 – Estimativas dos parâmetros e respectivos erros padrão do modelo de série temporal semiparamétrico limitado para os dados da taxa de desemprego.

	Quase+MM		Simulação	
covariáveis/par.	estimativas	erro padrão	estimativas	erro padrão
Intercepto	2,683	0,029	2,899	0,144
$\lvert t-118\rvert/158$	$-1{,}142$	0,066	$-2{,}210$	0,245
$\cos(2\pi t/6)$	$-0{,}016$	0,021	$-0{,}009$	0,007
$\sin(2\pi t/6)$	$-0{,}026$	0,021	$-0{,}040$	0,008
ϕ	$1,9 \cdot 10^{-4}$	–	$2,2 \cdot 10^{-4}$	$1,2 \cdot 10^{-4}$
σ^2	0,034	–	0,042	0,021
ρ	0,952	–	0,910	0,031

Realizamos uma simulação pelo método de Monte Carlo com 1000 reaplicações. Simulamos uma série temporal de tamanho 158 de uma distribuição Beta, com média igual a $\widetilde{\mu}_t = e^{-x_{nt}^\top \beta} e^{-\alpha_t}$ e variância $\lambda\widetilde{\mu}_t(1-\widetilde{\mu}_t)$, onde $\lambda = \frac{1-\phi}{\phi}$. O fator latente $\{\alpha_t\}$ é assumido ser um processo Gamma deslocado. Utilizamos $\widehat{\boldsymbol{\beta}}, \widehat{\phi}, \widehat{\sigma}^2$ e $\widehat{\rho}$ como os valores dos parâmetros para a simulação.

As médias e desvios padrão empíricos estão nas duas últimas colunas da Tabela 5. Como pode ser visto, há uma enorme diferença entre os erros padrão baseados na simulação de Monte Carlo (considerando a presença do processo latente) e os erros da abordagem de quase-verossimilhança. Isso também é bem discutido nos artigos de Davis et al. (2000) e Davis & Wu (2009) , onde é considerada uma abordagem de modelo linear generalizado.

Na Tabela 5, também é possível observar uma boa concordância das estimativas entre quase-verossimilhança e método de momentos e aquelas capturados na simulação de Monte Carlo, com exceção do coeficiente de tendência. Tivemos esse mesmo problema em nossos resultados simulados no capítulo anterior. Quanto a significância das covariáveis, quando consideramos o fator latente no modelo todas as covariáveis são significativas, porém, pelo modelo de quase-verossimilhança (onde o fator latente é ignorado) somente a covariável $|t-118|/158$ é significativa. Reforçando a importância de especificar corretamente o modelo.

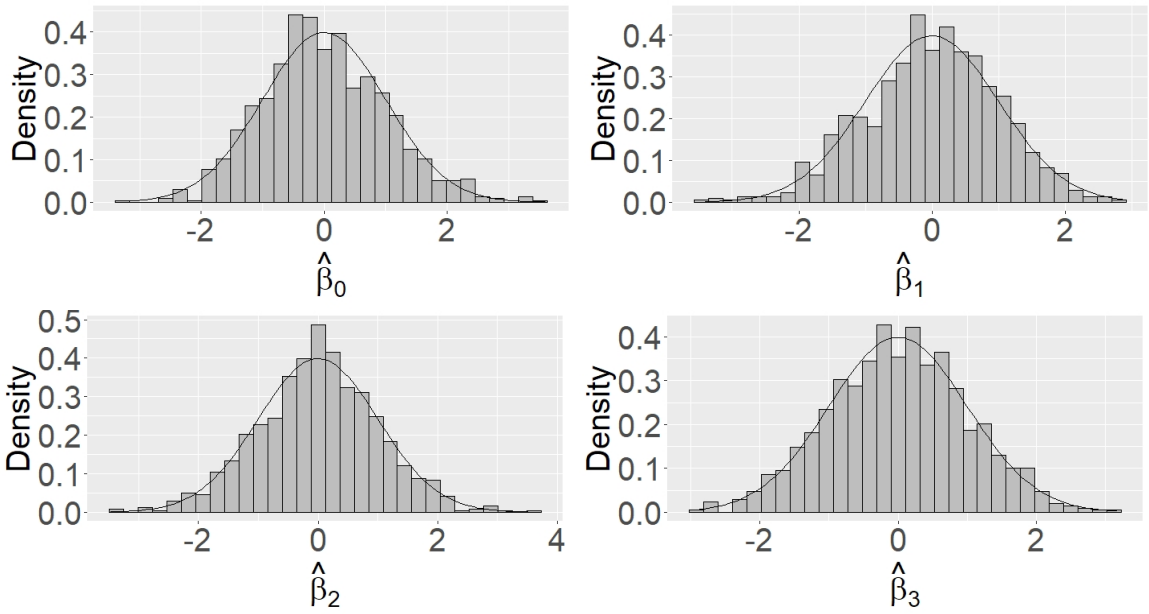

Figura 14 – Histogramas das estimativas padronizadas de quase-verossimilhança dos β's para os dados de desemprego.

As figuras 14 e 15, respectivamente, mostram os histogramas e qq plots das estimativas padronizadas de Monte Carlo dos β's, que indicam aproximações normais satisfatórias.

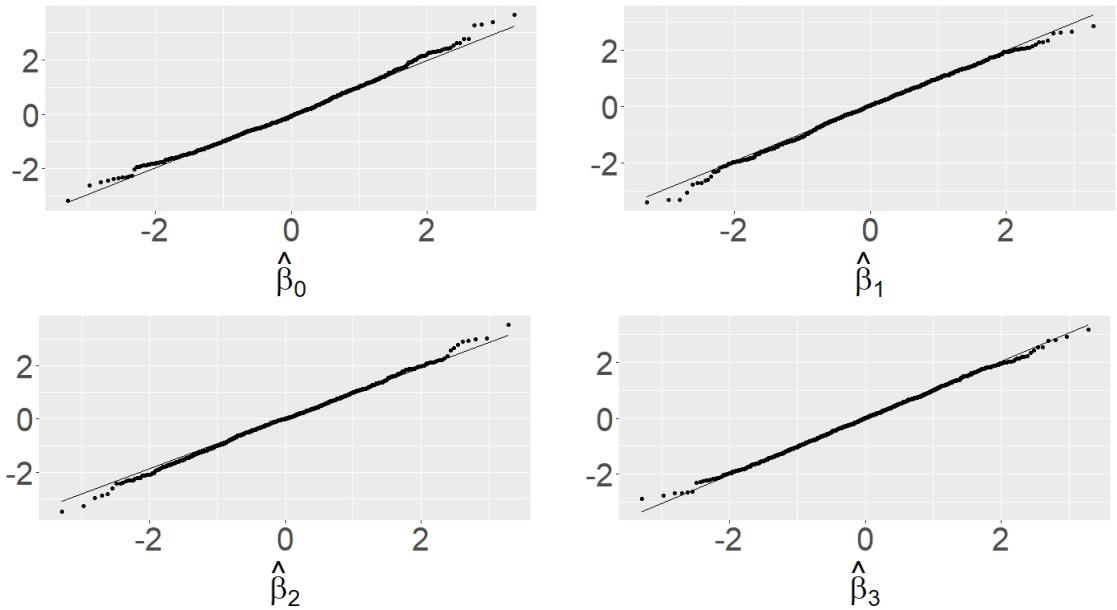

Figura 15 – QQ-plots das estimativas padronizadas de quase-verossimilhança dos β's para os dados de desemprego.

Na Figura 16, fornecemos predições de intervalos um passo à frente para a taxa de desemprego de Recife de maio de 2015 a fevereiro de 2016. Os intervalos foram construídos

simulando trajetórias com base nas estimativas dos parâmetros nas séries temporais observadas e tomando os quantis 97.5% e 2,5%. Como podemos observar, nosso modelo fornece uma previsão satisfatória.

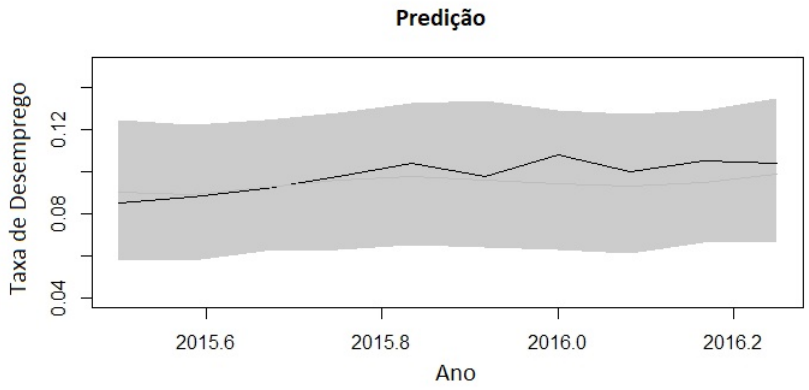

Figura 16 – Predição de um passo à frente para os dados de desemprego.

5 Conclusão

Uma classe flexível de modelos de séries temporais semiparamétricos foi proposta assumindo um modelo de quase-verossimilhança conduzido por um processo latente. Nossa metodologia proposta é capaz de lidar com séries temporais não-negativas contínuas, contagens, limitadas, binárias e com valores reais. Inferência sobre os parâmetros do modelo foi discutida e simulações de Monte Carlo foram abordadas para verificar o desempenho das estimativas. Aplicações em dados de séries temporais de taxas de desemprego e insolação total ilustraram a utilidade da metodologia proposta em situações práticas.

Um ponto desafiador parece ser a estimativa dos parâmetros relacionados à média para o caso limitado, onde um viés considerável foi observado, o que também foi observado por Davis & Wu (2009) em um modelo de série temporal binário. Uma solução possível pode ser usar um procedimento de Bootstrap (Efron & Tibshirani, 1994) para obter o viés e, em seguida, corrigir as estimativas de quase-verossimilhança.

Outro ponto que gostaríamos de chamar atenção é que outras formas para a função de variância podem ser consideradas e os resultados discutidos neste artigo podem ser facilmente adaptados. Por exemplo, no caso limitado, pode-se estar interessado em considerar a função de variância $V(\mu) = \mu^3(1-\mu)^3$, com $\mu \in (0,1)$. Os momentos marginais e a função de autocorrelação para este caso são obtidos seguindo as mesmas etapas fornecidas na Subseção 2.2.1.

Outros pontos que acreditamos que merecem ser investigados em pesquisas futuras são: (i) estudo mais aprofundado sobre predição; (ii) ferramentas de diagnóstico e (iii) extensão multivariada.

Referências

BARRETO-SOUZA, W.; OMBAO, H. Negative binomial process: A tractable model with composite likelihood-based inference. *Submitted for publication*, 2019. Citado na página 17.

BENJAMIN, M. A.; RIGBY, R. A.; STASINOPOULOS, D. M. Generalized autoregressive moving average models. *Journal of the American Statistical association*, Taylor & Francis, v. 98, n. 461, p. 214–223, 2003. Citado na página 8.

BERKES, I.; HORVÁTH, L. The rate of consistency of the quasi-maximum likelihood estimator. *Statistics & probability letters*, Elsevier, v. 61, n. 2, p. 133–143, 2003. Citado na página 10.

BOX, G. E.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964. Citado na página 11.

BOX GEORGE; JENKINS, G. *Time Series Analysis: Forecasting and Control*. [S.l.]: San Francisco: Holden-Day, 1970. Citado na página 8.

BRÄNNÄS, K.; JOHANSSON, P. Time series count data regression. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 23, n. 10, p. 2907–2925, 1994. Citado 2 vezes nas páginas 9 e 10.

BRILLINGER, D. R. *Time series: data analysis and theory*. [S.l.]: Siam, 1981. v. 36. Citado na página 8.

BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods: Theory and Methods*. [S.l.]: Springer Science & Business Media, 1991. Citado na página 8.

CHAN, K.; LEDOLTER, J. Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, Taylor & Francis, v. 90, n. 429, p. 242–252, 1995. Citado na página 9.

CHRISTOU, V.; FOKIANOS, K. Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, Wiley Online Library, v. 35, n. 1, p. 55–78, 2014. Citado na página 10.

CHRISTOU, V.; FOKIANOS, K. et al. Estimation and testing linearity for non-linear mixed poisson autoregressions. *Electronic Journal of Statistics*, The Institute of Mathematical Statistics and the Bernoulli Society, v. 9, n. 1, p. 1357–1377, 2015. Citado na página 10.

COX, D. R. Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, JSTOR, p. 93–115, 1981. Citado 2 vezes nas páginas 8 e 9.

DAVIS, R. A.; DUNSMUIR, W. T.; STREETT, S. B. Observation-driven models for poisson counts. *Biometrika*, Oxford University Press, v. 90, n. 4, p. 777–790, 2003. Citado na página 8.

DAVIS, R. A.; DUNSMUIR, W. T.; WANG, Y. Modeling time series of count data. *Statistics Textbooks and Monographs*, MARCEL DEKKER AG, v. 158, p. 63–114, 1999. Citado na página 8.

DAVIS, R. A.; DUNSMUIR, W. T.; WANG, Y. On autocorrelation in a poisson regression model. *Biometrika*, Oxford University Press, v. 87, n. 3, p. 491–505, 2000. Citado 6 vezes nas páginas 8, 10, 14, 15, 22 e 36.

DAVIS, R. A.; RODRIGUEZ-YAM, G. Estimation for state-space models: an approximate likelihood approach. *Statistica Sinica*, Citeseer, v. 15, n. 381-406, p. 7, 2005. Citado na página 9.

DAVIS, R. A.; WU, R. A negative binomial model for time series of counts. *Biometrika*, Oxford University Press, v. 96, n. 3, p. 735–749, 2009. Citado 10 vezes nas páginas 8, 10, 14, 15, 17, 19, 22, 30, 36 e 39.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994. Citado na página 39.

FAHRMEIR, L.; TUTZ, G. Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 89, n. 428, p. 1438–1449, 1994. Citado na página 9.

FOKIANOS, K.; RAHBEK, A.; TJØSTHEIM, D. Poisson autoregression. *Journal of the American Statistical Association*, Taylor & Francis, v. 104, n. 488, p. 1430–1439, 2009. Citado na página 10.

FRANCQ, C.; ZAKOIAN, J.-M. et al. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, Bernoulli Society for Mathematical Statistics and Probability, v. 10, n. 4, p. 605–637, 2004. Citado na página 10.

GRUNWALD, G. K.; RAFTERY, A. E.; GUTTORP, P. Time series of continuous proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 55, n. 1, p. 103–116, 1993. Citado na página 8.

HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton university press Princeton, NJ, 1994. v. 2. Citado na página 8.

HANNAN, E. Non-linear time series regression. *Journal of Applied Probability*, Cambridge University Press, v. 8, n. 4, p. 767–780, 1971. Citado na página 8.

HEYDE, C. C. *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. [S.l.]: Springer Science & Business Media, 1997. Citado na página 10.

JØRGENSEN, B. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 49, n. 2, p. 127–145, 1987. Citado na página 9.

JØRGENSEN, B. et al. A state space model for multivariate longitudinal count data. University of British Columbia, v. 86, n. 1, p. 169–181, 1995. Citado na página 9.

JØRGENSEN, B. et al. State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics*, Wiley Online Library, v. 24, n. 3, p. 385–402, 1996. Citado na página 9.

JØSRGENSEN, B.; SONG, P. X.-K. Stationary state space models for longitudinal data. *Canadian Journal of Statistics*, Wiley Online Library, v. 35, n. 4, p. 461–483, 2007. Citado 3 vezes nas páginas 9, 10 e 16.

KALMAN, R. E. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, American Society of Mechanical Engineers, v. 82, n. 1, p. 35–45, 1960. Citado na página 16.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986. Citado 2 vezes nas páginas 8 e 10.

MCCABE, B. P.; MARTIN, G. M. Bayesian predictions of low count time series. *International Journal of Forecasting*, Elsevier, v. 21, n. 2, p. 315–330, 2005. Citado na página 8.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado na página 11.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado 2 vezes nas páginas 11 e 13.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>. Citado na página 25.

ROCHA, A. V.; CRIBARI-NETO, F. Beta autoregressive moving average models. *Test*, Springer, v. 18, n. 3, p. 529, 2009. Citado na página 8.

SHUMWAY, R. H.; STOFFER, D. S. *Time series analysis and its applications: with R examples*. [S.l.]: Springer, 2017. Citado na página 8.

SIM, C. H. First-order autoregressive models for gamma and exponential processes. *Journal of Applied Probability*, Applied Probability Trust, v. 27, n. 2, p. 325–332, 1990. ISSN 00219002. Disponível em: <http://www.jstor.org/stable/3214651>. Citado 2 vezes nas páginas 17 e 18.

STRAUMANN, D.; MIKOSCH, T. et al. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 34, n. 5, p. 2449–2495, 2006. Citado na página 10.

TERUI, N.; DIJK, H. K. V. Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, Elsevier, v. 18, n. 3, p. 421–438, 2002. Citado na página 8.

WEDDERBURN, R. W. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, Oxford University Press, v. 61, n. 3, p. 439–447, 1974. Citado 2 vezes nas páginas 11 e 22.

ZEGER, S. L. A regression model for time series of counts. *Biometrika*, Oxford University Press, v. 75, n. 4, p. 621–629, 1988. Citado 3 vezes nas páginas 8, 9 e 10.

ZEGER, S. L.; QAQISH, B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, JSTOR, p. 1019–1031, 1988. Citado 3 vezes nas páginas 8, 10 e 15.